Spoken Content Retrieval Beyond Pipeline Integration of Automatic Speech Recognition and Information Retrieval

David N. Racca

Bachelor's in Computer Science

A dissertation submitted in fulfilment of the requirements for the award of

Doctor of Philosophy (Ph.D.)

to the

Dublin City University

School of Computing

Supervisor: Prof. Gareth J.F. Jones

July 2018

I hereby certify that this material, which I now submit for assessment on the programme of study leading to the award of Ph.D. is entirely my own work, and that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge breach any law of copyright, and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

Signed:

(Candidate) ID No.:

Date:

Contents

List of Tables

List of Figures

Abstract

Spoken Content Retrieval Beyond Pipeline Integration of Automatic Speech Recognition and Information Retrieval

David N. Racca

The dramatic increase in the creation of multimedia content is leading to the development of large archives in which a substantial amount of the information is in spoken form. Efficient access to this information requires effective spoken content retrieval (SCR) methods. Traditionally, SCR systems have focused on a pipeline integration of two fundamental technologies: transcription using automatic speech recognition (ASR) and search supported using text-based information retrieval (IR).

Existing SCR approaches estimate the relevance of a spoken retrieval item based on the lexical overlap between a user's query and the textual transcriptions of the items. However, the speech signal contains other potentially valuable non-lexical information that remains largely unexploited by SCR approaches. Particularly, acoustic correlates of speech prosody, that have been shown useful to identify salient words and determine topic changes, have not been exploited by existing SCR approaches.

In addition, the temporal nature of multimedia content means that accessing content is a user intensive, time consuming process. In order to minimise user effort in locating relevant content, SCR systems could suggest playback points in retrieved content indicating the locations where the system believes relevant information may be found. This typically requires adopting a segmentation mechanism for splitting documents into smaller "elements" to be ranked and from which suitable playback points could be selected. Existing segmentation approaches do not generalise well to every possible information need or provide robustness to ASR errors.

This thesis extends SCR beyond the standard ASR and IR pipeline approach by: (i) exploring the utilisation of prosodic information as complementary evidence of topical relevance to enhance current SCR approaches; (ii) determining elements of content that, when retrieved, minimise user search effort and provide increased robustness to ASR errors; and (iii) developing enhanced evaluation measures that could better capture the factors that affect user satisfaction in SCR.

Acknowledgements

First and foremost, I would like to express my most sincere gratitude to my wife, Flor, for joining me in this crazy journey even when this meant leaving so many things behind in the motherland. Thank you for believing in me, for your unconditional support, motivational talks, afternoon mates, and for always reminding me what the important things in life truly are. I would have never been able to finish my thesis without all your help and love.

Secondly, I would like to thank my parents, Claudio and Silvia, my brother and sister, Ivan and July, and my grandparents, Roberto, Mirtha, Isabel and Pancho for taking care of me all times despite being so many kilometres away. Not to have you guys around has certainly been one of the most difficult challenges I have had to face in my years in Ireland. Thank you so much for all the effort you have made in supporting my education and professional career.

Studying in Ireland has been an amazing and culturally enriching experience. During the past few years, I have been lucky enough to meet some wonderful people who in one way or the other have made my PhD journey a delightful and enjoyable experience. I am specially thankful to my fellow lab and university colleagues: Dasha Bogdanova, Maria Eskevich, Piyush Arora, Iacer Calixto, Sheila Castillo, Chris Hockamp, Lorraine Goeuriot, Teresa Lynn, Eva Vanmassenhove, Dimitar Shterionov, Keith Curtis, Peyman Passban, Antonio Toral, Lamia Tounsi, Ali Hosseinzadeh, Ahmad Khwileh, Maria Alecu, Anna Kostekidou, Federico Fancellu and many other colleagues with whom I have had unforgettable times in Dublin.

I am also extremely grateful to my flatmates from VA101, Suzanne Mc Mahon for making us feel at home at all times and teaching us everything we know about Ireland, and Nicolo Fantoni for making sure my caffeine levels were up to standards. I will always be grateful to the Doyle family for all their support, and for making us feel as members of their family.

A huge sense of gratitude goes to my supervisor, Gareth J.F. Jones. Thank you Gareth for your insightful and sharp advice, and for introducing me to the wider research community. I would also like to thank my examiners, Alan Smeaton and Lori Lamel, for their interest in my work and invaluable suggestions for its improvement.

Last but not least, I am thankful to Science Foundation Ireland for the grant No: 12/CE/I2267 Centre for Next Generation Localisation II (CNGL II) and ADAPT that supported this research. Special thanks to Dublin City University, NTCIR, and SIGIR for providing student travelling grants that allowed me to present and discuss my work with the wider research community.

Chapter 1

Introduction

The past few decades have seen an explosion in the amount of multimedia content that is being created and stored in digital format. This accumulation of data has been facilitated by advances in new technologies which have provided individuals with relatively low-cost devices that are able to produce and process high-quality audiovisual material. Almost every person on the planet has now access to powerful recording devices that could fit in a pocket. In combination with advances in mobile networks, this is causing a true revolution in the amount of video and audio that people generate and consume. Instant communication on social media platforms, which had mostly been driven by text, is now more frequently being driven by the sharing of images, voice messages, and video.

In addition to personal users, there is a need to process the increasing volume of multimedia content produced in the enterprise and corporate sector. It is common nowadays to hear *"this call may be recorded for quality assurance purposes"* every time one tries to contact a bank, TV, or internet service providers. Apart from call centres, media professionals involved in the broadcast of radio and TV are interested in tools for the editing, clustering, and automatic transcription of audio and video. Universities and individuals around the world offer online courses based on video lectures and are interested in providing users with tools for browsing and searching through such collections. TV-on-demand services are greatly enhancing the experience of users by implementing, for instance, automatic categorisation of movies, content-based search, and personalised show recommendations. Many companies are now using advanced telecommunication systems with recording capabilities that enable them to store business meetings and oral presentations for later consumption.

In all these contexts, the large amounts of audiovisual content available exceed the capability of users to manually handle, manage, and access the information contained on it. Therefore, it is imperative to develop computational techniques to permit automatic, efficient, and effective access to the relevant information contained within large collections of multimedia recordings. Frequently, much of the information of interest contained in an audiovisual recording is principally encountered within its audio stream, that is, within the spoken content or speech, as opposed to its visual stream which, although important, may only provide non-critical information. Examples of this type of content may include documentaries, interviews, meetings, lectures, and broadcast news, where most of the information is conveyed through speech.

This thesis investigates several aspects relating to the automatic retrieval of relevant information from within collections of multimedia recordings, where most of the information of interest is in spoken form. More specifically, this thesis deals with aspects associated with the use of speech information that go beyond *which* words are spoken to *how* they are spoken, the challenge of recovering from potential errors in the automatic recognition of spoken words, and that of estimating user satisfaction and measuring the quality of a list of search results.

1.1 Overview of spoken content retrieval (SCR)

This section introduces the basic concepts related to SCR. It provides a brief description of the fundamental technologies that are needed for developing practical SCR systems, previous research carried out in the area, and highlights current challenges in the field.

1.1.1 Information access and retrieval from spoken content

Spoken content retrieval (SCR) is concerned with the development of automatic methods to facilitate the search for information in a collection of speech recordings that satisfies an information request from the user (Chelba et al., 2008; Larson and Jones, 2012a; Lee et al., 2015). The reason why a user turns to an SCR system in the search for information is the so-called *information need* (Larson and Jones, 2012b), a term borrowed from the field of information retrieval (IR) (Manning et al., 2008) to refer to the deficit of information which the user is seeking to satisfy by using a search tool.

Commonly, the audio content within a collection is organised as a set of individual audio files, each containing the audio stream of a single recording instance. Less frequently, the spoken collection is just a long continuous stream of unsegmented audio without any given file structure. Even when the collection is organised into individual files, the information contained in each file may well cover multiple topics that users may be interested in. In a meeting retrieval system, for instance, users may be interested in finding the particular location within a meeting where a decision was made or where a particular item from the agenda was discussed. In broadcast news retrieval, interests may vary between finding all recordings covering the same news story to finding all instances where a particular person is mentioned. When retrieving content from lectures or academic presentations, searchers may be interested in finding a lecture they missed, one where a new topic was presented, or the exact moment when the lecturer introduces a new topic.

In order to satisfy the specific information needs that users may have, a SCR system must then provide users with pointers to where the content requested is exactly located in the collection. These pointers can be as simple as a path in a file system indicating which audio file contains the relevant information, or as advanced as a playback tool with embedded audio that, once clicked, commences playback of the audio stream from the exact point in time where the relevant information is located. These set or list of playback pointers are also referred to as "jump-in" or "listen-in" points. The user is said to be satisfied with the pointers produced by a SCR system, if the information being sought can be found effectively within the audio streams by following the playback pointers within a reasonable amount of time.

Reducing the time that users need to spend listening to audio material is critically important for maximising user satisfaction in SCR applications. In fact, time is one of the main reasons why search systems are useful: if time was not a concern, then users could just find the required information by manually assessing every document in the collection. This approach would obviously be inefficient for users who will likely have to spend most of their time assessing irrelevant content. One of the goals in IR is thus to reduce the auditing of irrelevant content, with the ultimate goal of reducing the time and effort required by users to locate the relevant information.

Because information in audio format is less easily accessible than in text format, auditioning time plays a major role in SCR applications. As opposed to the consumption of text content, the consumption of speech requires sequential processing and thus additional time and effort from part of the user. By contrast, textual content is immediately accessible in the sense that the information contained need not be processed in sequence. In addition, text content usually contains explicit structural information (headings, paragraphs, sections) that can facilitate its navigation, permitting almost immediate random access to individual pieces of information. Although structural information may also be present in spoken content, for instance in the form of speaker turns, it remains tacit in the audio stream and is therefore not immediately available to the SCR system. Advanced playback interfaces that permit the increase of playback speed or random seeks may help users reduce auditioning time yet these cannot provide users with immediate access to speech content which still needs to be listened to by users.

The consideration of aspects related to the access of information in audio and text media establish a clear difference as to how user satisfaction or the effectiveness of a retrieval system should be measured. While in the text domain, "retrieval effectiveness" is frequently quantified by the amount of relevant material that is returned to the user, relative to the amount of irrelevant material. In the speech domain, "retrieval effectiveness" must additionally take into consideration factors related to the temporal characteristics of speech media such as the amount of time users waste in listening to non-relevant material.

1.1.2 SCR system overview

Stated naively, SCR is the application of automatic speech recognition (ASR) and textual information retrieval (IR) to collections of speech recordings (Larson and Jones, 2012a).

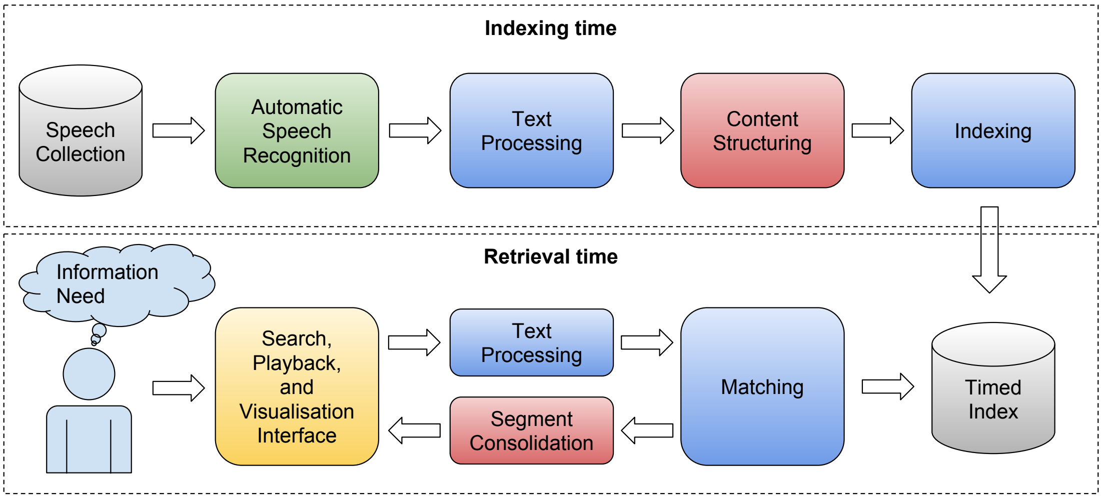


Figure 1.1: Block diagram showing the architecture and components of a conventional SCR system.

In the so-called "cascading" approach to SCR (Lee et al., 2015), the following processing steps are involved: (i) an ASR system is used to convert speech into text, particularly to obtain a text transcript of every spoken document in the collection; (ii) an IR engine is then used to create an index of the text collection and to rank transcripts in order of estimated relevance to the user query; and (iii) playback pointers corresponding to the top ranked transcripts are then generated and retrieved as search results to the user for further consumption.

Figure 1.1 depicts the architecture and main components of a standard SCR system. The dashed lines in the diagram divide components into two large groups: those that are used at indexing time to construct a timed search index (top group), and those that are used at retrieval time (bottom group) to generate the search results. The timed index file created in the indexing process is a series of data structures containing information about the occurrence of individual spoken words across documents, along with their time of incidence within the audio streams. Since ASR systems are incapable of recognising words that are not in the recognition vocabulary, the timed index may include additional information to help search for out-of-vocabulary terms. This includes word proxies (Chen et al., 2013), lattices or N-best list, or subword units such as morphemes or phonemes. While indexing components do not generally have major limitations in terms of processing time, retrieval components are designed so that search results can be produced as quickly as possible (less than a second in practice) to avoid wasting the time of the user.

Beyond their classification into indexing and retrieval time, the components of an SCR system can be additionally grouped by their functionality: IR components, which deal with the indexing, processing, and searching of textual data; ASR components, which perform speech-to-text conversions; and content structuring components, which provide the means for detecting relevant regions within large spoken documents and determining the location of playback pointers.

Information retrieval

Information retrieval (IR) (Manning et al., 2008) deals with the problem of finding content that is relevant to a user's information need within a collection of items. A user typically expresses their information need as a text query, which is most commonly formulated as a sequence of keywords or as a description in natural language. The query is then provided as an input to an IR system which searches for items in the collection containing one or more words from the query and presents them back to the user as a list of items ranked by their estimated likelihood of relevance.

In order to provide quick search response times, text indexing techniques (Zobel and Moffat, 2006; Manning et al., 2008) are used to construct a search index. The index is pre-populated with pointers and statistics about the occurrence of words in the documents so that documents that match the query can be efficiently identified at retrieval time. During this process, a lexicon containing the list of unique words found in the documents is created along with an inverted index, which stores information about the number of times a particular word occurs in a given document.

Prior to the construction of a search index, the text contained in the documents needs to be processed. This process usually consists of tokenisation, removal of punctuation symbols and stop words, and stemming (Manning et al., 2008). In the context of IR, tokenisation involves the identification of linguistic units to be used as indexing terms of documents. Normally, only semantically meaningful units of a language such as phrases, words, morphemes, phonemes, are used as indexing terms. Stop word removal is useful in IR because it reduces the size of the index without significantly harming retrieval performance. This is because terms that occur frequently in the collection are less useful in distinguishing relevant from non-relevant documents. Finally, stemming is used to cluster semantically similar terms with different suffixes into a single equivalence-class. When a query is provided to the IR system, the text of the query is processed in a similar manner as documents in order to maximise the overlap between them.

In IR jargon, "matching" refers to the process of scoring every document in the collection against the query. These scores are estimated based on the number of terms shared between the query and each document, so that they either reflect the probability of relevance of the documents, or their degree of semantic similarity with respect to the query. The total order induced over the collection of documents by these relevance scores can then be used to suggest the order in which documents should be inspected by the user.

Two popular ranking models are the vector space model (VSM) (Salton, 1979) and the probabilistic relevance model (Spärck Jones et al., 2000). Under the scope of these general models, several ranking functions have been proposed (Salton and Buckley, 1988; Zobel and Moffat, 1998; Robertson et al., 1994), most of which calculate a relevance score as a linear combination of weights associated to each term in the query matching the document.

The main principle governing how weights are assigned to terms in a document states that higher weight values should be given to terms that are representative of the content of the document and that can discriminate this document from others. Term weights are typically defined based on three fundamental statistics: (i) the number of times a term occurs in the document under consideration; (ii) the length of this document; (iii) and the number of documents in which a term appears across the whole collection. Several schemes have been proposed in the past that define functions for deriving effective term weights from frequency information (Zobel and Moffat, 1998). Usually, terms with high within-document frequencies relative to the length of the document, and with low document frequencies relative to the size of the collection are given larger weight values. Because relevance scores are calculated as the sum of term weights, those documents containing higher weighted terms in the current query are thus likely to appear at higher ranks in the list of results for this query.

Content structuring (segmentation)

In order to reduce the amount of time that users need to spend auditioning audio material, an SCR system should ideally indicate the most likely starting time of the relevant part in the audio file and also potentially the time span of material that contains likely relevant information. In practice, this is normally achieved by splitting documents into a set of sub-documents, referred to either *passages* or *segments*. The resulting sub-documents can be then treated as documents from a IR perspective, and be indexed and later ranked according to their relevance score against the query.

The process of splitting documents into passages for retrieval has long been the focus of research in the IR community and is known as passage retrieval (Callan, 1994; Kaszkiel and Zobel, 1997, 2001). A generalisation of passage retrieval is XML retrieval, where the passages to be ranked are organised in a hierarchical fashion into multiple levels of content granularity (Fuhr et al., 2002; Fuhr and Lalmas, 2007). The most effective passage retrieval and XML retrieval techniques exploit document-level as well as passage-level evidence at different granularity levels, for improved ranking of relevant passages or documents (Kaszkiel and Zobel, 2001; Ogilvie and Callan, 2005; Arvola et al., 2011). Techniques that seek to improve the ranking of relevant passages given the context from their container documents and that of their adjacent passages, are known as contextualisation techniques (Kekäläinen et al., 2009; Arvola et al., 2011).

Research in SCR has applied passage retrieval techniques for finding listen-in or jump-in, and listen-out or jump-out time points close to the beginning and respectively the end of relevant fragments in collections of spontaneous and conversational speech (Oard et al., 2006; Larson et al., 2011; Eskevich et al., 2013a; Akiba et al., 2011). In these approaches, the spoken collection is first segmented into short passages, which are then ranked by relevance to the query. The playback pointers to be shown to the user are then given by the time offsets of the ranked passages relative to the start of the documents where they occur in. Most approaches adopted for segmenting spoken collections into individual passages are based on windowing (Stanfill and Waltz, 1992; Callan, 1994; Kaszkiel and Zobel, 1997, 2001) or automatic text segmentation methods (Hearst and Plaunt, 1993; Choi, 2000; Malioutov and Barzilay, 2006).

Windowing consists of generating passages by moving a fixed-length window across the text document. The window is positioned at the beginning of the document and moved towards the end in steps given by a fixed length unit. A new passage containing the words that fall within the sliding window is generated at each step until the end of the document is reached. Additional improvements in retrieval performance can often be obtained by setting the step length to be smaller than the length of the window so that the resulting passages overlap (Stanfill and Waltz, 1992; Callan, 1994; Kaszkiel and Zobel, 1997). The length units are usually defined in terms of time or in number of words (Quinn and Smeaton, 1999).

Text segmentation algorithms seek to divide a text or speech document into semantically coherent units by exploiting features that are informative of topic shifts. These include methods based on lexical cohesion (Hearst, 1997; Reynar, 1998; Choi, 2000; Malioutov and Barzilay, 2006), and others that exploit multimodal features in either a supervised or unsupervised fashion (Reynar, 1998; Shriberg et al., 2000; Tür et al., 2001; Galuščáková and Pecina, 2014b).

When overlapping passages are indexed, the matching component may return pointers to passages that overlap in the result list. In the process of doing this, it may assign different ranks to passages that are adjacent in the original speech recordings. Depending on the application domain, users may be dissatisfied if presented with a list of similar playback pointers since these may be perceived as duplicate results. Two general segment consolidation strategies have been developed to deal with these issues: filtering (Wartena, 2012) and recombination (Abberley et al., 1999b; Johnson et al., 2000). Filtering consists of removing passages from the list of results that overlap or that are close to another passage ranked higher in the result list. In this strategy, only the result with the highest rank is kept. Recombination consists of merging passages that overlap or that are close to another passage ranked higher in the result list. In this case, the combined passage is normally assigned the rank of the highest scoring merged passage.

Automatic speech recognition (ASR)

Automatic speech recognition is concerned with the identification of words spoken in continuous speech, possibly by multiple speakers, across highly variable acoustic conditions (Levinson et al., 1983; Rabiner, 1989). Early ASR systems were only capable of recognising among a small number of words spoken in isolation, by a single speaker, in controlled recording environments. Subsequent improvements of ASR technology during the 1980s and 1990s gave rise to large vocabulary continuous speech recognition (LVCSR) systems capable of transcribing speech produced by multiple speakers and considering a much larger number of words (60,000 or more) (Gauvain et al., 1999; Rousseau et al., 2011).

The most effective speech recognition systems are based on statistical models that are able to handle the high complexities of the speech signal as well as the high variations that exist in spoken language. A popular statistical framework for ASR systems models the mapping between phonemes underlying spoken words and acoustic input from the speaker via hidden Markov models (HMMs) (Levinson et al., 1983), and the space of possible word sequences in a language via statistical language models (LMs) (Katz, 1987). The recognition process then consists of searching for the sequence of words that best explains the acoustic patterns observed and that has the highest language model probabilities. To make this inference practical, the number of possible words that can be recognised is fixed in advance, limited by the vocabulary of the language model.

ASR systems can produce predictions in multiple formats. A lattice is a graph that represents multiple hypotheses made by the recogniser, where nodes are points in time and arcs represent hypothesised words along with their confidence scores. The 1-best hypothesis is the sequence of words corresponding to the path in which the ASR system has greatest confidence. Typically, SCR systems only consider the 1-best hypothesis from the ASR in the indexing process, although advanced matching techniques (James and Young, 1994) may consider recognition units from less likely hypotheses in an attempt to match words from the query that may be missing from the 1-best hypothesis or the LM vocabulary.

Despite recent improvements in ASR technology (Hinton et al., 2012), transcription errors are still a common issue in modern ASR systems, especially in domains where speech is informal, unstructured, spontaneous, and conversational. The quality of ASR systems is frequently measured by estimating the word error rate (WER) of an ASR hypothesis, by counting the number of word deletions, substitutions, and insertions with respect to the perfect transcription of the utterance. State-of-the-art ASR systems can produce transcripts with WERs that range between 9%-11% for broadcast news (Bell et al., 2015; Wu et al., 2016), 10%-40% for multi-genre TV broadcast (Bell et al., 2015), 5%-40% for general spontaneous conversational speech (Lileikyte et al., 2015; Xiong et al., 2016; Chiu et al., 2017; Enarvi et al., 2017), and 45%-50% for YouTube videos (Hinton et al., 2012). Recognition rates can vary greatly depending on the domain, genre, spontaneity, language, and audio quality of the speech material as well as the amount of training data and computing resources available. With sufficient training and computing resources, ASR technology can attain WERs as low as 5% for relatively clean telephone conversations in American English (Xiong et al., 2016). By contrast, in more challenging conditions, practical ASR systems can transcribe conversational speech with WERs as high as 20%-40% (Lileikyte et al., 2015; Chiu et al., 2017; Enarvi et al., 2017).

As SCR systems principally rely on finding occurrences of query terms in ASR transcripts, ASR errors represent one of the main challenges in achieving effective retrieval.

1.1.3 Open problems in SCR

In order to motivate the research questions addressed in this thesis, this section describes in detail some of the limitations present in existing approaches to SCR, as well as aspects of SCR that have not been explored in full by previous research.

The problem of handling ASR errors in the speech transcripts

Due to the inherent difficulty of the speech recognition task, ASR systems produce erroneous transcriptions of the spoken material. This results in incorrect words being inserted and correct words being substituted or deleted in the predicted text. These errors complicate the task of the IR engine which principally relies on finding overlapping terms between the query and the documents to find relevant documents.

While human transcripts are free from ASR errors, transcripts produced by an ASR system are relatively inexpensive to obtain. They are also free from misspellings and, most importantly, contain word time information which is necessary in SCR for determining the potential location of candidate relevant regions within long spoken documents. Practical SCR thus requires the indexing of ASR transcripts which in turn necessitates of retrieval techniques that could handle recognition errors effectively.

Research in SCR has mainly focused on understanding how ASR errors affect the performance of IR models and on developing techniques to make retrieval robust to these errors. These aspects were extensively explored in the context of the Text REtrieval Conference on spoken document retrieval (TREC SDR) benchmarks (Voorhees and Harman, 2005) which evaluated the effectiveness of SCR systems over a collection of broadcast news speech recordings. Several techniques were then proposed to deal with ASR errors, notably including: the exploitation of multiple hypothesis produced by the ASR (Crestani et al., 1997; Siegler et al., 1997; Tsuge et al., 2011); the indexing of phonetic units instead of words (James and Young, 1994; Smeaton et al., 1997; Chelba et al., 2008); and the application of pseudo-relevance feedback (PRF) to expand the query and document's contents with terms extracted from external error-free corpora (Johnson et al., 1999b; Singhal and Pereira, 1999; Woodland et al., 2000).

The techniques developed at TREC SDR were found so effective at reducing the impact of ASR errors over IR performance that SCR was considered a "solved problem" (Garofolo et al., 2000). However, later analysis suggested that broadcast news speech does not present major difficulties for SCR since this type of speech content is normally planned, formal, redundant, and clearly delivered (Allan, 2001). For collections containing recordings of spontaneous or conversational speech, such as interviews, business meetings, or telephone conversations, it was later discovered that ASR errors can significantly decrease the effectiveness of SCR systems (White et al., 2005; Eskevich et al., 2012c; Akiba et al., 2011). The increased difficulty was attributed to the characteristics of casual speech, where information tends to be conveyed by a less diverse set of content-bearing words, and structural cues, such as topic shifts, are less clearly delivered.

In addition to the spontaneity levels of the speech content, it has been also pointed out that ASR errors may have a major impact on SCR effectiveness when the units of text to be ranked are short passages (60-100 words) extracted from an ASR transcript (Allan, 2001). The main reason being that short passages may not contain enough occurrences of query terms for the matching process to be able to recover from query terms being misrecognised by the ASR. Although considering longer excerpts of text, containing a greater number of terms, may seem like a reasonable solution, previous research has shown that the use of long passages can be detrimental to SCR performance when the granularity of the passages differs from that of the relevant content (Wartena, 2012; Eskevich et al., 2014).

As evidenced in previous research, expansion of passages with related terms extracted from in domain parallel corpora, can offer increased robustness to ASR errors (Johnson et al., 1999b; Singhal and Pereira, 1999; Woodland et al., 2000). Nonetheless, these techniques require the availability of an external text corpus with a domain similar to the target collection, which may be difficult to obtain. Contextualisation techniques (Kekäläinen et al., 2009; Arvola et al., 2011) can offer an alternative solution to passage expansion that does not require external corpora. In these techniques, the relevance score of a passage is computed based on the terms contained in the passage plus those contained in the remainder of the document. Although contextualisation techniques have been shown effective in textual passage and XML retrieval tasks (Carmel et al., 2013; Arvola et al., 2011), only a limited amount of work has explored their effectiveness in SCR (Nanjo et al., 2014; Shiang et al., 2014), while none of these has properly evaluated the capabilities of these techniques for reducing the impact of ASR errors.

The challenge of exploiting speech beyond lexical information

Current approaches for retrieval and content structuring for SCR have mainly sought to exploit the lexical representation of the spoken content that results from the ASR process, omitting other valuable information that is encoded in the speech signal. However, beyond its lexical representation, speech is known to encode richer information about what is said through the way words are pronounced. This is known as prosody and includes the pitch, duration, and loudness of speech.

Variations in pitch, duration and loudness have frequently been associated with various aspects of spoken communication. They are used for marking emphasis or focus on particular words, indicating the intentions or speech acts of an utterance, expressing emotions and attitudes, and facilitating the understanding of ambiguous syntactic expressions (Wagner and Watson, 2010; Hirschberg, 2002). Furthermore, prosody is believed to encode the information status of words and how this status changes over time. There is evidence that words considered “new”, “important”, “focused”, “not given”, “unpredictable”, “inaccessible”, or “informative” in a discourse are more likely to be emphasised acoustically than others (Prince, 1981; Hirschberg and Grosz, 1992; Silipo and Crestani, 2000; Hirschberg, 2002; Wagner and Watson, 2010; Ward and Richart-Ruiz, 2013; Röhr, 2013). Acoustic emphasis given to a particular word in speech is known as acoustic/prosodic prominence. Thus, while ASR transcripts are generally noisy and content originating in spoken form is likely to be more informally structured, spoken content has significant amounts of expressive information available that might also potentially be exploited in the segmentation and retrieval processes.

Although a considerable amount of research has been done in the exploration of the utility of prosodic information in various speech processing tasks, such as speech summarisation (Chen and Withgott, 1992; Koumpis and Renals, 2005) and segmentation (Hirschberg and Grosz, 1992; Shriberg et al., 2000), little research has been done to explore its direct utility in SCR. It was suggested that prosody can be used to improve the ranking of relevant content in SCR (Silipo and Crestani, 2000). This is because acoustically prominent words tend to be also those that are most descriptive of the content being conveyed, according to the term weights produced by a ranking function (Crestani, 2001).

Building upon Silipo and Crestani's findings, other researchers have attempted to exploit prominence information in SCR and topic tracking tasks by combining lexical information of words with acoustic features for the calculation of enhanced term weights (Chen et al., 2001; Guinaudeau and Hirschberg, 2011). Their approach consists of implementing an alternative term weighting scheme which increases the lexical weight of terms whenever their individual occurrences are found to be emphasised in the speech content. Researchers obtained mixed results with this technique. While prominence information was found useful for improving the retrieval of speech fragments discussing similar topics in a French corpus (Guinaudeau and Hirschberg, 2011), preliminary SCR experiments conducted over broadcast news speech in Mandarin Chinese showed no benefits from using these enhanced term weights in an SCR task.

Considering the ambivalence of these findings, it is thus unclear if prosodic prominence information could be effectively used to improve existing term weighting techniques for SCR. Furthermore, limitations of the speech collections available at the time put a restriction in the set of SCR experiments that researchers could conduct and limited their ability to address this problem in more detail. Recently, researchers have collected and released new test collections for SCR research that contain a large number of speech documents transcribed with improved ASR systems, numerous examples of search queries, as well as high-quality relevance assessments in various levels of spoken content granularity. It is therefore worthwhile to revisit the problem of exploiting prosodic information over these new datasets to seek for definitive answers and extend previous analysis to new languages, genres, and SCR tasks.

The problem of structuring content and of measuring user satisfaction

Content structuring still remains an open issue in SCR despite having been the focus of a number of existing studies (Eskevich et al., 2012b; Wartena, 2012; Eskevich et al., 2013c, 2014; Galuščáková and Pecina, 2014b). While the segmentation problem is generally unambiguous for formal and planned speech in which the information is explicitly presented in a structured manner, this is not the case in domains where speech is conversational and spontaneous. In such cases, segmentation into unambiguous semantically meaningful units may not be possible. For instance, in broadcast news, information is normally presented as a sequence of distinct news stories where boundaries between stories are easily recognisable. By contrast, the structure of a business meeting or a lecture may be less obvious and therefore harder to recognise automatically.

Popular approaches to automatically segment spoken material for SCR purposes fall into two broad categories. The first seeks to identify topic boundaries based on the lexical and acoustic properties of the transcribed spoken material (Hearst, 1997; Shriberg et al., 2000; Malioutov and Barzilay, 2006). The second, based on sliding windows, disregards topic structure and seeks to divide speech into arbitrary passages of similar length (Stanfill and Waltz, 1992; Kaszkiel and Zobel, 1997, 2001). Surprisingly, the latter approach has proven considerably more effective in work to date (Tiedemann and Mur, 2008; Wartena, 2012; Eskevich et al., 2012b; Galuščáková and Pecina, 2014b). The reason being that arbitrary passages are less affected by ASR errors; they can alleviate the difficulties associated with estimating relevance scores for passages that vary in length; and they can adapt better to different information requests. However, careful investigation of window-based approaches reveals them often to be sub-optimal, to provide poor playback listen-in points with consequential poor user experience, and to negatively affect the effectiveness of retrieval models compared to an optimal segmentation (Kaszkiel and Zobel, 2001; Eskevich et al., 2012b; Wartena, 2012).

Much of the difficulty in determining if there is a single superior content structuring approach for SCR, has been associated with the problem of evaluating the output of SCR systems. Early evaluation measures proposed for estimating the quality of SCR results were based on adaptations of standard measures originally developed for the evaluation of document (Harman, 1993), passage (Allan, 2004), or XML retrieval (Kamps et al., 2007) tasks, which estimate the proportion of relevant content retrieved at top ranks relative to the amount of irrelevant material. Although these measures may be appropriate in the context of text retrieval, they do not account for the temporal aspects involved in the auditioning of spoken content, namely, the time a user must invest in listening to the audio snippets retrieved.

Improved adaptations of evaluation measures have been proposed by a number of researchers (Liu and Oard, 2006; Galuščáková et al., 2012; Eskevich et al., 2012c; Aly et al., 2013a) to take account of a number of dimensions that are believed to affect user satisfaction in SCR. The main aspects considered being: the amount of relevant content retrieved measured in time units, its ranking, and additional time constrains such as the distance between the time pointers returned by the system and the beginning of the relevant content. Despite these improvements, most of these measures tend to assign disproportionate importance levels to the relevance, ranking and time dimensions, and can thus only offer a partial solution to the evaluation problem. Novel measures for IR evaluation proposed recently (Moffat and Zobel, 2008; Chapelle et al., 2009; Smucker and Clarke, 2012) attempt to model the behaviour of users when assessing a ranked list of results, but have not been fully explored in the context of SCR.

1.2 Research questions

Considering the open problems in SCR discussed in Section 1.1.3, as well as the previous research carried out in the area, this thesis investigates existing and proposes novel techniques for SCR along three directions: (i) the utilisation of non-lexical acoustic information for the detection of informative keywords; (ii) the adoption of contextualisation techniques for increasing SCR robustness to ASR errors; and (iii) the development of novel evaluation measures that could permit a fair comparison of different content structuring methods in SCR.

With regards to the challenge of exploiting non-lexical information, this work advances the investigations of Crestani (2001), Chen et al. (2001), and Guinaudeau and Hirschberg (2011), and studies the utility of acoustic/prosodic prominence features for improving existing SCR indexing techniques and term weighting schemes. In particular, the focus is on determining whether acoustic features derived at the word-level can be effectively used to estimate important mentions of indexing terms, and whether this acoustic evidence can be further combined with lexical features to improve SCR effectiveness. These objectives can be summarised in the following research questions:

RQ-1: Can information about which prosodic units are made prominent in speech be combined with lexical information to derive improved term weighting schemes and retrieval functions that could enhance SCR effectiveness?

This thesis seeks to answer RQ-1 empirically by conducting SCR experiments with retrieval functions that combine prosodic prominence and lexical information about terms to calculate relevance scores.

With respect to the challenge of handling ASR errors in the speech transcripts, this thesis investigates if contextualisation techniques can make the ranking process more robust to ASR errors. In this regard, the task under investigation is passage retrieval, in which the units to be retrieved are relatively short in length and may not contain sufficient terms to compensate for speech recognition or segmentation errors. This objective can be stated more formally as:

RQ-2: Can contextualisation techniques increase the robustness of standard text retrieval approaches to ASR errors when the retrieval units are made from short fragments of speech transcripts?

In order to answer RQ-2, the effects on retrieval effectiveness produced by different contextualisation techniques are analysed under various conditions of speech recognition errors in the transcripts.

Lastly, in relation to the problems of content structuring and evaluation in unstructured collections, this thesis first provides a critical overview of existing evaluation measures for SCR, and then investigates alternative measures that could provide more appropriate estimates of user satisfaction in the context of SCR. These alternative evaluation measures are then used to carry out an unbiased comparison of different content structuring techniques applied to SCR with the goal of determining which technique results more effective in terms of maximising user satisfaction. This set of goals can be summarised in the following research questions:

RQ-3-A: Can existing evaluation measures for SCR estimate levels of user satisfaction appropriately?

RQ-3-B: Can enhanced evaluation measures be developed to address the shortcomings of existing evaluation measures for SCR?

RQ-3-C: Which content structuring techniques are most effective in SCR in terms of maximising user satisfaction?

Answers to these research questions are first sought by reviewing previous research in IR and SCR evaluation, emphasising work that has focused on aspects related to the modelling of user browsing behaviour when scanning a ranked list of search results. Based on this analysis, a novel framework for SCR evaluation is developed and finally used to study the effectiveness of different content structuring approaches.

1.3 Thesis structure

This thesis begins by describing the principal technologies underlying modern SCR systems: information retrieval for text collections (IR), automatic speech recognition (ASR), and content structuring applied to IR tasks. It then continues with an in depth overview of previous research conducted in SCR, emphasising previous studies that focused on the interactions between ASR errors and IR techniques, the comparison of content structuring methods, and the exploitation of acoustic/prosodic information. This is followed by a description of the collections and software used for the experimental work in this thesis. The development of techniques and experimental work carried out in this thesis are then presented, followed by the conclusions and suggestions for future work.

The remainder of this thesis is structured into the following chapters:

Chapter 2 overviews the fundamental technologies needed for SCR. It starts by describing basic concepts and existing techniques used for creating indexes and retrieving relevant content within large collections of text documents, including those used in the experimental work of this thesis. This is followed by a description of the fundamental aspects related to ASR technology, including a high-level overview of the individual components required for an operational ASR system. Most content structuring approaches adopted in SCR are based upon research on the application of automatic text segmentation techniques to text retrieval tasks. Chapter 2 thus examines these techniques in detail.

Chapter 3 provides a critical review of previous and current research in SCR. The earliest experimental studies in SCR focused on relatively small collections of voice mail and broadcast news, and then switched onto more challenging conversational speech content such as interviews, general TV broadcasts, and lectures. Much of previous research in the area has mainly been driven by evaluation campaigns and research benchmarks, and has focused on the challenges of handling ASR errors and structuring content. Although not part of mainstream research, previous studies have investigated the potential benefits of using acoustic/prosodic information to improve SCR effectiveness. All of these studies are discussed in Chapter 3.

Chapter 4 describes the speech collections, queries, and software used in the experimental work of this thesis. Ideally, test collections for SCR research must: (i) be large enough to account for a varied number of interesting topics to search for; (ii) have available queries with associated relevant judgements, preferably carried out at sub-document granularity levels; (iii) be transcribed automatically by at least one ASR system. Due to the lack of availability and high costs associated to the creation of these data sets, the experimental work in this thesis is based on spoken collections that, despite not meeting all requirements outlined above, are still useful for the goals set in this thesis. In particular, the BBC collection is a relatively large (3000 hours) dataset containing English recordings of general TV content (talk shows, documentaries, series, etc) with 100 queries and low-quality fine-grained relevance assessments. The Spoken Document Processing Workshop (SDPWS) collection is a small (30 hours) data set of lecture recordings in Japanese, that contains a 230 queries, high-quality fine-grained relevance assessments, and a large number of transcripts produced by ASR systems of different quality.

Chapter 5 describes a series of experiments that seek to determine whether acoustic/prosodic information can be used to improve current lexical-based term indexing techniques. This chapter first describes the approach adopted for feature extraction and their posterior word-alignment against speech transcripts. It then elaborates on the derivation of heuristic-based prominence scores for individual indexing terms, and on their integration into a ranking function for speech content. Two groups of experiments are then described. The first group investigates if prominence scores can provide a meaningful increase in retrieval effectiveness when integrated via the heuristic-based approach. The second group uses machine learning techniques to study the relationship between prominent and informative terms, as well as the value that prominence information might have for improving content ranking in SCR.

Chapter 6 investigates the benefits of using contextualisation techniques for improving the ranking of speech passages in adverse conditions of ASR errors. This chapter begins by motivating the adoption of these techniques in SCR. Existing contextualisation techniques are then described and their ability to improve retrieval effectiveness evaluated under different conditions of ASR errors in the speech transcripts.

Chapter 7 introduces a novel user-centric framework for the evaluation of spoken passage retrieval. Evaluation measures under this framework are then used to carry out a large-scale comparison of existing content structuring approaches.

Chapter 8 describes the conclusions of this thesis, provides concrete answers to the research questions stated in Section 1.2, and suggests directions for future work.

Appendix A provides a list with all publications derived from this dissertation.

Appendix B describes a series of index similarity metrics used to measure the quality of a search index built from ASR transcripts.

Appendix C provides a detailed description of LambdaMART, a learning-to-rank method based on regression trees that was used in the experiments presented in Chapter 5.

Appendix D describes the general optimisation method used to tune the parameters of retrieval models in the experiments presented in Chapters 5 and 6.

Appendix E describes the results of retrieval experiments with a SCR method that exploits prosodic/acoustic features by leveraging the output of a binary classifier.

Chapter 2

Review of Fundamental Technologies in SCR

Information retrieval (IR) is the study and development of automatic indexing and ranking techniques that permit searching for relevant information within a collection of information sources. These techniques seek to solve the problem of “content overload” in which searching for a particular piece of information by browsing becomes impractical as the size of the collection grows over time. Content overload is more severe in spoken collections, since the browsing of speech material is more time consuming than the browsing of text. For this reason, IR is a fundamental technology to enable practical SCR systems for collections of more than trivial size.

Applying automatic text indexing and ranking techniques to collections of speech recordings requires the ability to recognise and quantify important keywords or indexing terms that are spoken in the audio streams. For this purpose, current SCR applications make use of Large Vocabulary Continuous Speech Recognition (LVCSR) technology, or Automatic Speech Recognition (ASR) in short.

When the spoken documents to be indexed discuss more than one topic or when they are too long to be auditioned within a reasonable amount of time, it is convenient to segment documents into shorter units that could be individually indexed and retrieved by the SCR system. Decisions involving how to best divide a spoken document into topically homogeneous retrieval units with the objective of maximising retrieval effectiveness while minimising user-auditioning time lie in the realms of content structuring technologies for SCR.

This chapter presents a review of these three technologies that are fundamental for SCR applications. Section 2.1 describes automatic text indexing and retrieval. Section 2.2 reviews fundamental concepts on ASR technology. Finally, Section 2.3 examines content structuring and topic segmentation techniques, while their applications to text retrieval tasks are reviewed in Section 2.4.

2.1 Information retrieval (IR)

In a broad sense, IR deals with the problem of finding documents that are relevant to an information need provided by the user in the form of a query. When documents and queries are given in natural language, and when the goal is to produce a ranking of the most relevant documents to a query, this task is commonly known as *ranked* retrieval. To solve this task efficiently, a standard IR system first constructs a search index of the document collection, which permits fast access to term statistics at querying time. These aspects related to indexing and document representation are described in Section 2.1.1.

When a query is issued by the user, a retrieval model is then used to produce a ranking of matching documents. In this regard, Section 2.1.2 describes some important models for ranked retrieval, including the one used in the experiments described in this thesis. Finally, Section 2.1.3 reviews some of the evaluation measures introduced in previous research which seek to measure the quality of the document rankings produced by a model.

2.1.1 Text pre-processing and indexing

Scaling the application of ranked retrieval to collections of hundreds of millions of documents is only possible in practice through the construction of efficient search indices. In a general sense, a search index is a data structure that stores information about the documents that comprise the collection to be searched. The most important property about a document that is stored is the number of times a particular indexing feature "points to" or "appears in" the document. In this context, an indexing feature refers to some quantifiable property of the document that may be also present in other documents in the collection. When dealing with documents in natural language, the most commonly used indexing feature is the word. A query issued to the IR system can then be characterised in terms of the set of indexing features that should preferably be present in the highest ranked documents returned by the system or that should influence how such ranking is constructed.

The first step towards the construction of a search index is to identify and extract indexing features from the documents that comprise the collection to be searched. This step usually requires processing the text with a tokeniser or text segmenter, designed to divide a document string into a sequence of tokens. Each token identified by a tokeniser roughly corresponds to a particular word from the language the text is written in. For text in most European languages, in which words are separated by spaces, text tokenisation is a fairly simple task and can be done with a carefully designed set of regular expressions to handle the uses of apostrophes, hyphens, and punctuation symbols. However, for *scripto continua* languages like Thai, Japanese, or Chinese, where words boundaries are not explicitly marked, tokenisation is a less trivial task. In these difficult cases, it is common to perform tokenisation by using statistical sequential models (Zhang et al., 2003; Kudo et al., 2004; Shao et al., 2017). The tokenisation process may additionally involve the standardisation of numbers, proper names, and other special words or symbols considered important for retrieval.

As an additional pre-processing step, it is common practise to discard tokens that have little or no value for retrieval. These tokens generally correspond to punctuation symbols and stop words. Stop words are generally function words that are frequently used in the language and consequently less useful in distinguishing relevant from non-relevant documents. Finally, stemming or lemmatisation are linguistic processing techniques used to map semantically related tokens that differ in their surface form into a single equivalence class. Lemmatisation consists of mapping a token to its base form or dictionary entry form (e.g. "walking" to "walk"). While stemming can be seen as a cheap alternative to lemmatisation and consists of removing/replacing the endings of tokens in order to reduce their inflectional variations (e.g. "walking" to "walkin"). A popular and effective stemming algorithm for English is Porter's algorithm (Porter, 1980). For each document, the stemming or lemmatisation processes produce the ultimate sequence of modified tokens that will be included in the search index. These resulting tokens are called "indexing terms" or just "terms", and the set of all terms in the collection is known as the index vocabulary or lexicon.

Several indexing algorithms have been designed for the construction of search indices (Zobel and Moffat, 2006). The main objective of indexing is then to build data structures that could be later used at querying time to score millions of documents efficiently. Two important data structures generated are the lexicon and the inverted index. The lexicon is a mapping of terms to term IDs with possibly additional information about the terms such as their document frequency and a pointer to its location in the inverted index. The inverted index holds a list of postings for each term in the lexicon. Each posting consists of a document frequency (d, tf) pair where tf indicates the number of times the term appears in a document d. In order to support phrase queries and proximity search, postings are commonly augmented with the positions at which the terms appear within the documents. Also, when indexing speech transcripts, postings can be extended with acoustic features that may be available for each term, such as confidence scores or word time-stamps.

2.1.2 Frameworks for ranked retrieval

A framework for ranked retrieval consists of a set of ideas, methods, and principles that specify how a set of documents may be ranked in order of relevance to a query. Most standard IR frameworks stipulate that this ranking can be constructed via a function, designed to calculate a numeric score for each document that reflects its degree of relevance with respect to the query. In the IR literature, this score is commonly referred to as a retrieval status value (RSV) or ranking score (S). To implement such a function, most standard frameworks adopt a "bag of words" representation for queries and documents in which these elements are represented by a set of indexing features (terms) taken from a fixed vocabulary. In a "bag of words" representation, the order in which terms appear in an element is completely ignored, as is the fact that some terms may condition the presence or absence of others within or across elements in the collection. Two major frameworks that have been used extensively in SCR research are the vector space model (VSM) (Salton, 1979), and the probabilistic model (Spärck Jones et al., 2000) for ranked retrieval.

The Vector Space Model (VSM)

The vector space model (VSM) (Salton et al., 1975) is one of the oldest and most widely adopted models in IR. In this model, queries and documents are represented as vectors in which every component is associated to a particular term in the vocabulary. More particularly, the vector of a document (query) is constructed so that its i-th component contains a score or weight that reflects the extent to which its associated term is considered representative of the topic of the document (query). For a collection C with M distinct terms indexed by $i : 1 \leq i \leq M$, a VSM represents a document by a vector $\vec{d} = \langle d_1, \ldots, d_M \rangle \in \mathbb{R}^M$, where d_i is the weight associated to the i-th term in the document. Similarly, a query is represented by a vector $\vec{q} = \langle q_1, \ldots, q_M \rangle \in \mathbb{R}^M$ with q_i denoting the query vector's i-th component. In the application of the VSM to IR, it is common to assign positive weights to terms that are present in a document (query) and zero weights to terms that are absent.

The underlying assumption in a VSM is that elements that are semantically similar will lie in similar regions in the vector space. Based on this assumption, the relevance of a document d with respect to a query q can be computed as the distance between their vector representations in $\mathbb{R}^M$. When the cosine similarity is used as a measure of distance, the ranking score of d for q is calculated as shown in Equation 2.1.

$$S_{VSM}(q, d) = \frac{\vec{q} \cdot \vec{d}}{\|\vec{q}\| \; \|\vec{d}\|} = \frac{\sum_i^M q_i \; d_i}{\sqrt{\sum_i^M q_i^2} \; \sqrt{\sum_i^M d_i^2}} \tag{2.1}$$

The set of functions that establishes how term weights are calculated is known as a weighting scheme. Weighting schemes are defined in such a way that terms that are more informative of the topic of a document obtain higher weights for that document. For retrieval purposes, a term is considered informative for a document if it represents the topic of the document and if it is effective in discriminating this topic from others that may be also present in the collection.

The weighting scheme generally adopted in a VSM involves the product of two factors: the within-document term frequency $tf_d(i)$, based on the number of times that the i-th term occurs in the document d; and the inverse document frequency $idf(i)$, based on the number of documents in C that contain the i-th term. The product between $tf_d(i)$ and $idf(i)$ is commonly known as the term-frequency inverse document frequency (TF-

IDF) score. Query terms are assigned weights similarly, as the product between a query frequency $qf(i)$ and a query inverse document frequency $qidf(i)$ factors. Most weighting schemes also incorporate normalisation factors that scale term frequencies depending on the total number of terms contained in d or q respectively. When the ranking score is defined as in Equation 2.1, the Euclidean norms of $\vec{q}$ and $\vec{d}$ in the denominator act as length normalisation factors.

The effectiveness of the VSM depends heavily on the selection of a good weighting scheme. A wide range of possible weighting schemes were explored by Salton and Buckley (1988), while Zobel and Moffat (1998) later presented an even more complete survey of existing schemes. When the cosine similarity is used to measure the distance between vectors, a simple and popular weighting scheme is formed by combining $tf_d(i)$ and $idf(i)$, and $qf(i)$ and $qidf(i)$, as shown in Equation 2.2.

$$\begin{aligned} tf_d(i) &= 1 + \log\, tf_i, & idf(i) &= \log \frac{N}{n_i}, \\ qf(i) &= qf_i, & qidf(i) &= 1, \end{aligned} \tag{2.2}$$

where tf_i, and qf_i are the number of times that the i-th term occurs in d and q respectively, N denotes the total number of documents in C, and n_i the number of documents in C containing the i-th term.

The Binary Independence Model (BIM)

The Binary Independence Model (BIM) (Spärck Jones et al., 2000) is an important model based on the Probability Ranking Principle (PRP) (Robertson, 1977), which states that optimal retrieval effectiveness can be obtained if documents are ranked in decreasing order of their probability of relevance based on whatever evidence is available about the information need and document collection.

In this model, every document is assumed to be either relevant (rel) or non-relevant ($\overline{rel}$) to the query. A document is then represented by a vector of binary random variables $\vec{d} = \langle d_1, \ldots, d_M \rangle$, where each component variable d_i can be 1 if the i-th term is present in the document and 0 otherwise. Considering a similar representation for a query $\vec{q} = \langle q_1, \ldots, q_M \rangle$, documents can then be ranked according to their odds of being relevant to q, as shown in Equation 2.3.

$$S_{PRP}(q,d) = \frac{P(rel \mid \vec{d}, \vec{q})}{P(\overline{rel} \mid \vec{d}, \vec{q})} = \frac{P(rel \mid \vec{q})}{P(\overline{rel} \mid \vec{q})} \frac{P(\vec{d} \mid rel\,, \vec{q})}{P(\vec{d} \mid \overline{rel}\,, \vec{q})} \overset{rank}{=} \log \frac{P(\vec{d} \mid rel\,, \vec{q})}{P(\vec{d} \mid \overline{rel}\,, \vec{q})} \tag{2.3}$$

The last equation is obtained by applying Bayes' rule twice, removing the components that only depend on $\vec{q}$, and by applying a log transformation which does not alter the final ranking of documents.

Under the assumptions that terms occur independently and that the probabilities are not affected by terms not present in the query, the ranking function of the BIM can be obtained from Equation 2.3 as shown in Equations 2.4 to 2.6,

$$\begin{aligned} S_{BIM}(q,d) &= \sum_{i \in q,d} \log \frac{P(d_i = 1 \mid rel)\ P(d_i = 0 \mid \overline{rel})}{P(d_i = 1 \mid \overline{rel})\ P(d_i = 0 \mid rel)} && (2.4) \\ &\approx \sum_{i \in q,d} \log \frac{(r_i + 0.5)(N - n_i - R + r_i + 0.5)}{(R - r_i + 0.5)(n_i - r_i + 0.5)} && (2.5) \\ &= \sum_{i \in q,d} w_{RSJ}(i) && (2.6) \end{aligned}$$

where the resulting weight $w_{RSJ}(i)$ is known as the Robertson/Spärck Jones (RSJ) weight. In the equations above the expression $(i \in q, d)$ denotes the set $\{i : q_i = d_i = 1\}$ so that all summations are restricted to terms occurring in both q and d. In addition, R denotes the number of documents in C that are relevant to q, r_i the number of relevant documents containing the term i, while N, n_i are defined as in the description of the VSM.

Because in practice the exact values of R and r_i are unknown, an approximation of the RSJ weight for a term-document pair can be obtained by assuming that $R, r_i \approx 0$. Replacing R and r by 0 in Equation 2.5 results in the BIM ranking function, shown in Equation 2.7,

$$S_{BIM}(q,d) \ \approx \sum_{i \in q,d} \log \frac{N - n_i + 0.5}{n_i + 0.5} \ = \sum_{i \in q,d} cfw(i) \qquad (2.7)$$

defined as the summation of collection frequency weights $cfw(i)$ across the terms occurring in both the query q and the document d.

The 2-Poisson model and Okapi BM25

A popular and effective ranking function within the probabilistic approach is the Okapi BM25 (Robertson et al., 1994; Spärck Jones et al., 2000). This function originates as an approximation of the 2-Poisson model, originally proposed by Harter (1975) and subsequently developed by Robertson et al. (1980), Robertson and Walker (1994), and Robertson et al. (1994). The 2-Poisson model extends the BIM to consider term frequencies within the documents and the query, thus making a distinction between documents containing one from those containing multiple occurrences of a query term.

In the 2-Poisson model, the random variables d_i and q_i are re-defined so that they can take any positive value tf_i in $\mathbb{N}_0$, corresponding to the events of observing tf_i occurrences of the term i in a document or query respectively. Next, all documents containing the term i are assumed to belong to one of two classes: an "elite" class of documents (E_i) which refers to those that are about the topic denoted by the term; and a "non-elite" class $(\overline{E_i})$, in which the term is merely used in passing and whose content is not strictly about the "topic" induced by the term. The distribution of a term's frequency across documents is then modelled as a mixture of two Poisson distributions, each one considering the possibility that the document may belong to the term's elite or non-elite classes. More specifically, d_i is assumed to be distributed under two Poisson distributions: $\mathcal{P}(\lambda_{E_i})$ under the elite set; and $\mathcal{P}(\lambda_{\overline{E_i}})$ under the non-elite set.

Besides these distributional assumptions, further assumptions are made about the associations between term frequencies, eliteness, and relevance. By assuming that term frequencies are related to the documents' relevance throughout eliteness, this set of assumptions expresses that

$$\begin{aligned} P(d_i = tf_i \mid rel) &= P(d_i = tf_i \mid E_i)\, P(E_i \mid rel) \;+\; P(d_i = tf_i \mid \overline{E_i})\, P(\overline{E_i} \mid rel) \\ &= \lambda_{E_i}^{tf_i}\, \frac{e^{-\lambda_{E_i}}}{tf_i!}\, P(E_i \mid rel) \;+\; \lambda_{\overline{E_i}}^{tf_i}\, \frac{e^{-\lambda_{\overline{E_i}}}}{tf_i!}\, P(\overline{E_i} \mid rel). \end{aligned}$$

Under the assumptions that a term occurs more frequently in its elite than non-elite documents ($\lambda_{E_i} > \lambda_{\overline{E_i}}$), plus that the relevance of a document only depends on its elitness condition, the probability of a document d being relevant to a query q in this extended model can be approximated by Equation 2.8.

$$S_{BM}(q,d) = \sum_{i \in q,d} \log \frac{P(d_i = tf_i \mid rel)\, P(d_i = 0 \mid \overline{rel})}{P(d_i = tf_i \mid \overline{rel})\, P(d_i = 0 \mid rel)} \approx \sum_{i \in q,d} \frac{tf_i}{k_1 + tf_i}\; cfw(i) \tag{2.8}$$

Further developments of the previous approximation lead to the well known Okapi BM25 weighting function (Robertson et al., 1994), shown in Equation 2.9, that accounts for the issues of length normalisation and incorporates evidence from within-query term frequencies.

$$S_{BM25}(q,d) \;=\; \sum_{i \in q,d} \frac{(k_1+1)\, tf_i}{tf_i + k_1\,(1 - b + b\, \frac{docl}{avel})}\; \frac{(k_3+1)\, qf_i}{k_3 + qf_i}\; cfw(i) \tag{2.9}$$

In Equation 2.9, *docl* denotes the length of d equal to $\sum_i tf_i$, *avel* denotes the documents' average length in the collection, $0 \leq b \leq 1$ controls the impact of length normalisation, and $k_1, k_3 \geq 0$ control the rate of increase of the term frequency and query frequency factors respectively.

Considered as an isolated function, the within-document term frequency factor in Equation 2.9 is a monotonically increasing function of tf_i that approaches an asymptotic maximum of $k_1 + 1$ as $tf_i \to \infty$. The k_1 parameter influences how fast this function approaches its asymptote with every increase of tf_i. Large values of k_1 signify slower convergence rate w.r.t. tf_i, while small values of k_1 result in faster convergence.

In the BM25 formulation, the length normalisation factor was originally conceived around the scope and verbosity hypotheses. Under the verbosity hypothesis, authors decide to create relatively longer documents because they have the tendency to be verbose and repetitive. In such circumstances, using a large value for b to heavily normalise term frequencies based on document length is appropriate. Alternatively, under the scope hypothesis, documents are relatively long because they cover multiple topics or multiple facets of the same topic. In this latter case, using a small value for b seems more appropriate.

2.1.3 Evaluation of ranked retrieval

This section describes the general evaluation framework that is adopted in IR research to measure and compare the effectiveness of retrieval systems. The initial ideas related to formal evaluation of IR systems were pioneered by Cleverdon, in the context of the Cranfield experiments carried out in the early sixties (Cleverdon, 1962; Cleverdon et al., 1966). In order to enable rigorous, repeatable, and meaningful evaluation of ranked retrieval, the Cranfield methodology proposes to construct a test collection consisting of: a set of documents; a set of queries or topics; a set of relevance judgements, indicating which documents are relevant to each query; and a numeric measure for estimating the quality of a ranked list of documents for a query.

Early IR research focused on small document collections that made exhaustive relevance assessments possible. For instance, the test collection used in the Cranfield's experiments contained 1398 abstracts of scientific articles and a relevance judgement for every query-abstract pair. Since the beginning of the Text REtrieval Conference (TREC)[1], the size of the document collections used in IR research has grown in various orders of magnitude. From about half million documents in the collections used at the TREC ad-hoc tracks, to about 1 billion documents in the more recent ClueWeb12[2] collection used at the TREC Web (Collins-Thompson et al., 2015) track.

Conventionally, the set of queries used for evaluating a retrieval system are generated by potential users of the system or by a group of hired annotators who are preferably knowledgeable of the contents of the document collection. Because the formulated queries are sometimes ambiguous underspecifications of an information need, query creators are commonly asked to provide a more detailed description of their search needs. In TREC parlance, a topic consists in the query text, a query ID, and a narrative field that describes it more fully. The number of topics varies across test collections. Traditionally, TREC collections have contained on the order of 50 topics, which is the minimum number of queries needed for absolute differences in mean average precision (MAP) of 5% be significant across systems (Voorhees and Buckley, 2002). In combination with significance testing, this number can be reduced by half and still be useful for determining significant differences among IR systems (Zobel, 1998).

Relevance judgements for query-document pairs are obtained through manual assessments. The procedure for assessing a pair consists of verifying the extent to which the document is relevant to the information need associated with the query. To facilitate this task, assessors are provided with the narrative description of the information need. Relevance is conventionally given as a binary value (the document is either "relevant" or "not relevant") or in a multi-graded scale of values.

[1] `http://trec.nist.gov`

[2] `http://lemurproject.org/clueweb12/`

Because of the vast size of the document collections currently used in IR research, obtaining relevance judgements for every query-document pair is prohibitive if not impossible. To circumvent this issue, strategies for “pooling” small subsets of documents from the collection to be later assessed for relevance were proposed (Spärck Jones and van Rijsbergen C. J., 1975). In its most basic form, the pooling procedure consists of producing (for each query) multiple ranked lists of documents by using independent IR systems. The union of the top-ranked N results ($N = 100$ in most TREC collections) from each ranked list is then calculated to form the pool of documents which are finally assessed for relevance against the query. Normally, unjudged results that do not form part of the pool for a query are considered non-relevant by most standard evaluation measures.

Since it is unlikely for a set of pooled documents to contain all documents that are relevant to a query, concerns have been raised by researchers about whether existing test collections could be used to evaluate IR systems that did not necessarily participate in the creation of the pool. Fortunately, Zobel (1998) has shown that results based on a limited set of pooled documents can still provide a reliable account of the relative performance that may exist between IR different systems, even for those that did not originally contribute to the pool.

Evaluation measures for ranked retrieval

A popular evaluation measure used to quantify the quality of a ranked list of results when relevance judgements are binary is average precision (AP) (Harman, 1993). AP is based on Precision at rank k ($P(k)$), which measures the proportion of documents retrieved until rank k that are relevant to the query. Formally, for a ranked list of results produced for a query, $P(k)$ is defined as shown in Equation 2.10.

$$P(k) = \frac{1}{k} \sum_{i=1}^{k} r_i \qquad \text{where} \quad r_i = \begin{cases} 1 & \text{if the i-th ranked result is relevant} \\ 0 & \text{otherwise} \end{cases} \tag{2.10}$$

AP is then defined by taking the average across the points in the ranked list at which a relevant document is found, as shown in Equation 2.11,

$$AP = \frac{1}{R} \sum_{k} P(k)\, r_k \tag{2.11}$$

where R denotes the total number of documents that are known relevant to the query. In order to evaluate the performance of a retrieval system across a set of queries, AP is calculated for every query and the resulting scores averaged. The resulting average is referred to as the mean average precision (MAP).

Effectiveness measures such as AP and precision can only be used with binary relevance judgements. However, multiple degrees of relevance need to be considered if the focus of the evaluation is on the ability of a IR system to retrieve highly relevant documents on top of less relevant ones. Various effectiveness measures have been proposed to consider graded relevance judgements. One of these is a simple adaptation of precision at k known as generalised precision ($gP(k)$) (Kekäläinen and Järvelin, 2002), which considers continuous relevance scores $r_k \in [0,1]$. $gP(k)$ is then calculated as $P(k)$ by using these continuous relevance scores. Summing $gP(k)$ across all ranks k at which $r_k > 0$ and then dividing by R results in the generalised average precision gAP measure.

An effectiveness measure more widely used for graded relevance judgements is discounted cumulative gain (DCG) (Järvelin and Kekäläinen, 2002). The DCG at rank n is shown in Equation 2.12,

$$DCG(n) = \sum_{k=1}^{n} \frac{2^{r_k} - 1}{\log(k+1)}, \tag{2.12}$$

where r_k is an integer value representing the discrete grade of relevance of the document retrieved at rank k. In Equation 2.12, the numerator represents the gain associated with the document ranked at position k, while the denominator determines the discounting factor associated with rank k. To make DCG values comparable across different queries, it is common to use the normalised version of DCG (nDCG) which divides Equation 2.12 by the maximum DCG value obtainable for the query, equal to that obtained with an ideal ranking of documents.

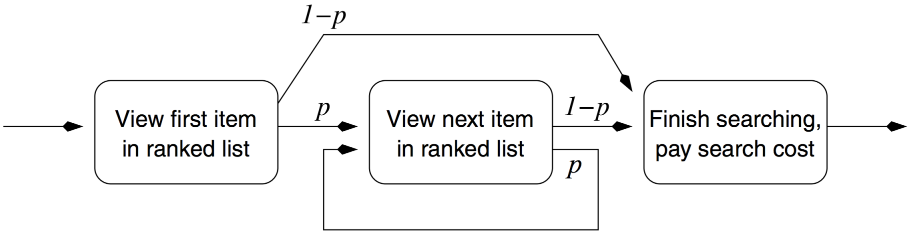
Moffat and Zobel (2008) propose an alternative effectiveness measure called ranked-biased precision (RBP) based on a probabilistic model of user behaviour. Figure 2.1 shows the states and transitions of this model. The user commences by viewing the document ranked at position 1 and then continues scanning the rest of the documents. At every position in the rank, the user can decide to view the next document, with probability p, or to stop its search, with probability $1-p$. Ranked-biased precision can then be written as shown in Equation 2.13,

$$RBP = (1-p) \sum_{k} r_k \, p^{k-1} \tag{2.13}$$

where p is the persistence probability and r_k is defined as in Equation 2.10. It has been shown that for $p = 0.7$ the geometric discounting factor from RBP can closely approximate the probability that a user would click on a document at a certain position in a web search results page (Chapelle et al., 2009).

While MAP, nDCG, and RBP apply a discounting function that only depends on the rank at which a document is located in the result list, the expected reciprocal rank (ERR) measure proposed by Chapelle et al. (2009) discounts according to the relevance of the documents located at previous ranks. In their development of ERR, Chapelle et al. (2009) propose a "cascade" model of user browsing behaviour which accounts for the fact that a user would be less interested in examining a fairly relevant document if it is ranked below

Figure 2.1: Underlying user model proposed in ranked-biased precision. Taken from (Moffat and Zobel, 2008).


1−p
View first item in ranked list
p
View next item in ranked list
1−p
p
Finish searching, pay search cost

a highly relevant one. ERR is then written as shown in Equation 2.14,

$$ERR = \sum_{k=1} \frac{1}{k} \prod_{i=1}^{k-1} (1 - R_i)\, R_k, \tag{2.14}$$

where R_k is the probability that the user is satisfied at rank k. The model induced by ERR assumes that the user continues viewing documents from the ranked list of results until finding a relevant document, at which point the user stops the search.

2.2 Automatic speech recognition (ASR)

In order to estimate the grade of relevance of a spoken document with respect to a natural language query using text retrieval techniques, an SCR system needs to quantify the amount of term overlap that exists between the query and the spoken words. A prerequisite for this is thus the ability to recognise the words spoken in recorded speech. The technology concerned with the problem of identifying all words spoken in a speech utterance is automatic speech recognition (ASR).

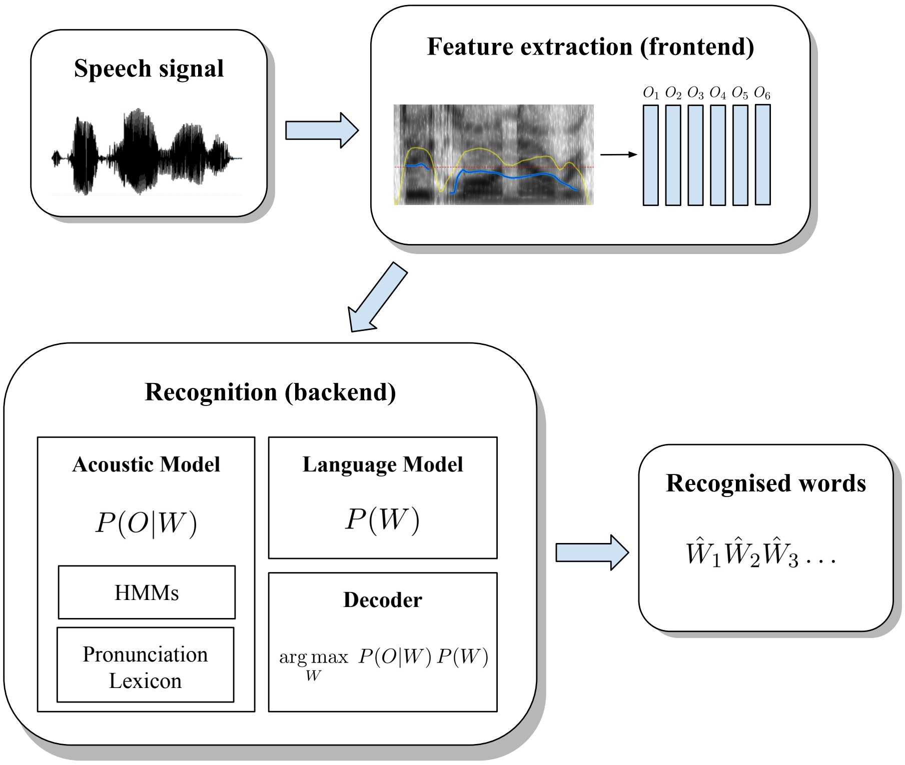
This section overviews the fundamentals of ASR technology. In particular, the section focuses on a specific type of ASR technology, which deals with the recognition of continuous speech as opposed to isolated words, unknown speakers as opposed to speech produced by a single known speaker, and open large vocabularies containing 60,000 distinct words or more. Systems that fall under this category are said to perform large vocabulary continuous speech recognition (LVCSR), and have become the standard ASR technology used in SCR applications.

2.2.1 Overview

The ASR problem is traditionally stated as of finding the most likely sequence of words $\hat{W} = \hat{W}_1 \hat{W}_2 \ldots \hat{W}_N$ spoken in some observed utterance O. More formally, this probabilistic specification of the problem can be written as shown in Equation 2.15,

$$\hat{W} = \underset{W \in \mathcal{L}}{\arg\max}\; P(W|O) \tag{2.15}$$

Figure 2.2: Main components and simplified architecture of a standard ASR system.


Speech signal
Feature extraction (frontend)
O_1 O_2 O_3 O_4 O_5 O_6
Recognition (backend)
Acoustic Model
$P(O|W)$
HMMs
Pronunciation Lexicon
Language Model
$P(W)$
Decoder
$\arg\max_W P(O|W)\,P(W)$
Recognised words
$\hat{W}_1\hat{W}_2\hat{W}_3\ldots$

that is, the problem of finding the word sequence $\hat{W}$ from among all sequences W in a language $\mathcal{L}$ that maximises the probability of W given the acoustic observation O. By applying Bayes' rule, the probability in Equation 2.15 can be broken down as shown in Equation 2.16,

$$\hat{W} = \underset{W\in\mathcal{L}}{\arg\max}\ \frac{P(O|W)\,P(W)}{P(O)} = \underset{W\in\mathcal{L}}{\arg\max}\ P(O|W)\,P(W), \tag{2.16}$$

where the prior $P(O)$ of the acoustic observation can the neglected because it is the same for every W. The rightmost expression in Equation 2.16 suggests that the ASR problem can be disentangled into three sub-tasks: (i) the task of calculating $P(O|W)$ given some acoustic observation O and word sequence W, known as acoustic modelling; (ii) the task of computing $P(W)$, termed as language modelling; and (iii) the task of decoding the word sequence $\hat{W}$ that maximises the product between the acoustic and language models probabilities.

Figure 2.2 shows how these components fit together in the architecture of a typical ASR system. The acoustic model (AM), language model (LM), and decoder components are charged with producing hypothesised word sequences given an acoustic observation. Together, these components comprise the "backend" of the ASR system. In addition to the backend components, an ASR system implements several "frontend" components whose main goal is to transform a speech waveform into a sequence of feature vectors

$O = O_1 O_2 \dots O_T$ upon which recognition is based. These vectors represent how the signal energy varies across its time and frequency dimensions. The following sections describe the individual components of a ASR system in more detail.

2.2.2 Speech units, signal processing, and feature extraction

Speech sounds are fluctuations of air pressure produced by vibrations of the vocal folds, which are excited by an uninterrupted flow of air coming from the lungs. The soundwave produced by these vibrations resonates in the vocal tract and is modified by the position and shape of different articulators, including the lips, jaws, tongue, and nose. Soundwaves are commonly visualised by plotting the change of air pressure over time. The amount of change in air pressure compared to that observed in normal conditions (atmospheric pressure) is the signal's amplitude. Another important characteristic of a speech signal is its frequency, corresponding to the number of times the signal repeats itself per second. Frequency is measured in Hertz (Hz) or cycles per second.

Speech can be digitally recorded by taking voltage samples from a microphone at regular time intervals. The frequency at which such samples are taken determines the maximum signal frequency that can be faithfully represented. This is determined by the Nyquist–Shannon sampling theorem, which indicates that a reliable representation of a signal can be obtained if sampling at twice the rate of the signal's frequency. Because human speech produced lies in lower-frequency bands below 8 KHz, speech recorded at 16 KHz is of sufficient quality for ASR purposes.

The individual speech sounds produced in a specific language can be categorised into a set of sub-word units called phonemes. Phonemes are the basic building blocks of speech. Words are then formed by composing phonemes, which together dictate how each word is pronounced in a particular language or dialect. While phonemes are used to distinguish between words with the same written form but that have different meaning, phones correspond to physical realisations of phonemes as instantiated in a specific speech signal and do not necessarily dictate the meaning of words. Phones and phonemes are commonly represented by symbols from the International Phonetic Alphabet (IPA) (Association, 1999). These sound units can then be seen as intermediate representations between acoustic patterns observed in the speech signal and words from a specific language. Thus, a requirement for solving the speech recognition problem is to find a function that could recognise the individual phonemes being spoken given acoustic patterns observed in the speech signal.

The frontend components of an ASR system are mainly concerned with the preprocessing of speech data prior to recognition. This process commonly involves the application of signal processing and feature extraction techniques over the input speech, with the goal of producing a set of descriptors that can effectively capture the characteristics of the individual phonemes produced by speakers at various points in time. The feature extraction process can be divided into two parts. The first is concerned with slicing the input signal into frames and extracting spectral features for each frame. The second process applies various transformations to these initial features in order to enhance their predictability power.

Spectral features

Spectral features refer to descriptors calculated from the spectrum of the speech signal, which contains information about the signal's amplitudes for different frequency rates at one particular point in time. The spectrum of a discrete-time signal can be obtained by calculating its discrete-time Fourier transform (DTFT) which separates the signal into its frequency components. The fast-Fourier transform (FFT) is an efficient algorithm that calculates the DTFT.

While the spectrum contains information for a single point in time, a spectogram provides a visual representation of the spectrum as it varies through time. Figure 5.1 shows a spectogram for an utterance extracted from a broadcast TV recording. The y-axis represents frequency, while frequency components with high amplitudes (peaks) are represented by darker colours. The reason why spectral features are useful for ASR is that phones can be well characterised by the trajectories of energy peaks and other patterns found in the spectrum. Most notably, vowels can be identified by analysing the location and trajectory of the strongest frequency components in the spectrum, called formants, which roughly correspond to a different resonance in the human vocal tract.

Prior to feature extraction, the individual samples of the speech signal are normally passed through a pre-emphasis filter which dampens low-frequency components in favour of high-frequency ones. The samples are then sliced into a sequence of equally-long overlapping frames of 20-30 ms. The step size or separation size between frames is frequently set to 10 ms to allow for overlapping frames that can capture sudden changes in the signal.

Several types of spectral features and extraction algorithms have been proposed that transform each frame of samples into a feature vector. A classical approach is to use linear-predictive coding (LPC) to characterise the frequency and intensity of a set of formants by regression coefficients (Atal and Hanauer, 1971). The LPC coefficients can then be used to obtain linear-predictive cepstral coefficients (LPCCs) from the signal's cepstrum (Huang et al., 2001). Cepstral features, like LPCC, tend to be more useful for ASR since the cepstrum representation can better discriminate between components related to the excitation of the signal (glottis) and its filters (vocal tract).

Another type of cepstral features widely used in speech recognition are Mel frequency cepstral coefficients (MFCCs) (Davis and Mermelstein, 1980). MFCCs are obtained by first warping the spectrum with a series of triangular bandpass filters, then applying a log transformation to the filters output, and finally taking the first 10-12 coefficients from a discrete cosine transformation (DCT). The set of triangular filters used in the MFCC calculation is known as the Mel scale filter bank and is designed to approximate the non-linear sensitivity of the human ear to different frequency bands.

Feature transformations

The feature processing stage produces a feature vector for each frame in a spoken utterance. ASR systems apply additional transformations to these acoustic vectors to facilitate phone classification. A common approach is to augment the feature vectors with delta and delta-delta coefficients. These are the first and second derivatives of each coefficient with respect to time, normally calculated as differences between successive frames.

It is also common to normalise each vector component based on its mean and standard deviation values by considering all frames available from a single speaker or audio file. This standardisation procedure seeks to cancel out variations across speakers and channels. More sophisticated methods exists for speaker adaptation. For instance, vocal tract length normalisation (VTLN) attempts to balance out differences in the vocal tract shape of male and female speakers (Lee and Rose, 1996), while speaker adaptive training (SAT) and maximum likelihood linear transformations (MLLT) (Gales, 1998) permit speaker-dependent transforms to be learnt and applied iteratively during training.

Prosodic features

In addition to the sets of spectral features described previously, feature vectors may be augmented with acoustic correlates of prosody. The logarithm of the signal energy is one of the conventional acoustic features used in ASR. This acoustic correlate of loudness is useful because it helps to distinguish between voiced and unvoiced sounds, and thus facilitates the distinction between vowels and consonants.

Besides acoustic correlates of loudness, pitch and duration features have also been found useful for various ASR related tasks. Kim and Woodland (2001) demonstrated that these features can be used to recover punctuation symbols in speech transcripts and even provide increased ASR accuracy. Similarly, Liu et al. (2006) describe an ASR system that exploits pitch, duration, and energy cues to improve the detection of sentence boundaries and prediction of filler words and speech disfluencies. Other research that suggests that prosodic cues can be directly used to reduce speech recognition errors include (Chen et al., 2006; Jeon et al., 2011; Chen et al., 2012).

2.2.3 Language and acoustic modelling

For each utterance, the frontend components produce a sequence of observations $O = O_1 \ldots O_T$, each describing the spectral characteristics of a particular frame. This sequence is subsequently received by the backend components of the ASR which search for the most likely sequence of words $\hat{W} = \hat{W}_1 \ldots \hat{W}_N$ that may explain these observations. Two important components used for this purpose are the language model (LM) and the acoustic model (AM).

The language model (LM)

Within the space of all word sequences that could possibly be generated by randomly appending individual words from a language, only a small subset of them will be grammatical, and only a small proportion of these will be meaningful and frequently used in spoken language. ASR systems can then take advantage of the fact that some word combinations are more frequent than others to limit the search space of possible word sequences in the search for the optimal $\hat{W}$.

A language model (LM) assigns a probability to a sequence of words $P(W_1 \ldots W_N)$. A good LM assigns higher probabilities to sequences that are highly used in a language, and low ones to sequences that are less frequently used. The most common type of LMs used in ASR are the so-called n-gram language models, in which the probability of the occurrence of the next word in a sequence is based on the $n-1$ words that occur before it. That is

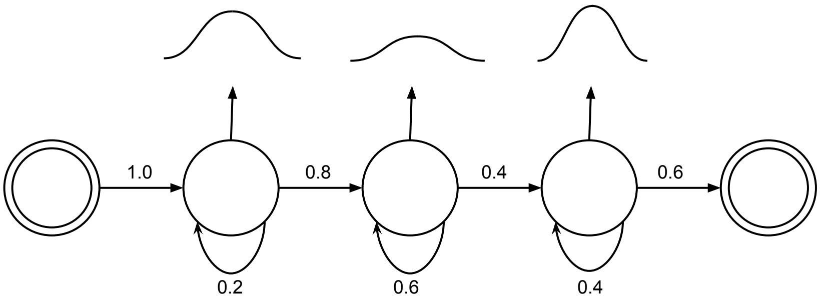
$$P(W_1 \ldots W_N) = \prod_{i=1}^{N} P(W_i \,|\, W_{i-n+1} \ldots W_{i-1})$$

These conditional probabilities are usually estimated by counting the number of occurrences of n-grams in a large corpus of text.

In practice, estimating n-gram probabilities based on observed counts has the issue that a large number of valid n-grams in a language may not appear at all in the training data and would be thus assigned a probability of 0. This is a potential problem for ASR since, given Equation 2.16, sequences with 0 LM probability would never be recognised even if they obtain high acoustic probability. Several smoothing techniques have been proposed to tackle this problem, most of which introduce adjustments to the occurrence counts of rare n-grams so that they acquire non-zero probabilities. The most simple technique is additive smoothing (Chen and Goodman, 1996) which adds a fixed pseudo-count to non-occurring n-grams. More sophisticated techniques include Jelinek-Mercer (Jelinek and Mercer, 1980), Katz (Katz, 1987), and Kneser-Ney (Kneser and Ney, 1995) smoothing, which use a weighted linear interpolation of decreasingly lower order models (back-off) for improving the estimation of rare n-grams.

Modern ASR systems use a combination of n-gram LMs and neural language models (NLMs), also known as continuous-space language models (Mulder et al., 2015). NLMs are based on artificial neural networks, more specifically, on deep neural networks (DNNs), including feed-forward neural networks (FNNs) (Bengio et al., 2003) and recurrent neural networks (RNNs) (Mikolov et al., 2010), trained using the back-propagation algorithm to predict the identity of the n-th word in a sentence given its preceding words. Much of the success of these approaches lies in their use of continuous vector representations of words, commonly known as word embeddings (Mikolov et al., 2013). In this respect, neural network approaches have the capability to map words from the vocabulary onto a latent vector space so that words used in similar contexts are clustered. This helps to alleviate the data sparsity problem as the predictions of the model are implicitly based

Figure 2.3: Graphical representation of a hidden Markov model used to model an individual phone.



on such word clusters. In addition, RNNs can in theory handle arbitrary context lengths, thus unlike n-gram models, they need not be designed for a fixed number of preceding words.

The acoustic model (AM)

The main goal of acoustic modelling is to estimate $P(O|W)$ for a given sequence of acoustic observations O and words W. The traditional approach to calculate these probabilities makes use of hidden Markov models (HMMs) (Levinson et al., 1983; Rabiner, 1989). An HMM models a process that produces sequences of symbols probabilistically. An HMM has a set of hidden states $Q = \{q_1, \ldots, q_N\}$, special starting and ending states, and transition probabilities a_{ij} between each pair of states. Some of the states in an HMM are regarded as emitting states from which the model can produce an observed value. Each emitting state q_i defines a probability distribution $b_i(O_t)$ over some set of possible observation values O_t.

Figure 2.3 shows a left-to-right HMM with five states. The generation process begins at the left-most state of the diagram. At each step, the model decides to transition to its right state with some probability or to remain in its current state. While visiting an emitting state, the model produces an observation based on a probability distribution, depicted in the figure as an arrow pointing to a density function. This particular HMM structure with three emitting states and left-to-right transitions is typically used in ASR systems to model individual phones. The states in the HMM represent some intermediate step in a phone's pronunciation, while the self-transitions (loops) model duration variations. In order to consider variations produced by preceding and following phones, context dependent systems represent a single phone with three HMM states concatenated in sequence. Further, it is common to append the HMM structures of various phones together into sets of triphones to account for acoustic variability. This appending process can be used to produce word level HMMs based on a pronunciation dictionary or lexicon which contain transcriptions of word strings into phonemes.

Traditional acoustic modelling techniques use Gaussian mixture models (GMMs) to model emission probabilities over continuous acoustic vectors $O \in \mathbb{R}^N$. Under this approach, the output probability distribution for a state $b_i(O)$ based on a GMM with K

components is given by

$$b_i(O) = \sum_{k=1}^{K} \phi_k \mathcal{N}(O, \mu_{ik}, \sigma_{ik}),$$

where $\mathcal{N}(O, \mu_{ik}, \sigma_{ik})$ is the probability density function of the k-th multivariate Gaussian component in the mixture. More recently, DNNs have been used instead of GMMs for estimating emission probabilities (Hinton et al., 2012; Deng et al., 2014; Yu and Deng, 2014). The superiority of DNNs over GMMs for phone recognition can be attributed to their ability to discover useful features from more primitive spectral descriptors than MFCCs, their high robustness to small noise perturbations in the inputs, and their effective exploitation of contextual input features.

The use of GMMs or DNNs to model emission probabilities in combination with HMMs for acoustic modelling is regarded as the hybrid GMM-HMM or DNN-HMM frameworks. In these frameworks, the likelihood of an acoustic observation $O_1 \dots O_T$ given an HMM, $\mathcal{M}$, is given by Equation 2.17,

$$P(O|\mathcal{M}) = \sum_{S} \prod_{t}^{T} b_{s(t)}(O_t)\; a_{s(t)\, s(t+1)} \tag{2.17}$$

where the summation ranges over all possible sequences of states $S = s(1) \dots s(T)$ in the model. The process of recognising the most likely sequence of phones spoken in a given utterance O then consists of finding a HMM model $\hat{\mathcal{M}}$ that maximises the likelihood from Equation 2.17, corresponding to the model that best explains the acoustic observations.

2.2.4 Decoding, output representation, and evaluation

Decoding refers to the process of finding the most likely sequence of words W that maximises the product between the acoustic likelihood $P(O|W)$ and language model probability $P(W)$. The traditional decoding algorithm used for this purpose is the Viterbi algorithm (Viterbi, 1967), which uses dynamic programming to efficiently infer the most likely state sequence from all word HMMs that best match the given observations. In practice, implementations of this algorithm perform some type of pruning mechanism to reduce the size of the search space by discarding state paths with low probabilities. A common pruning technique is beam search in which only the top K scoring paths (hypotheses) are kept while advancing the search from one time step to the next.

The optimal word sequence found by Viterbi is usually termed a 1-best hypothesis. For many applications however, it is more convenient to consider more than one recognition hypothesis. Several decoding algorithms have been developed for this purpose, most of which extend Viterbi to generate the top N-best recognition hypothesis besides the 1-best (Schwartz and Austin, 1991; Soong and Huang, 1991). Considering alternative hypothesis permits the application of increasingly complex models to iteratively refine the ASR output in a process called multi-pass decoding. For instance, it is common to perform

Figure 2.4: An example of a recognition lattice (without confidence scores) taken from (Larson and Jones, 2012b).

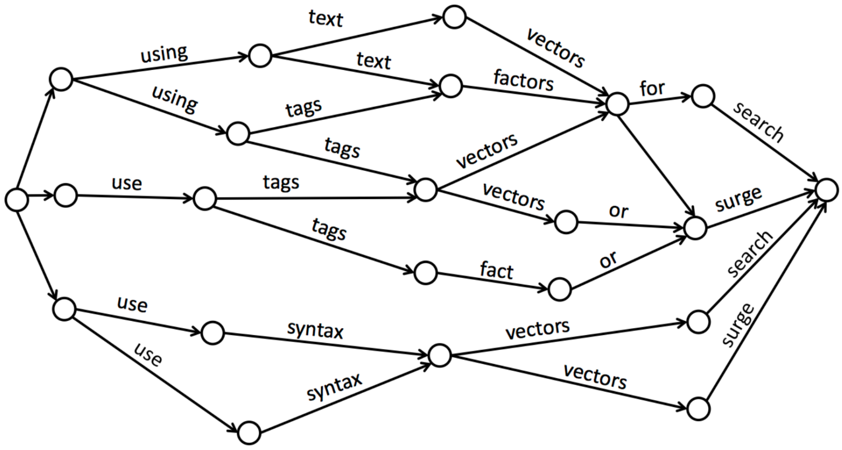
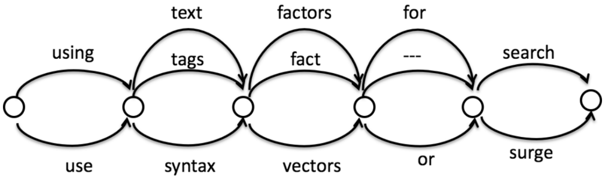


Figure 2.5: An example of a word confusion network (without confidence scores) taken from (Larson and Jones, 2012b).



a first-pass decoding with a bigram LM to obtain a list of N-best hypothesis and use a trigram LM to re-score the hypotheses in a second-pass.

An alternative representation of the most likely recognition hypotheses is a lattice. An example is shown in in Figure 2.4. A lattice is a weighted directed acyclic graph that encodes alternative recognition hypotheses. Each complete path through a lattice represents an alternative hypothesis weighted by its recognition score. The nodes in a lattice represent points in time and the arcs represent hypothesised words or other recognition units like phones or HMM states. Arcs are also labelled with a score that represents the confidence level of the ASR about the recognition of a particular word.

Word confusion networks (WCNs) provide yet another compact representation of recognition hypotheses (Mangu et al., 1999). An example of WCN is shown in Figure 2.5. A WCN is a conflated version of a word lattice in which exact time information is discarded in favour of providing more direct information about the relative position of each word in the recognised sentence along with its set of competing words.

Transcription errors produced by an ASR system at the word level can be classified into in-vocabulary and out-of-vocabulary (OOV) errors. The first type occur for words that despite being included in the LM of the ASR are not recognised correctly. The second type occurs when the words to be recognised are not included in the LM of the ASR, and thus have 0 probability of being recognised. Word error rate (WER) is the main evaluation measure used to estimate the quality of an ASR 1-best hypothesis against a reference (perfect) transcription. WER is calculated by first aligning the hypothesis (hyp) with the perfect transcript of the utterance (ref), and then counting the number of insertions (I), substitutions (S), and deletions (D) errors in hyp relative to ref. The alignment between hypothesis and reference transcripts is done so that the number of errors is minimal. WER is then defined as the ratio between the sum of these errors and the total number of words in the reference transcription, as shown in Equation 2.18.

$$WER = \frac{I + S + D}{|ref|} 100 \tag{2.18}$$

Speech recognition accuracy is known to vary greatly across tasks and, in particular, across domains, genres, languages, and speech types. Recognition accuracy can also vary depending on the amount of data available for training as well as the amount of computational resources available for training and decoding. Typical averaged WER values reported in the literature for various tasks are: read speech (3-5%), broadcast news (9-11%) (Bell et al., 2015; Wu et al., 2016), multi-genre TV broadcasts (10-40%) (Bell et al., 2015), conversational telephone speech (5-40%) (Lileikyte et al., 2015; Xiong et al., 2016; Chiu et al., 2017; Enarvi et al., 2017), lectures/public talks (17%-30) (Rousseau et al., 2012; Akiba et al., 2016), YouTube (45-50%) (Hinton et al., 2012). Although recent advances in DNN-based modelling have significantly reduced error rates across a wide range of tasks (Hinton et al., 2012), speech recognition remains difficult in situations where there is insufficient training data for a particular task or language, or when there are substantial differences between the data used for training and evaluation.

2.3 Content structuring

In the context of this thesis, content structuring is concerned with the problem of identifying coherent units of information in text or spoken documents which could represent a good target for retrieval. The overall objective of a content structuring method is then to find structural components within multi-topical unstructured documents so that each component found is aligned with a single concept, idea, or topic which could potentially serve to satisfy a single information need from the user.

The main motivation behind structuring the content of a search collection is to enable the retrieval of smaller retrieval units, and with that, to reduce the amount of effort a user has to invest in order to consume the information of interest. By retrieving focused, smaller, units of relevant content or by pointing out to the user where such relevant content begins within the original document, the hope is that the user will save valuable time that they would otherwise have to spend skimming or navigating the document in order to locate the relevant information.

Because textual content can be skimmed and browsed more easily than speech, content structuring techniques present more potential benefits for SCR applications, where the time a user needs to audition a search result is not negligible. Most content structuring methods used in SCR can be classified into two broad categories: automatic text or topic segmentation methods, originally designed for the processing of text documents; and spoken document segmentation methods, which besides text transcripts can make use of other structural cues that are prominent in speech.

This section provides a detailed overview of existing content structuring methods. Emphasis is given to methods that have been used in passage retrieval and SCR research. This includes a large number of approaches originally designed for the automatic segmentation of text material.

2.3.1 Automatic segmentation of text documents

Most automatic segmentation methods proposed in the literature attempt to measure the degree of lexical cohesion that exists between adjacent elements in a piece of text. The standard method for measuring lexical cohesion consists of quantifying the amount of term overlap that exists between two or more contiguous elements. The higher the lexical overlap between the elements, the higher their assumed degree of cohesion. These estimates of lexical cohesion are then used to make decisions about whether contiguous elements in a document should be treated as separate segments or merged into a single one so as to maximise the inter-segment cohesion.

Among the various segmentation methods proposed in the literature, the remainder of this section describes those that have been influential in subsequent work, and have been widely used in IR and SCR research.

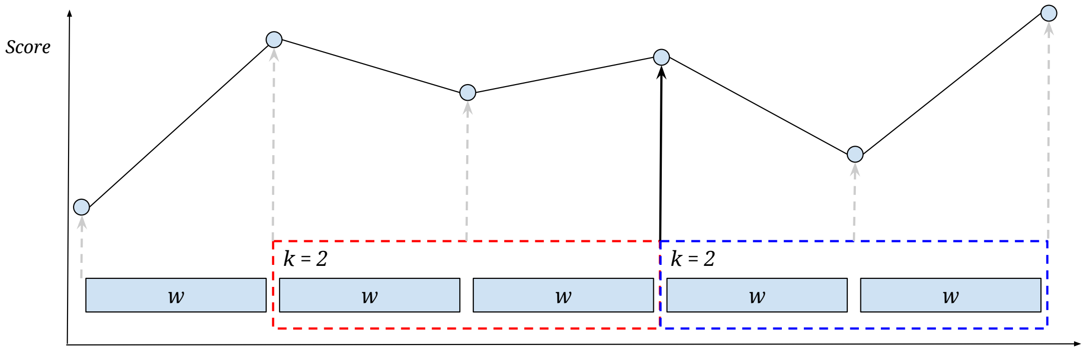
Sliding windows

The most trivial approach to text segmentation is to divide a document into arbitrary passages of equal length. A common approach to do this consists of sliding a window of length L, measured in words, over the text document, one word at a time, and extract a segment or passage every time the window has been shifted by S steps. The L parameter then determines the length or size of the segments to extract, while S determines the amount of overlap between adjacent segments. Thus, under these definitions, setting $S = L$ would result in non-overlapping passages being created, whilst for $S = L/2$ there would be 50% overlap between consecutive passages.

TextTiling (TT)

TextTiling performs segmentation of text documents by identifying strong changes in vocabulary usage between adjacent fragments of text (Hearst, 1993, 1994, 1997). The algorithm can be broken down into three processing steps. The first step consists of tokenising the input text, followed by lowercasing, removal of stop words, and stemming. The second step consists of computing a similarity score, called the lexical score, between

Figure 2.6: The TextTiling algorithm applied to blocks of $k = 2$ pseudo-sentences of size w. A depth score is calculated at the boundaries of each pseudo-sentence based on the similarity between adjacent blocks (red and blue dashed boxes).



adjacent pairs of text blocks. The last step of the algorithm selects the most promising segment boundaries in the document by identifying pairs of blocks with minimal lexical scores.

In the lexical score computation step, blocks are formed by grouping k adjacent pseudo-sentences, which are in turn formed by sequences of w consecutive terms. A sliding window is then passed over the pseudo-sentences of the document to create the blocks, and a lexical score is computed at each pseudo-sentence boundary. Hearst proposed two methods for calculating the lexical score at a boundary: block similarity and vocabulary introduction. In block similarity scoring, blocks are represented by vectors of term frequencies and the lexical score between two blocks is calculated as the cosine distance between their vectors. In the vocabulary introduction method, the lexical score between a pair of adjacent blocks is given by the number of terms contained in the blocks that are seen for the first time in the text. The intuition is that blocks that introduce new vocabulary are more likely to signal the beginning of a new topic.

The last step of the algorithm is to identify block pairs showing low lexical scores corresponding to the most likely topical boundaries. This problem can be seen as that of identifying the deepest valleys in the lexical score contour. Instead of just selecting the valleys with the lowest absolute scores, a "depth" score for each valley is calculated as the sum of relative differences between the lexical score of the valley and that of its left and right peaks. Boundaries are then ranked by their depth scores and the lowest ones returned as output. Figure 2.6 shows how TextTiling is applied to a sequence of pseudo-sentences. The number of desired boundaries can be automatically determined by selecting valleys whose depth scores surpass a specified threshold. Instead of using arbitrary thresholds, Hearst proposed to calculate per document thresholds based on the average (μ) and standard deviations (σ) of depth scores.

Feature-based approaches

A number of suggested approaches to text segmentation make use of statistical models that can learn how to best combine a set of features to predict where topic boundaries may occur given some training examples of boundaries in text data (Beeferman et al., 1997; Reynar, 1998). The basic approach consists of extracting features that may be indicative of the presence of topic boundaries, such as cue phrases, lexical cohesion scores like those computed by the TextTiling algorithm, lexical chains based on named entities or word synonyms extracted from a thesaurus, location of the previous predicted boundary, etc. A machine learning model is then trained to learn associations between these features given examples of true and false topic boundaries. This model can later be used to predict the probability that a topical break exists at a particular location within a given document.

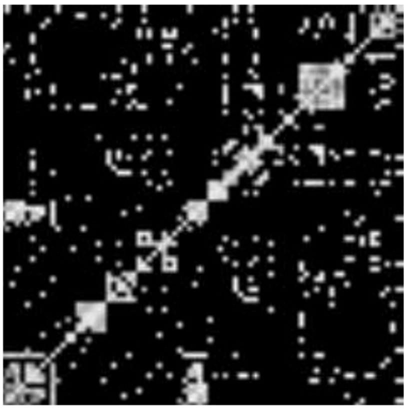
The technique described by Beeferman et al. (1997) trains a log-linear model with a set of lexical and visual features for segmenting a TV broadcast news video. Two important features used in their approach are given by a long-range language model, trained on selected words from the previous N sentences, and a short-range tri-gram language model, which only conditions its predictions on the previous two words in a sentence. At points in a document where a new topic is introduced, the changes in vocabulary cause the predictions from the short-range LM to be better than those from the long-range LM. Therefore, hypothetical topic boundaries can be predicted by comparing the performance between these two LMs.

DotPlot and C99

The segmentation approach proposed by Reynar (1998) calculates lexical similarity scores between every pair of text-blocks in a document. These values are then visually depicted in a dotplot, a 2D matrix showing the similarity scores of each pair of blocks (i, j) where high similarity scores between pairs are denoted by using a brighter colour. Figure 2.7 shows an example of a dotplot. Regions that show high cohesion are visible in the dotplot as small bright squares along the diagonal. Thus, the segmentation problem can be framed as one of detecting these type of patterns in a dotplot. This can in turn be seen as an optimisation problem, where the goal is to find a set of "splits" along the diagonal that maximise the intra-segment similarity or inter-segment dissimilarity.

Choi (2000) proposed yet another influential segmentation algorithm based on lexical cohesion called C99. In this work, Choi highlighted the fact that absolute cosine distances between block pairs are often unreliable for short blocks of text, and that only relative similarity differences can be considered meaningful. Based on this observation, C99 transforms a cosine similarity matrix by converting each of its values (i, j) into an integer which specifies the position at which (i, j) would rank if compared to its K closest neighbours. After the rank-similarity matrix is obtained, C99 performs divisive clustering. At each iteration, the algorithm selects the split which maximises a global intra-density criterion,

Figure 2.7: Example of a dotplot extracted from (Choi, 2000) showing the pairwise similarity matrix between blocks of text.



calculated for all of the segments in a segmentation plus the new segments that result from applying the split step. For a segment, this criterion is calculated as the ratio between the sum of ranks of the segment and its area in the rank-similarity matrix. The iterative clustering procedure continues until the global intra-density measure stabilises.

Utiyama and Isahara (UI)

The probabilistic approach proposed by Utiyama and Isahara (2001) attempts to find the segmentation that attains maximum probability given a text document. More formally, given a sequence of words $W = w_1, \ldots, w_n$ comprising the document, the goal is to find the most likely segmentation $S = S_1, \ldots, S_m$ that satisfies

$$\arg\max_S P(S|W) = \arg\max_S \; P(W|S)\, P(S).$$

In this equation, the likelihood $P(W|S)$ is approximated by $\prod_i \prod_j P(w^i_j|S_i)$, where S_i is a segment containing a subsequence of n_i consecutive words from W, i.e. $S_i = w^i_1, \ldots, w^i_{n_i}$. The individual probabilities of a word being generated by a segment $P(w^i_j|S_i)$ can be obtained by estimating a language model for each individual segment S_i. In the absence of any prior information about S, the authors suggest defining $P(S)$ as being proportional to n^{-m}, where n is the length of the document and m is the number of segments in S.

Under the above set-up, the likelihood $P(W|S)$ will be maximised when the segmentation S is constructed in such a way that a large number of terms of the same type are included in a single segment, which will occur when words are grouped into a small number of segments. This criteria goes against the prior probability objective, which is maximised when there are a large number of segments. Given these optimisation targets, Utiyama and Isahara cast the optimisation problem as the problem of finding the optimal path in a directed weighted graph, where nodes in this graph represent possible splitting points between words, and edges represent individual segments covering all words inbetween the connected nodes.

Minimum Cut (MC)

Malioutov and Barzilay (2006) developed the Minimum Cut model for the segmentation of spoken lectures. A text document is represented as a undirected acyclical weighted graph, where the nodes in the graph correspond to atomic text blocks, and edges represent the similarity between a pair of blocks. The segmentation problem is then cast as a graph-partitioning problem, where the objective is to find a partition of the document graph which minimises the normalised-cut criterion, an objective function found useful for image segmentation tasks (Shi and Malik, 2000). This optimisation objective seeks to capture the within partition similarity of a candidate partition as well as the dissimilarity across different partitions.

The authors evaluated the Minimum Cut algorithm on a collection of ASR transcripts and observed that their method tended to perform more robustly than others in the presence of ASR errors. Despite this advantage, a major drawback of Minimum Cut is that the number of partitions produced is not automatically determined by the algorithm and instead needs to be provided in advance.

Bayesian segmentation (BayesSeg)

In follow-up work, Eisenstein and Barzilay (2008) developed a more general Bayesian framework for the definition of text segmentation algorithms and demonstrated that Utiyama and Isahara's method is a particular case of this framework. Besides using language models for estimating the probabilities of a segmentation based on word counts, Eisenstein and Barzilay (2008) proposed using an additional language model to account for cue phrases, which they found useful for the segmentation of transcribed meetings and a medical textbook. This new algorithm, called BayesSeg, was shown to outperform UI and Minimum Cut in terms of segmentation quality for both transcribed speech and written documents. Despite this increased performance, BayesSeg assumes that the number of topics, and therefore the number of desired segments, is given in advance.

2.3.2 Segmentation of spoken content

The most common approach to finding topic boundaries in speech consists of running an automatic text segmentation algorithm over the transcripts generated by an ASR system. Compared to the segmentation of text documents, performing topic segmentation over noisy speech transcripts is arguably a more difficult task. In addition to transcription errors and the lack of punctuation symbols, spoken language tends to be more informally structured than written language. Spoken language tends to show smoother topic transitions and a general reduction in the usage of content bearing words which make the segmentation task more difficult.

In spite of its increased difficulty, spoken content contains additional information that can potentially be helpful for the segmentation process. Prosodic features derived from the speech signal, that capture variations in pitch, loudness, speech rate, as well as duration of pauses between words and utterance-ending syllables, have been shown to correlate well with the occurrence of topical boundaries (Hirschberg and Grosz, 1992; Hirschberg, 2002; Wagner and Watson, 2010). Previous work on topical segmentation of speech has successfully incorporated many of these prosodic features into segmentation approaches. This work typically relies on supervised machine learning techniques to combine prosodic/acoustic features with lexical cues.

The seminal work on exploiting prosodic information for topic segmentation is that of Shriberg et al. (2000) and Tür et al. (2001). In this work, prosodic features were extracted around each word boundary in the ASR transcripts with a window that included the preceding and following words around each boundary. This set of features included pause durations, phone durations, and various hand-crafted pitch and voice quality descriptors. A decision tree classifier was then trained on this set of features to estimate the probability of a topic break occurring at an inter-word boundary. The probabilities estimated by this classification tree were then combined under a HMM framework with a topic segmenter, independently trained to predict topic assignments from lexical information. Shriberg et al. experimented with this model on a sentence and topic segmentation tasks and found it to perform substantially better than a model that did not make use of prosodic information.

Subsequent work on prosodic-based speech segmentation include that of Kolář et al. (2006), who observed that besides pitch and pause features, energy features can also be beneficial for speech segmentation. Also, Malioutov et al. (2007) devised an unsupervised approach that detects putative topic boundaries in a spoken document without requiring any lexical information. This approach attempts to approximate the lexical cohesion score that a pair of utterances would attain based on their acoustic similarity. These similarity scores are then used along with Minimum Cut to segment the spoken document.

In addition to prosodic information, which is always present in spoken language, there are other domain specific cues which can be used to enhance the quality of topic segmentation algorithms. In the broadcast news domain, for example, topics generally correspond to news stories. Frequently, news stories are interleaved with commercials which can be automatically detected by looking for significant changes in energy levels. Other important features for the identification of story boundaries in broadcast news are cue words/phrases such as "Good morning" and "reporting from ...". Additionally, if the content is known to be produced by multiple speakers, speaker turns are another feature that are often indicative of topic shifts. Since most ASR systems perform a fine-grained segmentation of the input speech based on voice activity recognition and speaker diarisation methods, the outputs of these components are frequently available and can be used as additional features. Finally, if video information is also available, segmentation algorithms can make use of visual structuring cues to guide the identification of topic boundaries. Common visual features include the beginning of shots, produced by a single camera, and scenes, corresponding to groups of visually similar shots.

2.4 The application of content structuring methods to text retrieval

Knowing the topical structure of a document can be beneficial for a number of text retrieval tasks, including document, passage, and XML retrieval. This section describes prior work that has made use of segmentation methods to improve text retrieval techniques. A more detailed review of previous studies that have applied segmentation methods to SCR tasks is given later in Chapter 3.

2.4.1 Document retrieval

Early attempts to exploiting sub-document structure in IR focused principally on improving the effectiveness of full-document retrieval techniques. The standard technique adopted for doing this among researchers is comprised of three basic steps: (i) segment each document in the collection into short passages; (ii) calculate a relevance score for each passage against the query; (iii) rank the documents based on a combination of passage scores. Researchers have explored different segmentation algorithms and strategies to produce the document scores from various combinations of passage scores.

An early application of sliding windows is mentioned in (Stanfill and Waltz, 1992) as part of a description of the now extinct CMDRS retrieval system. In this system, documents were split into non-overlapping contiguous passages of 30 words each. When a query was issued, the system would score each passage for each document in the collection and rank the documents based on the score of their highest scoring passage. Stanfill and Waltz motivated this passage-level approach by stating that: (a) it facilitated the retrieval of very long documents; (b) it provided a better normalisation mechanism for collections that contained examples of both extremely long and short documents. Standard IR models would usually assign low scores to long documents that contain a relative small relevant part. By scoring passages, retrieval models can be made more sensitive to short sections containing a high density of query terms and thus improve the retrievability of long documents. In order to avoid splitting a high scoring document region in half, Stanfill and Waltz applied passage "blurring", by combining adjacent passages into a longer overlapping passage.

During this time, several researchers highlighted the benefits of considering passage-level evidence for improving full-document retrieval (Hearst and Plaunt, 1993; Salton et al., 1993; Callan, 1994). Notably, Hearst and Plaunt (1993) experimented with the

TextTiling algorithm for dividing documents into multi-paragraph passages. For retrieval, documents were ranked based on the sum of scores of the top 200 passages retrieved in an initial retrieval pass. Their experiments showed that passages generated by TextTiling were not more effective than those from paragraphs. In both cases, passage-based retrieval provided better document rankings than if performing full-document retrieval alone.

Work by Callan (1994) evaluated document scoring techniques based on passages generated from paragraphs and fixed-length overlapping windows. He evaluated three retrieval approaches that varied depending on which source of evidence was used for estimating the relevance score of a document: (i) evidence from the document only; (ii) evidence from its best scoring passage; (iii) evidence from both the document and its best passage. Conclusions from this work indicated that paragraphs perform poorly compared to sliding windows, mainly because the former do not always align well with the boundaries of relevant sections. Instead, the overlapping windows approach provides an extra degree of flexibility and adapts better to different relevant regions with arbitrary starting points. The author also observed that exploiting both document and passage level information in combination (iii) resulted in improved search effectiveness compared to using document or passage evidences alone.

Subsequent research investigated optimal strategies for the combination of multiple sources of evidence from independent searches (Bartell et al., 1994; Fox and Shaw, 1993; Belkin et al., 1995). Bartell et al. (1994) proposed learning optimal weights for a linear combination of relevance scores by using a gradient-based optimisation approach. Fox and Shaw (1993) and Belkin et al. (1995) investigated the performance of simple aggregations of relevance scores from multiple ranked lists, consisting of adding the different scores (CombSUM), dividing or multiplying the sum of scores by the number of ranked lists in which a document appears (CombANZ and CombMNZ), and taking the maximum or minimum values across rankings (CombMAX and CombMIN). Among these strategies, experimental results showed that the summation of scores (CombSUM) was the most effective at combining evidence from multiple rankings (Belkin et al., 1995).

The seminal work on using content structuring techniques for improving document retrieval is that of Kaszkiel and Zobel (1997, 2001), who performed an in-depth comparison of existing segmentation methods proposed at the time for scoring documents. The effectiveness of a segmentation method was based on its ability to produce passages that could serve to rank documents effectively, where documents were ranked according to their highest scoring passages. The set of methods compared included: discourse segments such as those from a book's paragraphs, sections, and pages; segments produced by TextTiling; fixed-length non-overlapping windows; and fixed-length and variable-length arbitrary passages. The latter two types correspond, respectively, to passages of fixed or any length that could start at any word position within a document.

The experiments conducted by Kaszkiel and Zobel indicated that variable-length arbitrary passages performed best among all passage types considered, although only by a small fraction over fixed-length arbitrary passages. Both variable-length and fixed-length arbitrary passages were shown to outperform other types of non-overlapping pre-defined passages and to enhance the quality of the document rankings overall. Depending on the length chosen, the performance of fixed-length arbitrary passages varied widely across test collections, indicating that there is not a single optimal passage length that could "fit" every possible query and collection. In fact, an oracle approach that selected the best passage length per query was shown to perform significantly better than the rest of approaches under study. This motivated the authors to conclude that, despite providing better results than a fixed-length approach, their variable-length strategy failed at finding the optimal passage for every query. Another important result from this work is the observation that the application of length normalisation mechanisms to adjust the relevance scores can considerably improve document retrieval effectiveness when passages vary greatly in length.

Subsequent research in this area focused on exploiting sub-document structure for ranking passages, instead of documents, and for improving the ranking of semi-structured documents specified in extensible mark-up language (XML).

2.4.2 Passage retrieval

Passage retrieval refers to the task of finding the portions of documents that are relevant to a query. Due to the high costs associated with the collection of relevance judgements for arbitrary text fragments, rigorous evaluation of passage retrieval techniques did not commence before shared-tasks and benchmarking initiatives, such as those organised by the Text REtrieval Conference (TREC), provided a test collection with passage-level relevance judgements. In particular, much research in passage retrieval was done in the context of the TREC question answering (QA), the high-accuracy retrieval from documents (HARD), and the spoken document retrieval (SDR) tracks (Voorhees, 2001; Allan, 2003; Voorhees and Harman, 2005). The TREC HARD track posed a passage retrieval task, where systems were evaluated in terms of their ability to rank passages with relevant content at high ranks.

Most existing approaches to passage retrieval have been based on techniques previously shown to be effective for full-document retrieval using sub-document structure. The most effective approaches usually rely on fixed-length overlapping sliding windows to define the passages to be retrieved, and apply additional post-processing techniques to either improve the quality of the initial ranking of passages or to adjust the passage boundaries. The approach described in (Huang et al., 2004) first ranked non-overlapping passages and then combined adjacent highly scoring passages from the same document into a single passage. After merging, the scores of the passages were also updated by summing the scores of the merged passages with that of their document. In general, combining document and passage level evidence has generally been found to improve passage retrieval effectiveness (Huang et al., 2004; Abdul-Jaleel et al., 2004).

While most approaches proposed for document retrieval perform fixed-length segmentation at indexing-time, some researchers explored the idea of forming retrieval units of variable-length dynamically, at querying-time, to take advantage of the extra information from the query. One of such approaches was implemented in the MultiText system (Clarke et al., 2000a,b), which detected passages dynamically by identifying the shortest word-sequences in a document containing all, or a subset, of terms from the query. Documents were then scored based on the length and number of distinct sub-sequences found in the documents.

Another query-dependent approach was proposed by (Mittendorf and Schäuble, 1994) for document and passage retrieval. In this approach, documents are assumed to be produced by two HMMs: one that emits words that are relevant to the query; and another one which generates words that are unrelated to the query. Documents can then be ranked by their odds of being generated by the "relevant" HMM relative to the "background" HMM. For passage retrieval, Mittendorf and Schäuble (1994) considered a sequential model resulting from the concatenation of a relevant HMM in the middle of two background HMMs. Relevant passages can then be identified by detecting fragments that are likely to be generated by the relevant state and whose neighbouring words have high probability of being generated by the background states of the HMM.

Jiang and Zhai (2004, 2006) built upon Mittendorf and Schäuble's work and experimented with improved HMM structures and with language models to estimate word-emission probabilities. Their experiments on the TREC HARD tracks showed that an HMM-based approach was effective at refining the boundaries of an initial list of pre-segmented passages and found that these adjusted boundaries correlated better with those determined by the true relevant passages.

Yet another technique for constructing variable length passages at retrieval time was investigated by Abdul-Jaleel et al. (2004), based on the locality-similarity approach previously proposed by de Kretser and Moffat (1999). In the locality-based approach, individual occurrences of query terms appearing in a document are scored according to their query and inverse document frequencies. Each term in a document is then assumed to affect the scores of its neighbouring terms falling within a pre-defined region of influence. Passages can then be determined by identifying high scoring regions of influence containing a high density of query terms. In the work of Abdul-Jaleel et al., every region of influence was treated as a possible passage to be retrieved for a query. Despite its ability to find variable length passages dynamically, this technique did not perform better than using fixed-length overlapping passages with a standard retrieval model at TREC HARD (Abdul-Jaleel et al., 2004).

Beyond the TREC HARD campaigns, Tiedemann and Mur (2008) compared the utility of various types of segmentation approaches for question answering, including TextTiling, fixed-length overlapping windows, and a segmentation method based on co-reference chains over named-entities. The conclusions from this work suggests that passages based on windowing approaches provide the highest QA effectiveness. The authors emphasised that the gains that could be achieved by using semantically motivated passages of variable-length are outweighted by the use of passages of uniform length, on which standard IR models perform better. In a similar study to Kaszkiel and Zobel (2001), Lamprier et al. (2008) revisited this issue and showed that semantically motivated passages can be as effective as fixed-length arbitrary passages if appropriate length normalisation is applied to control for length variations.

Besides the development of new passage retrieval techniques, a substantial amount of research effort has been devoted to the development of novel measures for evaluating passage retrieval effectiveness (Allan, 2001, 2004; Wade and Allan, 2005) based on relevance assessments collected for arbitrary sections of a document. Compared to the evaluation of document retrieval, evaluation of unsegmented retrieval poses several additional challenges. First, the passages retrieved for a document may not perfectly align with those that have been marked as such in the ground truth, but instead have an "overlapping" section, in which case it is not clear whether the passage should be considered relevant or not. A second fundamental problem is how to deal with redundant results in the ranked list which may arise if the system under evaluation returns overlapping passages from the same document. Many aspects related to passage retrieval evaluation were later revisited in the context of XML retrieval and SCR, and are covered more extensively in the Chapter 7 of this thesis.

2.4.3 XML retrieval

XML retrieval (Luk et al., 2002; Fuhr et al., 2002) refers to the task of finding relevant information from within collections of semi-structured text documents, specified in the eXtensible Mark-up Language (XML). An XML document specifies a set of nodes and elements organised into a tree-like hierarchical structure. The internal elements of the tree specify structural information, while external elements (leaves) contain the textual content. For instance, a book might be specified in XML format by a root element `<book>` containing one or more `<chapter>` elements, which in turn may contain multiple `<section>` elements. At the deepest level of a book's schema, a section could contain several `<paragraph>` elements each containing the actual text content of a specific paragraph.

In an XML document, each internal element can be seen as a passage representing the contents of its children, so that different levels in a XML tree correspond to different levels of content granularity. The goal of XML retrieval is then to retrieve the most appropriate elements from within a collection of XML documents in order of relevance to a query. Appropriateness in this context refers to the granularity of the retrieved content. The ideal element to be retrieved for a query is the most specific element that contains just enough information to satisfy the information need from the user, without including any additional irrelevant information. The query in this case may be free text or optionally impose structural constraints over the type and granularity of the content being sought by the user.

Much of the research in XML retrieval has been driven by shared tasks organised by the INitiative for the Evaluation of XML retrieval (INEX) (Fuhr et al., 2002; Fuhr and Lalmas, 2007) Researchers experimented with several approaches during these benchmarks, most of which attempted to extend standard IR models to consider the structural information of the documents. Carmel et al. (2003) presented an extension of the vector space model (VSM) that represents XML elements and structured queries as vectors of pairs $(tf, path)$ where tf refers to the frequency of a term in an element located at a given $path$ within the element's hierarchy. A modified cosine similarity function is then used to score elements against a query, that multiplies vector components that have similar paths.

The approach proposed by Gövert et al. (2002) attempts to propagate the weights assigned to terms in the leaf elements onto their parent elements. In order to avoid the elements at higher levels in the document tree from always obtaining greater scores than their children, the propagation procedure down-weights the transferred weights at each level in the hierarchy by some pre-defined factor. Similar approaches were later proposed by other researchers, who achieved similar propagation effects by using more principled techniques. Most of these were based on language models (Kamps et al., 2004; Ogilvie and Callan, 2004, 2005). In particular, Ogilvie and Callan (2004, 2005) estimate a language model for each element in a document tree, which they later use to calculate the relevance score of an element as the probability that its language model generates the query. The language model of an element is estimated via linear interpolation of language models obtained from: the text of the element itself, that of its children, parent, document, and document collection.

In general, techniques that exploited evidence from multiple levels of content granularity performed best at the INEX benchmarks. This result goes inline with previous results observed in document and passage retrieval research, that showed that using evidence from documents and passages provided improved retrieval effectiveness for both tasks. Kekäläinen et al. (2009) and Arvola et al. (2011) re-branded this set of techniques as "contextualisation" approaches, to emphasise the fact that elements can be ranked more effectively when considered within the context of their container (parents) or neighbouring (siblings) elements. Arvola et al. (2011) analysed the effects of different contextualisation approaches on retrieving relevant elements at three predefined levels of content granularity: paragraphs, subsections, and sections. The results of their experiments demonstrated that vertical contextualisation (parents/children) as well as horizontal contextualisation (siblings) can improve the retrieval elements at any of these granularity levels.

2.5 Summary

This chapter reviewed fundamental technologies in SCR. IR provides methods for indexing and searching relevant material within large collections of text documents. ASR provides the tools needed to convert speech to text. Content structuring provides approaches to derive small text fragments from longer pieces of multi-topical documents that are more appropriate as retrieval units.

Two major frameworks for ranked retrieval were presented in detail. The vector space model (VSM) represents text as vectors in a vector space, where components correspond to individual terms and weights are derived from the product between within-document and inverse document term frequencies. The VSM ranks documents based on their cosine similarity against the query in the vector space. In the probabilistic framework for IR, documents are ranked based on their probability of being relevant to the query. Various models were developed to estimate these probabilities. The binary independent model (BIM) considers presence and absence of terms and assumes term independence. The state-of-the-art Okapi BM25 model extends this to consider within-document and within-query term frequencies and applies a length normalisation mechanism. Several measures have been developed for evaluating the quality of a ranked list of search results, including: average precision (AP), normalised discounted cummulative gain (nDCG), ranked-biased precision (RBP) and expected reciprocal rank (ERR). A recent trend in IR research is to interpret (and develop) evaluation measures as models of user behaviour.

The most effective speech recognition systems use statistical models to identify speech sounds in the speech signal and to represent valid words combinations in a language. Speech is processed by a sequence of components. The front-end component converts speech into a sequence of spectral feature vectors. The acoustic model (AM) component treats vectors as the observations of a hidden Markov model (HMM), used to estimate probabilities of phone sequences. While the language model estimates probabilities of word sequences. Lastly, a decoding algorithm searches within the vast space of possible word sequences for the most likely words spoken, and represents the output as a N-best list of hypothesis, a lattice, or as a confusion network.

Several content structuring methods were reviewed for automatically segmenting a piece of text into topically homogeneous units. Windowing approaches divide the text into arbitrary fragments. Other techniques attempt to detect topic-shifts in the text by finding boundaries that maximise the intra-segment cohesion and minimise inter-segment similarity. TextTiling (TT) uses a VSM to compute similarity scores between adjacent passages and then finds local minima along the document. C99 improves upon TT by considering relative ranks instead of raw cosine scores and optimising a global cost function; Utiyama and Isahara (UI) propose a probabilistic approach to segmentation which was later given a Bayesian formulation in the BayesSeg algorithm. The Minimum Cut algorithm cast the problem as one of finding an optimal partition of a graph that minimises the normalised-cut criterion. Content structuring methods have also been applied to speech transcripts, either in isolation or in combination with other acoustic/prosodic features that are indicative of topic shifts in spoken content. Lexical and acoustic features are then used to train machine learning models to predict the locations of possible topic boundaries.

This chapter also reviewed previous research on exploiting sub-document structure for improving the retrieval quality of documents, passages, and XML elements. Although methods based on lexical-cohesion are able to produce more topically cohesive segments, arbitrary overlapping passages have commonly been found more effective when used as evidence of relevance in the ranking of documents and passages. The main reason attributed to this effect is that standard IR techniques are less effective at scoring elements with highly variable length, even if length normalisation mechanisms are applied. Further research in XML retrieval suggests that elements can be ranked more effectively when considered within the context of their container document and related elements.

Chapter 3

Review of SCR Research

Chapter 2 described the three fundamental technologies needed to enable SCR: automatic speech recognition (ASR), to convert speech to text; text indexing and retrieval, to provide efficient ranking of relevant text documents; and content structuring, to fragment long documents into short topically-coherent excerpts. This chapter focuses on how these three technologies have been combined together in past and recent research seeking to maximise the effectiveness of SCR systems.

The chapter begins with Sections 3.1 and 3.2, which review literature on past SCR research. Throughout these sections, particular emphasis is given to research that has explored methods for handling ASR errors and structuring of spoken content to enable immediate access to relevant material. Section 3.3 discusses previous research that has attempted to exploit acoustic/prosodic information in SCR and related speech retrieval applications. This work tries to move beyond lexical-based retrieval techniques to incorporate additional acoustic/prosodic information from the speech signal. This non-lexical information has mainly been used for increasing the quality of content structuring techniques, and for identifying acoustically-emphasised keywords in speech.

3.1 Experiments with formal speech

This section reviews initial research on SCR, conducted inbetween 1990-2001. This preliminary work focused on collections of formal speech, mainly voice mail and radio and TV broadcast news.

3.1.1 Early work: voice mail and private collections

In the early nineties, the widespread use of the hidden Markov model (HMM) framework (Levinson et al., 1983) for speech recognition allowed the creation of ASR systems capable of recognising words in continuous speech from a relatively short fixed-vocabulary. This led to the appearance of the first commercial applications for filtering and classifying speech messages based on word-spotting techniques.

Word or keyword spotting is the task of determining whether a word from a given vocabulary is present or absent in some speech sample. Some of the earliest research in spoken document indexing was based on word-spotting techniques (Rose, 1991; Wilcox et al., 1992). The basic approach consisted of classifying speech messages into a set of topic categories based on spotted keywords in the speech stream. Identified topics could then be used for re-routing telephone calls or as indexing terms for post-retrieval and organisation of voice messages.

Early SCR systems based on word spotting techniques could only process queries that resembled one of the topic categories initially provided to the word spotters at indexing time, that is, at the time when the spoken material had to be recognised. This restricted the number of possible queries that a system could handle at retrieval time or forced systems to re-index the entire collection every time a new topic category or term was provided in a search request. Subsequent research in SCR focused on removing this practical limitation.

Glavitsch and Schäuble (1992) presented the first prototype of a modern SCR system based on large-vocabulary speaker-independent continuous speech recognition (LVCSR). This method used sub-words as indexing terms and performed retrieval by matching the sub-words in the query against those recognised in the spoken documents. More importantly, this method set the basis for designing SCR systems which, more in line with conventional text retrieval techniques, permitted efficient retrieval for queries whose vocabulary did not need to be provided in advance of the indexing process.

Significant contributions to the field were made in the context of the Voice Mail Retrieval (VMR) project led by researchers at Cambridge University (Jones et al., 1997). In this work the idea of vocabulary-independent word-spotting was proposed, later known as Phone Lattice Spotting (PLS) (James and Young, 1994). In PLS, a lattice of hypothesised phone-transitions is generated after a first recognition pass over the speech data. The presence of arbitrary query terms can then be determined at query-time by searching for the terms' phonetic-transcriptions in the pre-computed phone-lattices. Subsequent research at Cambridge University investigated hybrid approaches which combined word-level LVCSR with PLS to account for out-of-vocabulary (OOV) terms in the query (James, 1995, 1996; Brown et al., 1996; Jones et al., 1996). Despite recent advancements in ASR technology, techniques such as PLS may still be useful for SCR, especially for low-resource languages for which there may be insufficient data to train a complete ASR system.

The Informedia Project was another important research initiative at the time, with focus on developing content-based retrieval techniques to support search in video collections (Wactlar et al., 1996, 1999). The system produced in the context of this project was one of the first to provide large scale multimedia retrieval and browsing capabilities by exploiting LVCSR technology for speech indexing and visual analysis for content segmentation.

During this period, the effectiveness of the SCR techniques was evaluated over small collections of privately owned speech material, typically consisting of no more than a couple of hours of radio news or voice mail messages. Most of the speech content used for evaluation was characterised as being formal, read or scripted, produced by a relatively small number of speakers, in silent and controlled recording conditions and by using good-quality recording devices. Also, standard document retrieval measures like precision and mean average precision (MAP) were generally used to evaluate the effectiveness of the SCR methods. Cross-comparisons of performance across research labs were rare during this period. It was not until the late nineties, with the first Text REtrieval Conference (TREC) spoken document retrieval (SDR) campaigns (Garofolo et al., 2000), that the research focus began shifting towards cross-lab evaluations over larger spoken collections and more challenging types of speech data.

3.1.2 Broadcast news: the TREC-SDR campaigns

The Text REtrieval Conference (TREC) is a series of workshops and shared tasks that provide a common framework for the evaluation and comparison of large-scale text retrieval experiments[1]. Every year, TREC organises several tasks or tracks that pose particularly interesting research problems to the IR community. Organising teams are in charge of designing the task and providing the document collection and queries to the participants. Participant teams must develop IR systems that address the task and submit their retrieval results (runs) for quality estimation.

The first TREC track that focused on SCR was held in 1997 (Voorhees et al., 1997; Voorhees and Harman, 2005) as part of the TREC-6 workshop. TREC-6 SDR was a known-item retrieval task. As opposed to an ad-hoc retrieval task, where multiple documents from the collection can be relevant to a query, in a known-item task there is only one known relevant document per query. In TREC-6 SDR, systems were evaluated over a collection 50 hours of broadcast news speech. The collection was manually pre-segmented into 1,451 news stories, each corresponding to the presentation of a single news event. The task consisted of ranking news stories given a text query so that the single known relevant story for that query was ranked on top. Precision-based effectiveness measures like mean reciprocal rank (MRR) were used for estimating the quality of the retrieval runs.

An important conclusion drawn from TREC-6 SDR was that standard text-retrieval techniques are robust to relatively high word error rates (35-40% WER) in the document transcripts, but that there is a significant decrease in retrieval effectiveness when more erroneous transcripts (above 50% WER) are used (Garofolo et al., 2000). A wide range of techniques were proposed at TREC-6 SDR to cope with the ASR errors present in the document transcripts. Most notably, the use of word confidence scores from the ASR to calculate expected term frequencies, the exploitation of N-best recognition hypotheses from a single or multiple ASR systems, and phonetic-based matching (Crestani et al., 1997; Siegler et al., 1997; Smeaton et al., 1997). Experiments with these techniques indicated

[1] `http://trec.nist.gov/`

that using word confidence scores did not provide gains in retrieval effectiveness, while considering additional terms in the N-best lists or multiple ASR outputs demonstrated potential at recovering from deletion or substitution errors.

The TREC-7 SDR track ran a year after TREC-6. In contrast to TREC-6, TREC-7 SDR posed an ad-hoc retrieval task over a larger document collection comprising 87 hours of broadcast news speech divided into 2,866 news stories (Garofolo et al., 1998). One of the main focuses of study in TREC-7 SDR was the correlation between ASR errors and retrieval effectiveness. Analysis of the results submitted at this track showed that there is a negative, albeit gentle, linear-correlation between ASR errors and retrieval effectiveness. In fact, WER was found to negatively correlate with retrieval effectiveness but an even stronger negative correlation was found when WER was restricted to named entities (Garofolo et al., 1998), which were commonly present in the queries. This observation indicated that the misrecognition of highly informative terms, like proper names, had a major impact on retrieval effectiveness.

A large number of groups investigated document and query expansion techniques based on pseudo-relevance feedback (PRF) to cope with ASR errors in the broadcast transcripts. Expansion techniques were generally found effective at reducing the performance gap between the retrieval from automatic and perfect transcripts, especially when expansion terms were extracted from external in-domain resources (Singhal et al., 1999; Singhal and Pereira, 1999). In its simplest form, this document expansion technique consisted of augmenting the document transcripts with topically-related terms extracted from an external collection that is free from transcription errors and that contains similar topics. To find topically-related terms for a spoken document, a text query is first constructed by selecting terms from the document transcript. This query is later used to rank documents from within the external corpus in order of relevance. Terms are then selected from the top K ranked documents in the external corpus and used to expand the contents of the target document transcript. This expansion approach was found to be effective at reducing the number of term mismatches between the query and the document transcripts from the TREC-7 SDR collection since it helped to recover important terms that might haven been deleted or substituted from the transcripts during the ASR process. Besides document expansion, researchers experimented with similar strategies to augment the query text with terms extracted from the collection of noisy transcripts or from parallel corpora (Abberley et al., 1998). Due to their demonstrated effectiveness, expansion techniques were regularly applied by research groups in subsequent editions of the TREC SDR tracks (Gauvain et al., 1999; Johnson et al., 2000; Renals and Abberley, 2000).

In the TREC-8 and TREC-9 SDR tracks, the Topic Detection and Tracking corpus (TDT-2) of broadcast news (Cieri et al., 1999) was used as the target collection for retrieval. This corpus contains 557 hours of speech content from 21,754 TV and radio broadcast news stories, transcribed with WERs ranging from 20-30%. TREC-9 introduced a new ad-hoc retrieval condition that required systems to retrieve relevant news stories with no prior knowledge about the exact location of story boundaries. In this case, systems were required to retrieve jump-in or playback time-points falling within the boundaries of a relevant story. This criteria was used in precision-based effectiveness measures to determine whether a given search result should be treated as relevant or not in the measure calculation. In order to avoid rewarding systems for returning near-duplicates in the result lists, which would occur when multiple results pointed to the same relevant story, the evaluation procedure discarded any returned jump-in points falling in the same relevant story than some other better-ranked result.

Participating teams of the unknown-boundary condition at TREC-8 and TREC-9 SDR experimented with different content structuring techniques for dividing a spoken document into multiple retrieval units as well as segment consolidation techniques for removing near-duplicate results from the ranked lists. A simple and commonly adopted structuring approach based on the work from Hearst and Plaunt (1993), previously applied to SCR by Brown et al. (1995), consists of slicing a document transcript into a sequence of overlapping windows (Smeaton et al., 1997; Abberley et al., 1999b; Johnson et al., 2000; Renals and Abberley, 2000). The extracted windows were then ranked in order of relevance to the search query and their starting time-points retrieved as the jump-in points to be inspected for relevance.

Various segment consolidation strategies were proposed to avoid returning near-duplicates at top positions in the rankings. Johnson et al. (2000) proposed to filter out lower ranked near-duplicate segments that overlapped with a higher ranked one in the results list. While Abberley et al. (1999b) and Renals and Abberley (2000) adopted a recombination strategy, whereby segments overlapping in time were merged into a single one if their ranks lied within some fixed distance r. This recombination process was carried out for a number of iterations, shrinking the value of r each time, until no more segments could be merged. Abberley et al. also experimented with alternative functions to re-estimate the relevance score of merged results, including taking the maximum score across all overlapping segments, re-calculating the relevance status value for the combined segment, and using an average score that corrected for segments' length and amount of overlap. Among these, taking the maximum score resulted in increased retrieval effectiveness over the rest.

In related work, Abberley et al. (1999a) compared time-based versus word-count-based sliding windows for content segmentation and arrived at the conclusion that the two approaches provide similar levels of retrieval effectiveness. Additionally, Abberley et al. (1999a) and Quinn and Smeaton (1999) explored the effects of varying the size and amount of overlap between adjacent windows and concluded that short windows of 30 seconds with an 33-50% of overlap performed best on TREC-SDR tasks. Gauvain et al. (2000) analysed the length distribution of relevant news stories and hypothesised that a two-level windowing approach that simultaneously targets short and long stories could benefit retrieval. Experiments with this multi-level approach showed minor gains in retrieval effectiveness over a single-level windowing strategy. In addition to windowing approaches, (Johnson et al., 1999b) experimented with the TextTiling segmentation algorithm (Hearst, 1997), but found this approach less effective in practice than using fixed-length overlapping windows.

Evaluation results of the submitted runs at TREC-8 and TREC-9 SDR showed further evidence that text retrieval techniques are fairly robust to high WER conditions. Later analysis showed that this effect could be attributed to the characteristics of broadcast news speech, in which topically important keywords are mentioned multiple times during the coverage of a news story. Results also reflected that the unknown-boundary condition significantly increased the difficulty of the retrieval task. Retrieval effectiveness was on average 20% lower when story boundaries were unknown to the retrieval systems.

This period ended with the publication of Garofolo et al. (2000) that stated that SDR was a solved problem and that research should shift focus towards more challenging tasks such as question-answering, spoken-queries, video retrieval, or on exploiting paralinguistic information to improve the navigation of spoken documents. This hasty conclusion was primarily driven by the following facts: LVCSR systems could produce 1-best transcripts with relatively low WER (c.a. 20%); text retrieval techniques were robust to relatively high WERs conditions; and speech recognition errors could to some extent be alleviated using expansion techniques, which was seen to even produce comparable performance to that obtained when using perfect transcripts (Johnson et al., 1999b; Singhal and Pereira, 1999; Woodland et al., 2000). Although these are valid conclusions, the nature of broadcast news speech facilitated the recognition and retrieval of relevant content, and masked other issues that were encountered in later experiments with less formal speech material.

3.2 Experiments with conversational spontaneous speech

In 2001, Allan (2001) stated that there was still room for research in SCR, since TREC SDR had mainly focused on long documents and long queries for which standard retrieval techniques were not dramatically affected by speech recognition errors. Instead, it was proposed that research should focus on short or spoken queries, or on tasks like question answering, where the boundaries of the ideal passages containing an answer are unknown and passages may not contain enough terms to compensate well for ASR errors. Furthermore, Allan pointed out the importance of using non-linguistic information and that of moving beyond scripted speech to less formal spontaneous conversational speech.

Spontaneous speech does not present the same characteristics as the speech found in broadcasts of TV and radio (Ward, 1989). Spontaneous speech usually contains disfluencies, such as filled pauses, repetitions, repairs, and false-starts. Other important differences include the presence of ungrammatical constructions or ill-formed sentences and the frequent use of ellipsis and interjections. Vocabulary usage also differs significantly depending on the level of spontaneity. While formal speech usually contains more content bearing words that may describe the central topic of a conversation more precisely, casual speech contains more words that provide an implicit and inexact description of the main topic (Larson and Jones, 2012a). Furthermore, topical boundaries are less clearly specified in spontaneous conversational discourse, where even rhetorical topics are common. Further complications are present in speech that contains multi-party dialogues, utterances from non-native speakers, background noise or music, or that are recorded in poor acoustic conditions.

Subsequent research on SCR focused on more challenging speech collections containing a higher degree of spontaneity than broadcast news. Research then focused on retrieval from collections of interviews, lectures, meetings, academic talks, and TV content.

3.2.1 Interviews: the CLEF-SR campaigns

The Conference and Labs of the Evaluation Forum (CLEF)[2], formerly the Cross-Language Evaluation Forum, organised a cross-language speech retrieval (CL-SR) task from 2005 to 2007 over a collection of spontaneous conversational speech, consisting of interviews in English and Czech with survivors of the Holocaust (White et al., 2005; Oard et al., 2006; Pecina et al., 2007a).

In CL-SR 2005 (White et al., 2005), topically coherent segments of speech were manually labelled by subject matter experts for each interview and the task was designed as a known-boundary retrieval task. The document collection comprised 589 hours of speech divided into 8,104 topically homogeneous segments. The collection was automatically transcribed with WER of approximately 38%, showing evidence that ASR of interview speech was more difficult than the recognition of broadcast news. Besides ASR transcripts, metadata about the interviews including summaries, lists of keywords, and mentions of important people were manually annotated in each interview and made available to the task participants.

Overall, the evaluation results of CL-SR 2005 showed that retrieval of interview segments was substantially more difficult than the retrieval of news stories from the previous TREC SDR tasks (Wang and Oard, 2005). Researchers hypothesised that this was due to important keywords and named entities being misrecognised in the ASR transcripts or not even spoken by the participants in the interviews. Since informative topical related words were frequently not present in the transcripts, systems had to rely on the manually generated metadata to maximise retrieval performance. Even by exploiting metadata, the performance of the SCR systems was considerably lower than in the previous experiments with broadcast news speech.

CL-SR 2006 and 2007 included an unknown-boundary condition where systems were required to produce a ranked list of starting point suggestions instead of manually predefined segments (Oard et al., 2006). For these editions of the CL-SR task, a collection of interviews in Czech was used. As a baseline collection of segments, the document transcripts were automatically segmented into overlapping windows of 3 minutes length

[2] http://www.clef-initiative.eu/

and 2 minutes of overlap. To measure the quality of a ranked list of entry points returned by a system, an adaptation of generalised average precision (gAP) was used (Liu and Oard, 2006). Recall from Section 2.1.3 that gAP is an extension of average precision (AP) to graded relevance assessments. In the CL-SR tasks, the relevance grade of a retrieved entry point is a continuous value that depends on the temporal distance between this point's and an ideal jump-in point indicating the beginning of a relevant speech fragment.

Evaluation results across participating teams at CL-SR 2006 and 2007 indicated that stemming is important for SCR in Czech (Pecina et al., 2007a; Levow, 2007) and, more importantly, that segmentation granularity affects retrieval performance (Ircing and Müller, 2007). Regarding the latter observation, windowing approaches that generated segments with a length that corresponded better to the real length of the relevant content were found to be the most effective. In the known-boundary condition, researchers obtained improved results with techniques that combined evidence from multiple ASR transcripts and manually generated metadata using techniques such as field weighting and XML retrieval (Oard et al., 2006; Jones et al., 2006; Hiemstra et al., 2006).

3.2.2 Broadcast TV: the MediaEval campaigns

The benchmark initiative for Multimedia Evaluation (MediaEval)[3] has organised a yearly task devoted to SCR since 2010. The Rich Speech Retrieval (RSR) task at MediaEval 2011 was a known-item task that required systems to retrieve jump-in points within relevant portions of semi-professional TV shows (Larson et al., 2011; Schmiedeke et al., 2013). Participating teams experimented with different automatic segmentation methods, including windowing approaches (Wartena, 2012), segments generated by the speaker diarisation module of an ASR (Wartena, 2012; Alink and Cornacchia, 2011; Schmidt et al., 2011; Aly et al., 2011), and the C99 (Choi, 2000), TextTilling (Hearst, 1997), and MinCut (Malioutov and Barzilay, 2006) algorithms (Eskevich and Jones, 2011a; Wartena, 2012).

The work described by Wartena (2012) compared different segmentation strategies and showed that a windowing approach with filtering of lower-ranked near-duplicates provides improved SCR performance relative to a segmentation generated by the MinCut algorithm. This work also highlighted the fact that SCR effectiveness decreases significantly when windowing segmentation is used with a window length that differs considerably from the average length of the relevant material and suggest that topically-motivated segments, as those produced by MinCut, may require less parameter tuning than windowing approaches.

The approach described by Aly et al. (2011) performed retrieval of speech segments by considering evidence from the full contents of the document in which the segment occurs. In this work, the relevance score of a segment was obtained by linearly combining the segment's relevance score with that of its containing document. This work also proposed the selection of alternative jump-in points in the vicinity of the returned segments as a segment consolidation strategy. Similarly, Alink and Cornacchia (2011) and Schmidt et al.

[3] `http://www.multimediaeval.org/`

(2011) proposed a two-stage cascaded approach. In the first stage, the top-N documents that best match the query were ranked. In a second stage, a ranking was produced for the segments contained in the documents retrieved in the first stage. Overall, the highest SCR effectiveness in the RSR task was obtained by using windowing based segmentation instead of text segmentation algorithms Wartena (2012), and by exploiting user-generated metadata (Eskevich and Jones, 2011a).

Subsequent analysis of the RSR results showed that, independently of the IR model used and irrespective of the WER of the transcripts, SCR systems were able to return relevant content at high ranks as long as they implemented a segmentation strategy that fully captures the topic of the relevant content in a single segment without including too much irrelevant material (Eskevich et al., 2012b). Thus, an ideal segmentation strategy for SCR should not undersegment nor oversegment the relevant content or, in other words, should maximise the within-segment precision and recall of the returned segment with respect to the relevant material. This observation motivated the development of alternative evaluation measures for retrieval of unsegmented speech content based on temporal precision: mean average segment precision (MASP) (Eskevich et al., 2012c). In this family of measures, precision is estimated as the proportion of relevant content that is captured by a retrieved segment, measured in units of time, relative to the temporal length of the segment.

In subsequent years, the RSR task was renamed as the Search and Hyperlinking (S&H). The S&H 2012 task was a known-item task that evaluated systems over an extended subset of the RSR collection with 2,125 hours of semi-professional TV content (Eskevich et al., 2012a). The best evaluation results were obtained with a combination of query expansion, expansion of segments with metadata, windowing-based segmentation, and filtering of lower-ranked overlapping results (Galuščáková and Pecina, 2012; Eskevich et al., 2013c).

Subsequent iterations of the S&H task in 2013, 2014, and 2015 posed an ad-hoc retrieval task over a large document collection of circa 2700 hours of TV broadcast from the British Broadcasting Corporation (BBC) (Eskevich et al., 2013a, 2014, 2015). Approaches that determine the best granularity level of the segments to be retrieved at query time were proposed in (Preston et al., 2013; Schouten et al., 2013). In particular, Preston et al. (2013) evaluated a kernel density function along the timeline of a video to represents local variations of retrieval scores throughout time. The process of estimating a density function for a video consisted of clustering query terms based on their temporal distance in the video transcript. A hierarchical agglomerative clustering algorithm was used for this purpose, forming term clusters every time the time distance between the middle points of two clusters surpassed a threshold. A Gaussian function was then estimated at the centre of each of the resulting clusters, with amplitude given by the retrieval score of the cluster against the query, and width equal to 30% of the cluster's duration. The final density contour for a video was calculated by summing all Gaussians corresponding to all clusters in the video. Finally, a ranked list of segments was constructed based on the regions delimited by the clusters and their scores. Schouten et al. (2013) proposed a similar approach in which individual density functions were estimated for each query term and then summed. Potential segment boundaries were then detected by locating valleys in the density function that are above a certain threshold.

An appealing characteristic of the approaches proposed by Preston et al. and Schouten et al. is their ability to construct variable-length segments dynamically, based on the contents of the query. Despite this theoretical advantage, these dynamic content structuring approaches performed poorly at the S&H 2013 task compared to simple windowing methods. Similarly, the approach described by Galuščáková and Pecina (2014b) that used decision tree classifiers to predict putative segment boundaries from features such as cue phrases, length of pauses, speaker diarisation boundaries, and TextTilling boundaries, did not provide clear gains in retrieval effectiveness over windowing approaches. Interestingly, Sahuguet et al. (2013) showed that segmentation based on visual scenes can perform as well as those based on fixed-length windows suggesting that a multimodal approach could perform better than one that solely relies on speech or linguistic features. The best retrieval performance at S&H 2013 was obtained by Eskevich and Jones (2013b) using fixed-length overlapping windows and adjusting the jump-in points of the retrieved segments. In the latter, alternative jump-in points were chosen by selecting nearby speaker diarisation boundaries or pauses longer than 500 milliseconds.

3.2.3 Lecture recordings: the NTCIR campaigns

The NII Testbeds and Community for Information access Research Project (NTCIR)[4] organised the "IR for Spoken Documents" (SpokenDoc) Task in 2011 (Akiba et al., 2011) and 2013 (Akiba et al., 2013a) offering SDR and SCR ad-hoc tasks over a collection of lecture recordings in Japanese. In the SpokenDoc-1 task (Akiba et al., 2011), the Corpus of Spontaneous Japanese (CSJ) (Maekawa et al., 2000) was used as the document collection. This corpus contains 612 hours of speech recordings of academic presentations. In the SpokenDoc-2 Akiba et al. (2013a) task, a smaller corpus containing 27 hours was used and the task switched to a passage retrieval task instead of full SDR.

Most participating teams at the SpokenDoc benchmarks focused on techniques to reduce the impact of ASR errors on retrieval effectiveness. Notably, the work by Tsuge et al. (2011) showed that using more than one hypothesis from the ASR is sometimes beneficial for SDR, while Kaneko et al. (2011) and Akiba et al. (2013b) showed that performing matching at the syllable level can help in overcoming OOV errors in the query.

In the passage retrieval task, researchers compared the performance of windowing approaches (Nanjo et al., 2011) and lexical cohesion algorithms (Eskevich and Jones, 2011b, 2013a). The evaluation results from these experiments provided further evidence that windowing segmentation is more effective for SCR than semantically motivated segments obtained by automatic segmentation algorithms. Moreover, the method described

[4] `http://http://research.nii.ac.jp/ntcir`

in (Akiba et al., 2013b) that redefined the boundaries of retrieved passages by searching for boundaries that could maximise the relevance score between the query and the passage underperformed a simpler windowing approach based on fixed-length segments.

Besides the CLEF and NTCIR campaigns, which evaluated SCR techniques over Japanese and Czech speech, other research has focused on porting traditional SCR approaches to languages other than English, including Mandarin (Chen et al., 2001), French (Guinaudeau and Hirschberg, 2011), and Spanish (Varona Fernández et al., 2011). On this regard, researchers have frequently highlighted the importance of applying language-specific text processing methods, principally tokenisation and stemming, for an effective application of IR techniques to ASR transcripts (Pecina et al., 2007b; Nanjo et al., 2014).

3.2.4 Final remarks on content structuring and ASR errors

Previous research in SCR suggests that windowing approaches can produce retrieval units that provide increased retrieval effectiveness on top of segmentation algorithms that are based on lexical cohesion. Using fixed-length overlapping segments as the basic unit of retrieval is beneficial for a number of reasons. First, IR models are known to perform better when the collection of items to be ranked are of similar length (Singhal et al., 1996). This is because the ranking of equal-length documents removes the need to apply length normalisation mechanisms in IR, and with that the need to adjust the relevance scores of documents relative to their length.

Second, the overlap introduced between consecutive windows can avoid splitting a topically consistent piece of relevant information into disjoint segments, thus reducing the chances of separating query terms that may appear in close proximity in the document. This increases the probability of capturing term phrases that may appear in the query in a single segment, and term proximity information in general which has long been considered a useful indicator of document relevance (Büttcher et al., 2006).

Third, windowing approaches are not affected by recognition errors, whereas ASR errors may have a direct impact on the quality of lexical cohesion segmentation methods. The fact that ASR errors tend to occur rather randomly across a transcript may disrupt non-random chains of related terms appearing in adjacent sentences and may in turn encourage segmentation algorithms to produce spurious breaks. In this regard, there is empirical evidence that the retrieval effectiveness associated with lexical cohesion segmentation methods degrades when the segments used for retrieval are produced with ASR instead of perfect transcriptions (Eskevich et al., 2015).

Despite the advantages mentioned above, standard windowing approaches present some major drawbacks. Increasing the length of the extracted passages or the amount of overlap increases the possibility of fully capturing the contents of a relevant excerpt (high within-segment recall), at the expense of losing topical "focus" by increasing the amount of irrelevant content added to the passage (low within-segment precision). A system that considers excessively long passages could potentially make the task of identifying an appropriate jump-in point more difficult, since there is a high risk that the beginning of a passage would be too far away from the putative relevant section to be of any use for the user. Variations in the ranking of relevant segments caused by improper segmentation were studied in (Eskevich, 2014; Eskevich et al., 2015), which concluded that retrieval techniques fail to retrieve relevant content at top ranks when such content is not fully contained within the boundaries of a segment (low-recall) or when there is too much irrelevant content included in the container segment (low within-segment precision).

Another important issue of window-based approaches is that is unclear what the best values for the window and step lengths should be for a given set of queries and collection. Previous research has evidenced that the optimal values for these parameters may depend on the retrieval task and underlying structure of the spoken material. Even if the optimal configurations are known for a collection, it is not guaranteed that the resulting window-based segmentation will perform well for all queries. In general terms, it is reasonable to think that no single static segmentation could possibly satisfy every information need that a user may have when interacting with a SCR system. The content sought by a user may well be spread over a wide array of lengths and granularity levels across the collection and be covered at different levels of detail across different documents.

Query independent structuring approaches for SCR in which retrieval units are defined at indexing time, prior to retrieval, are thus less likely to generalise well across a diverse set of search requests. This observation also applies to content structuring approaches based on topic segmentation algorithms, despite the fact that these have been shown to produce segments that align better with the length of relevant sections (Wartena, 2012). On the contrary, query dependent or dynamic structuring approaches, like segment recombination/merging strategies, or clustering based on query term density estimation, seek to determine or refine segments at retrieval time and are thus capable of adapting search results to relevant regions of variable length.

The majority of segmentation approaches explored in SCR research produce a flat, linear, sequential structure of segments, where each segment is assumed to represent a single "topic" that may become the target of a search request. Flat structures make the assumption that topics do not have sub-topics and that, consequently, topics have a similar level of information specificity. A more realistic approach would consider a hierarchical structure of topics in which levels in the hierarchy could represent different levels of topic specificity. Segments representing more specific topics could be arranged at lower levels in the hierarchy than those representing broader, more general, topics. Different levels in this hierarchy of segments could then be targeted for retrieval depending on the user's query and the characteristics of the content. Recent research in this direction includes the work on hierarchical topic segmentation, carried out by Simon et al. (2015b). Experiments conducted by Simon et al. at the Search & Hyperlinking (S&H) tasks showed that by targeting elements located at different granularity levels in the hierarchy, an IR model could retrieve segments that were highly diverse in terms of both length and content specificity compared to other segmentation approaches. However, the effectiveness of their approach could not be properly determined at the S&H task, as the evaluation procedure tended to favour segments that were considerably longer than those returned by the hierarchical approach.

Besides an overall degradation in the segmentation quality of methods based on lexical-cohesion, recognition errors have been found to affect the ranking of speech transcripts in interesting ways. In particular, Shou et al. (2003) and Sanderson and Shou (2007) observed that relevant transcripts with low WER tend to be ranked higher by standard IR models than relevant transcripts with high WER. In other words, the higher the number of ASR errors in the transcripts, the lower will likely be its rank in the results list. The main reason for this effect is that the presence of ASR errors can reduce the frequency and diversity of query-related terms in the documents, diluting and hindering regions that would otherwise contain a high density and high amount of distinct query-term occurrences (Sanderson and Shou, 2007). Later, Eskevich et al. (2015) revisited this problem and found that the effects seen by Sanderson and Shou also apply to non-relevant documents. Hence, independently of the relevant status of the document, highly noisy documents are generally ranked lower than less errorful documents.

3.3 SCR beyond lexical matching: exploiting acoustic features and prosody

Previous research in SCR has mostly focused on reducing the impact of ASR errors on retrieval performance and on reducing user auditioning effort by structuring the speech content into smaller audio excerpts to enable passage retrieval. With the exception of some content structuring approaches reviewed in Section 2.3, most SCR methods proposed in the past only rely on the lexical information recognised by the ASR, neglecting other sources of information that are also present in the speech signal. An important, potentially useful, source of information for SCR is given in the prosody of the speech which characterises variations in the way words are spoken. This section reviews past research on the use of prosodic information for SCR and other related speech processing tasks.

3.3.1 Speech Prosody

Prosodic information has been shown useful in various speech processing tasks, including SCR tasks (Chen et al., 2001; Guinaudeau and Hirschberg, 2011; Ward et al., 2015). This section provides general background about prosody and presents two important aspects of prosody, prominence and phrasing. These two facets of prosodic information could potentially be used to improve retrieval effectiveness in SCR.

Pitch, duration, and loudness

In linguistics, prosody is defined as the "suprasegmental" characteristics of speech (Lehiste, 1970). These are features that cannot be characterised as discrete speech units (segments), such as vowels, consonants, or syllables, but that rather occur simultaneously with them, spanning across multiple units, and describing their intonational and rhythmical properties.

Prosody is more informally defined as the variations of pitch, duration, and loudness of the speech units across time. The acoustic correlates of these features, which can be extracted automatically from the speech signal, are respectively: the fundamental frequency (F_0), duration, and signal amplitude. The fundamental frequency refers to the value of the lowest frequency-component of a speech waveform, mostly influenced by the vibrations of the vocal folds. Duration is the relative length of a speech sound. Loudness is the perceived volume of a speech sound and is mostly correlated with descriptors of signal amplitude, such as energy and intensity. Apart from pitch, duration, and loudness, aspects related to voice quality, such as creaky, breathy, whispery or lax speech, are also considered to be prosodic, with their main acoustic correlates being jitter, shimmer, and harmonic-to-noise ratio.

Prosodic features are not considered absolute characteristics of a single speech unit, but rather they describe relative differences. For instance, duration can vary depending on whether the speaker is speaking faster than usual at a particular moment in time. Prosody is used for a wide range of purposes in human-to-human communication, including, disambiguation of syntactic structures, marking of contrastive emphasis or focus, indication of the speech act of an utterance, and expression of the speaker's emotions and attitudes (Wagner and Watson, 2010; Hirschberg, 2002). Two aspects of speech prosody widely studied in linguistics are prosodic prominence and prosodic phrasing.

Prosodic prominence

A speech unit (phoneme, syllable, word, etc.) is prosodically prominent or stressed when it stands out from neighbouring units by differences in pitch, duration, or loudness (Terken and Hermes, 2000). Unlike lexical stress, whose main purpose is to help listeners distinguish between the identity of different words with equal pronunciations, prosodic stress is mostly concerned with how different stress levels are assigned to the different words in an utterance, to make particular words more prominent than others.

Speakers can make a word or syllable prominent in order to perform different communicative functions in spoken language (Hirschberg, 2002; Wagner and Watson, 2010). Among these, prosodic prominence is used to convey the information structure of the discourse. This covers aspects such as focus, emphasis, contrast, giveness, and topicality (Krifka, 2008). For example, prosodic prominence can be used to alter the meaning of utterances. Consider, for instance, the utterance *"I didn't use your laptop yesterday"*.

By stressing *"I"*, the speaker may want to emphasise that the person who used the laptop was somebody else, while stressing *"yesterday"* would indicate that the speaker did use the laptop but on a different day.

Besides intent-related clarifications, prosodic prominence may also be used in a conversation to highlight words that include new or previously non given information (Prince, 1981; Hirschberg and Grosz, 1992; Röhr, 2013). The general trend is that words carrying new information are more likely to be accented, while words that present old or redundant information to the topic being discussed are more likely to be de-accented. Additionally, there is evidence that more frequent or predictable words, as well as function words and subsequent repetitions of content words, have shorter de-emphasised pronunciations (Bell et al., 2009; Röhr, 2013). Previous research has pointed out that there are exceptions to all these trends (Hirschberg, 2002; Wagner and Watson, 2010), principally because information structure can be also conveyed by other means, for instance, by the grammar and position of words in a sentence (Terken and Hirschberg, 1994). Despite this, the idea that prosodic information may help to signal informative words is appealing for tasks such as SCR, where commonly only lexical information is used to identify words that are descriptive of the topic of a document.

Prosodic phrasing

Prosodic grouping or phrasing refers to the strength with which speech units are separated and on how these units are structurally organised in speech. The grade of disjuncture between speech units characterises a distinctive boundary type. A common list of boundary types in increasing order of strength is: phoneme, syllable, foot, phonological word, intermediate phrase, intonational phrase, and utterance (Selkirk, 1984). Speech units can then be arranged in a phonological hierarchy according to their prosodic boundary types, in a similar way words, clauses, and sentences can be arranged in a syntactical hierarchy. The set of acoustic features found to correlate with the strength of prosodic boundaries are excursions in F0 around the boundary, lengthening of the last syllable preceding the boundary, the presence and length of pauses, and intensity (Wagner and Watson, 2010).

In human-to-human communication, prosodic phrasing is used to disambiguate semantically ambiguous utterances in read and spontaneous speech (Lehiste, 1973; Cutler et al., 1997; Hirschberg, 2002). For instance, the utterance *"When Roger leaves the house it is dark"* can convey different meanings depending on the position where prosodic boundaries are placed. Making a pause between *"leaves"* and *"the"* would imply that the house is dark, whilst a pause between *"house"* and *"it"* would indicate that Roger left the house in the night. More importantly for SCR, prosodic phrasing has been shown helpful in practice for identifying topic structure and sentence boundaries in spoken content (Shriberg et al., 2000; Kolář et al., 2006; Malioutov et al., 2007). Prosodic boundaries could therefore be used by a SCR system as an additional source of evidence for content structuring and to identify potential playback entry points to return to the user.

3.3.2 Prosody and informativeness

The relationship between the prominence of spoken words and their level of "informativeness", that is, the extent to which words are significant and descriptive of the information conveyed in speech, has been studied in previous research. This section describes previous work that has explored the correlation between prominent and informative words in the context of SCR applications, where the identification of informative words plays a major role.

Prominence and BM25 weights

Silipo and Crestani (2000); Crestani (2001) investigated the relationship between prosodic prominence and BM25 weights of terms in the OGI Stories Corpus of telephone conversations (Muthusamy et al., 1992). In this study, the authors utilised 144 telephone calls each containing roughly 60 seconds of spontaneous speech produced by speakers of American English. Two trained linguists labelled every spoken syllable in the recordings as either containing a primary prominence stress, an intermediate stress or the absence of stress; these events were given numeric values of 1, 0.5, and 0, respectively. A stress score was then defined for a word mention as the sum of the stress scores from the word's syllables. Next, the overall stress score of a word in a call was defined as the average stress across all the occurrences of the word in the call.

Averaged stress scores of words were then compared against weights computed by the Okapi BM25 function, with collection frequency estimates calculated based on word occurrences across the entire corpus. From this comparison, Silipo and Crestani found that, in general, words with high (low) BM25 weights also tend to have high (low) stress scores. The analogous case, this is, that highly stressed words are associated with high BM25 weights could not be reliably determined by the authors due to the coarse granularity of the stress annotations used in the study. Despite this, Silipo and Crestani's work suggests that prosodic features may have the potential to identify acoustic "keywords" in the spoken content, and that this information could potentially be exploited in SCR to create a more effective index of spoken documents.

In the context of IR, a term is considered important or informative in a document if the term is significantly associated with the topic of the document, and if it is effective in discriminating this topic from others. If it is true that the most prominent spoken terms are those that best describe the topic in discourse, an SCR system could then exploit this fact to generate better estimates of the weight that a term is given for a document. In other words, terms that are made prominent could be considered more representative of the topic of the document, and hence given increased weights in the relevance scoring process.

Prominence and word importance

In a more recent study, Ward and Richart-Ruiz (2013) investigated the correlation of manual annotations of importance and prosodic features in a subset of 100 minutes from the Switchboard corpus of telephone conversations (Godfrey et al., 1992). In this study, annotators were asked to select and label short speech intervals according to a 5-point scale of importance. Each interval was then labelled as being highly important (5), typically important (4), less important (3, 2 and 1), and silence (0). Next, loudness, F_0, F_0 range, and speaker-rate were extracted from the speech signal considering windows of various widths, resulting in a sequence of 78 dimensional feature vectors along the timeline of a speech recording. Principal component analysis (PCA) was then performed to map all feature vectors into a reduced dimensional space, with the purpose of gaining further insight into the association between dialogue events and prosodic information (Ward and Vega, 2012). A linear regression model trained with this data was then used to predict levels of importance for unseen speech data. Predictions from this regression model were found to correlate well (Pearson's $\rho = 0.83$) with manual annotations of importance. These results provide further evidence about the apparent relationship between informative content and prosody. In particular, this study suggests that the relative importance of words in a speech stream may be predicted using a linear combination of prosodic features extracted from the speech signal.

Prosody in speech summarisation

Speech summarisation is concerned with producing self-contained abstracts of spoken documents (Furui, 2007). Ideal summaries are those which only retain the important information conveyed in the original document without including redundant material. Standard approaches to speech summarisation use ASR technology to convert speech into text. The resulting speech transcript is then segmented into a collection of sentences from where important sentences are selected to be included in the document summary. Previous work on automatic speech summarisation has made use of prosodic information for both the identification of sentence boundaries and the selection of important sentences. Chen and Withgott (1992) trained a HMM on hand labelled data to detect emphasised speech regions based on pitch and energy features. Emphatic speech regions predicted by the model were then extracted to generate speech summaries.

In the work described in (Koumpis and Renals, 2005), important words were extracted from short voice mails by using a binary classifier trained to predict whether an individual word was worth including in the summary. The classifier was trained on a set of lexical and prosodic features for each word, including ASR confidence scores, duration, length of pauses before/after the word, energy, and F_0 derived features. An analysis of the discriminating power of each individual feature indicated that lexical features were more useful than prosodic features at detecting important words, yet prosodic features were found useful when combined with lexical descriptors.

Maskey and Hirschberg (2005) used sentence classifiers instead of word classifiers to determine which sentences from the document should be included in a summary, and compared the predictive power of lexical, prosodic, structure, and discourse (giveness) features. An important conclusion drawn from this work is that a model trained with prosodic plus structural features can perform comparably to a model trained with lexical features alone, indicating that "the importance of what is said correlates with how it is said" (Maskey and Hirschberg, 2005). Again, the most effective model was obtained when the classifiers were trained on a combination of lexical and prosodic features.

In similar work, Xie et al. (2009) investigated the utility of prosodic features for automatic speech summarisation of meeting recordings. The summarisation approach adopted in this case consisted of classifying individual sentences, each represented by a vector of lexical and speaker-normalised acoustic features. Xie et al. concluded that models trained with prosodic features outperformed models trained with lexical features only. Additionally, normalisation of acoustic features based on speaker, topic, and local context proved to be more effective than using raw, unnormalised, acoustic features. In a more recent study, Jauhar et al. (2013) report that a random walk based approach can produce better summaries of academic meetings when using prosodic features than when only using lexical information.

Overall, besides their demonstrated effectiveness in topic segmentation, prosodic features have been found valuable at detecting important words or sentences in spoken content. These findings are consistent with previous studies of the relationship between prominent and informative words, suggesting that speakers tend to characterise content bearing keywords with particular acoustic patterns.

3.3.3 Prosody and ASR errors

Recently, Goldwater et al. (2010) conducted a major empirical study that investigated what kind of features may characterise spoken words that are hard to be recognised by an ASR system. In this study, two state-of-the-art LVCSR systems were used to transcribe a corpus of conversational speech. Then, prosodic, lexical, and disfluency features were extracted for each individual word in the corpus with the goal of analysing their influence over WER. WER was calculated for a particular group of words of interest by counting the number of times that each word in the group contributed to an ASR error.

Statistical analysis performed with this data indicated that words pronounced with atypical prosody are more likely to be misrecognised. The analysis reflected that words pronounced with extreme intensity and pitch values, that is, extremely high or low intensity and pitch values, are associated with high error rates. Interestingly, words with a wide intensity range were associated with low WER, while those with large pitch range were predictive of high WER. Furthermore, words with lower than average duration or with extremely high large duration were correlated with high WER. Other prosodic factors that were analysed included speech rate and jitter. Both were found to be correlated with high WER for extremely high or low values of the features. Among the various lexical factors that were analysed in the study, it was found that content words and frequent words are easier to recognise than those that are rare and less predictable.

In a follow up study, Stoyanchev et al. (2012) trained binary classifiers with lexical and prosodic features to detect misrecognised words in transcripts of dialogue speech. They showed that detection accuracy can be significantly improved if prosodic features are used in combination with lexical and confidence scores calculated by the ASR. The research findings described above suggest that prosodic information can be used to predict regions of speech that are likely to be misrecognised by the ASR. This information could potentially be exploited in SCR to decide when it is worth using alternative hypothesis from the ASR or when it is worth selecting alternative word hypothesis from the ASR lattice. This idea was partially explored by Stoyanchev et al. (2012) in a spoken term detection (STD) task, which consisted of identifying where and when a given term is mentioned in a spoken document. Most STD approaches try to exploit multiple hypothesis generated by the ASR by traversing the ASR lattices. In this study, duration, pitch, and intensity features were found to provide considerable improvements in performance for the STD task.

3.3.4 Previous attempts to use prosody in SCR

Following on Silipo and Crestani's (2000) findings, Chen et al. (2001) and Guinaudeau and Hirschberg (2011) experimented with various methods that sought to exploit the prosodic prominence of spoken terms to improve retrieval effectiveness in different speech retrieval tasks. These methods follow a similar approach which can be summarised as follows.

First, every term mention within a document was assigned an prominence/acoustic score. This score is generated from a combination of prosodic features extracted from the speech signal, and is assumed to reflect the grade of relative salience of the term mention in the context where it is pronounced. Among the features used for this purpose were signal magnitude and other correlates of intensity (Chen et al., 2001; Guinaudeau and Hirschberg, 2011), pitch (Guinaudeau and Hirschberg, 2011), duration (Chen et al., 2001), and ASR confidence scores (Chen et al., 2001). Prominence scores were computed in an unsupervised ad-hoc fashion, by making assumptions about how features should characterise prominent terms, such as "terms pronounced louder and with an expanded pitch range are prominent" (Guinaudeau and Hirschberg, 2011) or "terms pronounced louder, longer, and clearer are prominent" (Chen et al., 2001). Alternatively, the scores can be learnt from speech corpora annotated with levels of prominence by using supervised learning techniques (Guinaudeau and Hirschberg, 2011; Mishra et al., 2012; Christodoulides and Avanzi, 2014). In addition, the acoustic scores from multiple mentions of a term in a document may be optionally combined into a single score representing how prominent the term is for the document. To achieve this, Guinaudeau and Hirschberg averaged the scores across all term occurrences in a document or alternatively retained their maximum value.

Second, the combined or individual prominence scores were incorporated into the computation of the terms TF-IDF weights within a vector space model (VSM) for text retrieval. Different term weighting schemes were proposed. Chen et al. (2001) computed term weights with a standard IDF factor and a modified term frequency factor in which the term counts in a document were replaced by the sum of its acoustic scores. Instead, Guinaudeau and Hirschberg (2011) calculated the prosodic-based weights as the weighted sum between a standard TF-IDF weight and the acoustic score associated with the term. In both variations of weighting schemes, the objective was to increase the weight of terms that are prosodically prominent in the spoken document. Terms that are given increased weight in a document contribute more significantly to the document's relevance score, therefore promoting the final rank of this document in the result list.

Chen et al. (2001) carried out SDR experiments with prosodically-enhanced weights over a Mandarin Chinese subset of broadcast news speech recordings, from the Topic Detection and Tracking corpus (TDT-2 and TDT-3). In these experiments, newswire articles in Mandarin Chinese were used as queries, thus the task was more akin to a query-by-example task, where queries are much longer than in conventional SDR tasks. Retrieval experiments by using the prosodically-enhanced weighting scheme provided small, although not statistically significant, improvements over purely lexical-based weights.

In the experiments conducted by Guinaudeau and Hirschberg (2011), prosodically-motivated weights were evaluated in the context of a topic tracking task over a collection of broadcast news recordings in French. In this context, topic tracking refers to the task of finding links between speech segments that describe similar information, where the similarity between segments is usually computed with text-based retrieval models. Because Guinaudeau and Hirschberg used speech segments as queries, their experiments with prosodic-based weights were also akin to a query-by-example task. A comparison of two methods for computing prominence scores was made. Scores computed in an unsupervised ad-hoc fashion provided larger improvements than scores predicted by a supervised model trained with annotations of prosodic prominence. However, both types of acoustic scores provided significant improvements in the topic tracking detection task when combined with lexical-based TF-IDF weights.

In addition to the above studies, Ward et al. (2015), and Galuščáková and Pecina (2014a) tried to improve retrieval effectiveness in query-by-example tasks by finding spoken segments whose prosody is similar to that of a spoken query. These approaches differ in purpose from the prosodically-enhanced weighting schemes proposed by Chen et al., and Guinaudeau and Hirschberg. While Chen et al. and Guinaudeau and Hirschberg's methods seek to use prosodic information to enhance the retrieval of spoken content that is topically related to the query, approaches based on prosodic similarity seek to retrieve speech content in which speakers show similar emotions, attitudes, or intents.

3.4 Summary

This chapter reviewed previous research in SCR. Emphasis was given to techniques that attempt to tackle one of the main challenges of the field: the presence of ASR errors in the documents, the structuring of speech content to reduce access time to relevant information, and the exploitation of acoustic/prosodic information to improve SCR components.

Early work in the field explored the use of word spotting techniques to search for query keywords in small collections of voice mails and private collections. Large vocabulary continuous speech recognition (LVCSR) systems enabled the indexing of speech collections by using a large number of indexing terms, which removed the limitations of previous approaches based on word spotting, and permitted the retrieval of speech content from an unrestricted set of query terms.

The TREC SDR benchmarks provided researchers with a common framework for evaluating SCR techniques on large collections of radio and TV broadcast news. Research within TREC SDR studied the robustness of existing text retrieval techniques when applied to retrieve relevant news stories from noisy speech transcripts. The main conclusion drawn from TREC SDR campaigns was that query and document expansion techniques, mainly when selecting expansion terms from an external source of text data, helped reduce the negative effects that ASR errors have on retrieval effectiveness and could significantly reduce the performance gap between using perfect and noisy transcripts. In the last editions of these benchmarks, researchers experimented with various content structuring techniques, including overlapping sliding windows and Heart's TextTilling algorithm, and observed that the former is more effective for SCR.

Subsequent research in SCR investigated the retrieval of short excerpts of spoken content from conversational spontaneous speech collections, which were shown to be substantially more difficult than broadcast news speech for both recognition and retrieval. Research in this period was driven by various evaluation campaigns that benchmarked SCR systems over different speech types, genres and domains: interviews in the CLEF-CL-SR campaigns; internet and broadcast TV content in the MediaEval campaigns; and academic lectures in the NTCIR initiatives. Researchers evaluated several SCR techniques during this period. Notable work includes the comparison between windowing and automatic text segmentation for content structuring, post-retrieval adjustment of jump-in points, multi-field representations of documents to exploit additional document metadata, passage re-scoring based on document-level relevance scores, and relevance density estimation based on query term proximity.

A frequent observation within previous work in SCR is that simple windowing approaches, in combination with the application of passage recombination or filtering to remove near duplicates, provide increased retrieval effectiveness compared to passages defined via text segmentation methods such as C99, TextTilling, or MinCut. Other sophisticated segmentation approaches that do not use a pre-segmented collection and that instead perform segmentation at querying-time, or that seek to predict putative topic boundaries based on machine learning techniques have generally performed poorly compared to windowing approaches.

In addition to content structuring and expansion techniques, a fair amount of research has explored the utilisation of acoustic/prosodic information to improve the overall quality of SCR systems. Besides observed improvements on topic segmentation quality, prosodic information has been used in the past to identify emphatic words in spoken documents and detect words that are likely to be misrecognised by ASR systems. In relation to the former, much of the work done has studied the relationship between prosodically emphasised words and their level of importance or informativeness in spoken documents. This work has mainly been motivated by research in linguistics and speech prosody, which suggests that important words that are new, focused, and/or unpredictable are more likely to be accented than others.

Research on speech summarisation, plus additional studies about the relationship between acoustic stress and degrees of word importance suggests that prosodic information may encode meaningful information that could be potentially useful to characterise words by distinct degrees of informativeness. Based on these observations, researchers have proposed and tested alternative term weighting schemes for SCR that increase the score of words that are made prominent in speech. These experiments showed mixed results in two query-by-example search tasks over broadcast news speech in French and Chinese, and it is therefore unclear whether prosodically-motivated term weights could be effectively used in SCR to enhance the quality of search results in other speech genres and languages.

As in most of the empirical research reviewed in this chapter, a test collection consisting of spoken documents, queries, and relevance assessments were used for the experimental work of this thesis. The next chapter describes these test collections in more detail.

Chapter 4

Materials and Test Collections

Advances in SCR research would not have been possible without extensive and rigorous experimentation. Most of the research presented in Chapter 3 reflects this and shows that SCR research, and IR research in general, is mainly of an empirical nature. An essential component of empirical research is datasets in which certain aspects of interest are observed and quantified for the purpose of validating one or more research hypotheses. In the case of IR research, the use of a "test collection" has become standard practice for validating the effectiveness of new methods and making comparisons against established techniques. A test collection is often cited as being comprised of three key elements: a set of documents, a set of queries representing information needs, and relevance data for pairs of queries and documents. This chapter presents the datasets and test collections used in the experimental work of this PhD. Each test collection is characterised in terms of its constituent documents, queries, and relevance assessments. Additionally, the description of each test collection includes an overview of the various manual and automatic transcripts available for each speech collections.

Two speech collections were used in this PhD. The BBC collection of TV programmes (Eskevich et al., 2013b, 2014) and the Spoken Document Processing Workshop (SDPWS) collection of academic presentations (Akiba et al., 2008). The BBC collection contains audiovisual material from TV shows broadcast by the British Broadcasting Corporation (BBC). Most of the material from this collection contains professional multiparty speech produced by native speakers of English. Although most of its speech content is scripted, the high diversity of the BBC material, which includes movies, TV-series, documentaries, broadcast news, talk shows, and sports events, makes it a challenging test collection for SCR. The SDPWS collection contains academic presentations produced by native speakers of Japanese. The monologues from this collection cover a range of scientific topics, including subjects in computer science and speech technology. Because of the nature of its spoken content, the material from the SDPWS collection can be considered more spontaneous and somewhat more homogeneous in terms of domain and genre compared to the BBC material.

The remainder of this chapter describes in more detail the BBC and SDPWS collections. These are described in Sections 4.1 and 4.2 respectively.

4.1 The BBC collection of TV content

The BBC collection contains recordings of TV programmes in English broadcast for the UK audience in mid 2008. Much of this material was originally compiled to support research on video access and retrieval applications as part of the “Access to Audiovisual Archives” (AXES) project[1]. Although the complete dataset is not publicly available, various subsets of the collection were distributed to the participants of the MediaEval Search and Hyperlinking (SH) and Search and Anchoring in Video Archives (SAVA) tasks (Eskevich et al., 2013a, 2014, 2015).

4.1.1 Overview

The BBC collection consists of 5,843 recordings of TV shows, comprising a total of 4,322 hours of audiovisual material. For each programme recording, separate streams of video and audio are available plus additional data in text format, including titles, descriptions, synopsis, cast, subtitles, and various automatic transcripts produced by different ASR systems. Shots from commercials and breaks were removed from the recordings by the providers of the BBC data and are thus not present in any of videos from the dataset.

The programmes in the collection include multiple episodes of 872 shows broadcast on the channels BBC One, BBC Two, BBC Three, and BBC Four between April and July 2008. The collection is thus comprised of shows from a wide variety of formats and genres. The genres include news, series, soap operas, talk shows, reality shows, game shows, interviews, documentaries, sport events, comedy, cookery, cartoons and films. The speech material is thus highly diverse in terms of style, register, domain, background noise, and number of speakers who produced it.

The 5,843 recordings are split into two collections covering two time periods of TV-broadcast across the four channels of the BBC. The first of these (BBC1) contains 2,323 files and was used as training and testing data in the Search & Hyperlinking 2013 (SH13) task. Among these, there are 463 duplicate files corresponding to a subset of the TV shows that were re-broadcast by the BBC in the time period when the recordings were collected. In the Search & Hyperlinking 2014 (SH14) and Search and Anchoring in AudioVisual Archives (SAVA) tasks, the BBC1 collection was cleaned of duplicates and used as training data, while the remaining 3,520 of the videos (BBC2) were used as testing data.

Table 4.1 provides general statistics about the recordings from the BBC1 and BBC2 collections. As the table shows, the programmes are not only diverse in terms of genre and domain, but also vary substantially in terms of duration. Programmes presenting news and weather highlights can last 3 minutes or less while shows covering the results

[1] `http://www.axes-project.eu/`

		Duration				
Dataset	**Documents**	**Total**	**Avg.**	**S.D.**	**Min**	**Max**
BBC1	1860	1335 hrs	43 min	39 min	44 secs	6 hrs 23 min
BBC2	3520	2648 hrs	45 min	43 min	3 min	10 hrs 35 min

Table 4.1: Duration statistics of videos in the BBC collection after removing duplicates.

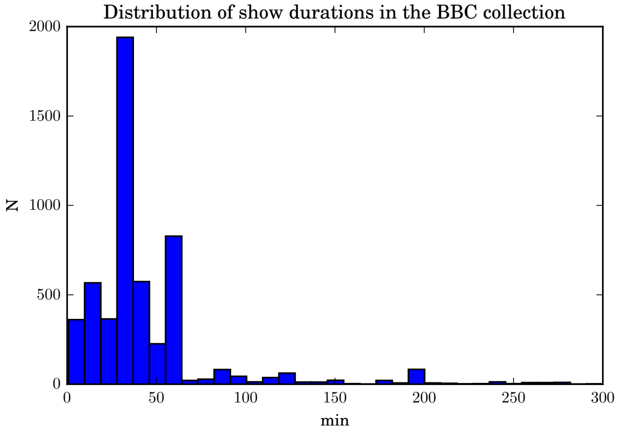


Figure 4.1: Distribution of video durations in the BBC collection.

of an election or certain sports events can easily extend up to 6 or 10 hours. Figure 4.1 shows the distribution of programmes by duration in the dataset. As is standard in TV broadcasts, most programmes are 30 or 60 minutes long.

The remainder of this section describes in detail the characteristics of the BBC1 and BBC2 collections. Since the main focus of this thesis is the study of SCR techniques in collections where most of the information is encountered within the spoken stream, the following description only covers aspects that are relevant to the processing of the speech material and text transcripts for ASR and SCR purposes. So, despite the visual nature of the BBC content, the scope of this thesis is only on the exploitation of the spoken content.

4.1.2 Speech collection and transcripts

This section overviews the characteristics of the speech recordings of the BBC collections as well as the set of transcripts available. Both manual and automatic transcripts of the BBC shows are available.

Manual transcripts

Subtitles generated by the BBC for the hearing impaired are available for every programme. These contain utterance transcriptions in ASCII and their timestamps, as well as other metadata such as indications of music and sounds played. Speaker identities are also indicated in the subtitles by different RGB colour codes, normally used when the captions are displayed in a TV broadcast. Although the subtitles contain manually curated text, they are not a verbatim transcription of the spoken material. This is because during the creation of closed-captions long phrases are sometimes shortened or rephrased by the transcribers to minimise read-time[2]. Also, utterance timestamps were set to abide by display constraints driving read-time broadcast and as such they can only be considered approximations of the true times when an utterance is produced.

ASR transcripts

The spoken content from the BBC1 and BBC2 collections was automatically transcribed by different ASR providers. In order to prepare the audio material for speech recognition, the audio track from each video file in the collections was first extracted by the dataset creators with the *ffmpeg*[3] tool. The audio was originally encoded in Vorbis with a sample rate of 48 kHz and 16 bits of precision in stereo format. This was later uncompressed into WAV format, down-sampled to 16 kHz, and reduced into a single channel to comply with the specifications of the ASR providers. The audio was subsequently processed by three providers: LIMSI-CNRS/Vocapia[4] (Gauvain et al., 2002), LIUM (Rousseau et al., 2011), and NST-Sheffield (Lanchantin et al., 2013).

Some of the research questions explored in this thesis require the analysis of acoustic information at the level of individual words, which in turn requires word-level timestamps to be available from the output of the ASR systems. Because LIUM transcripts did not always contain word-level time information, the experiments presented in this thesis with the BBC collections were restricted to LIMSI and NST transcripts only. What follows is a brief description of the main characteristics of LIMSI-CNRS/Vocapia and NST-Sheffield recognition systems.

The LIMSI-CNRS/Vocapia ASR system

The LIMSI-CNRS/Vocapia system is an enhanced version of the LIMSI-CNRS broadcast news transcription system (Gauvain et al., 2002), which has been under constant development since the late 1990's. The specific version of this system used to transcribe the BBC speech collections at the SH13, SH14, and SAVA tasks corresponds to the VoxSigma vrbs_trans system (version eng-usa 4.0), with models updated with support from the Quaero programme (Gauvain, 2010). The modelling techniques used in this system are described in (Lamel et al., 2011).

To cope with the problem of acoustic variability, audio files are first partitioned into homogeneous segments with speech samples produced by a single speaker. This partition step permits the system to perform adaptive recognition of speech fragments, plus the identification of speaker turns, identities, and gender. Audio segmentation is performed by

[2] For examples see `http://bbc.github.io/subtitle-guidelines/`

[3] `https://www.ffmpeg.org/`

[4] `http://www.vocapia.com`

an agglomerative clustering algorithm that iteratively classifies and merges audio segments based on a set of Gaussian mixture models (GMMs) (Gauvain et al., 1998).

For front-end processing, speech frames are represented by a vector consisting of 39 cepstral features derived from 12 linear prediction cepstral coefficients (LPCC) and a log-energy estimate plus their first and second derivatives. This feature vector is subsequently extended with 39 additional acoustic features learnt via a bottleneck feed-forward neural network. The back-end component consists of continuous-density HMMs with gaussian mixture models (GMMs) for acoustic modelling. For language modelling, a 4-gram back-off language model is used, whose probabilities are further interpolated with those estimated by a neural language model trained on a large amount of broadcast news transcriptions and news articles. Recognition is performed in multiple decoding passes in which recognition lattices are re-scored based on the interpolated n-gram and neural LMs.

With the specifications reported in (Gauvain et al., 2002), previous versions of the LIMSI-CNRS/Vocapia system were able to transcribe English broadcast news speech with 13-20% WER. The more recent version of this system described in (Lamel et al., 2011) is reported to transcribe French broadcast conversational speech with 19% WER.

The NST-Sheffield ASR system

In the context of the Natural Speech Technology (NST) project, researchers from various universities across the UK and the BBC R&D department collaborated in the development of a new ASR system for transcribing spoken material from the BBC archive (Lanchantin et al., 2013). For the purpose of transcribing the BBC1 and BBC2 speech collections used at the SH14 and SAVA tasks, the organisers of Search & Hyperlinking tasks used a version of the NST system trained with a different subset of BBC material that does not overlap with the contents of the BBC1 and BBC2 collections from Table 4.1. In the absence of perfect transcripts, the system was trained on subtitles.

The NST system has two distinctive characteristics. First, as subtitles have imperfect word timestamps, the authors used a slightly supervised approach to obtain more accurate word timing information. In this procedure, the output of a first decoding pass produced by using a generic acoustic model and a domain specific language model is used to identify a subset of partially well-aligned utterances which are then used to retrain the parameters of the acoustic model. The method used to identify candidate utterances with high recognition probability consists of selecting the utterances whose 1-best hypothesis have low WER against the subtitles. This process is repeated for a number of iterations until the number of correctly recognised utterances converges to a fixed value. Second, deep neural networks (DNNs) pre-trained with out-of-domain data for phone classification were used to enhance the feature representation of the acoustic observations. The latter approach, called Multi-level Adaptive Networks (MLAN) showed increased recognition accuracy in cross-domain cross-genre experiments on broadcast TV material from the BBC (Lanchantin et al., 2013).

Unsurprisingly, experiments with these techniques reported in (Lanchantin et al., 2013) showed that the overall recognition accuracy of the system is highly dependent on the characteristics of the spoken content. For speech recorded in studio (controlled) conditions, the NST system can obtain WERs as low as 9.8%, whereas for non-studio recorded speech, such as parliamentary proceedings, the WERs may increase up to 20-23%. On the other extreme, transcriptions of TV drama series proved to be the most difficult with WERs as high as 50%.

Processing and indexing of English transcripts

In all experiments reported in this thesis with the BBC collections, English transcripts are pre-processed prior to being indexed. First, a list of recognised words are extracted from the 1-best recognition hypothesis of each speech segment. Second, the resulting text is processed and indexed by using the Terrier IR platform[5] v4.0 (Ounis et al., 2007).

Terrier provides different text processing modules that can be combined to define custom processing pipelines. In the experiments reported in this thesis, Terrier is configured to process English text as follows:

- Text is tokenised (class `UTFTokeniser`) and resulting tokens are lower-cased (property `lowercase=true`).
- Tokens present in Terrier's default stop-word list are discarded (class `Stopwords`).
- Tokens are subsequently stemmed using the Porter algorithm (Porter, 1980) (class `PorterStemmer`).
- Stems appearing in more than half of the documents in the collection are discarded (property `ignore.low.idf.terms=true`).
- UTF-8 support is enabled (`string.use_utf=true` and `trec.encoding=utf-8`).

Under this set-up, no special standardisation is applied to numbers, dates, URLs, and other special tokens that may appear in the transcripts. Terrier is then used to create a separate search index for each type of transcripts from the BBC1 and BBC1 collections. Tables 4.2a and 4.2b present document length statistics, measured in number of term occurrences, calculated from each of these indices. It should be noted that the number of indexed documents is lower than the number of TV episodes in cases where the transcript files from a particular ASR provider were not available in the dataset. In order to be consistent with the experimental setup used in the SH13, SH14, and SAVA tasks, two separate indices were generated for the BBC1 and BBC2 collections. Also, duplicate transcripts from re-broadcasts were not included in these indices.

[5] `http://terrier.org`

Table 4.2: Length statistics of BBC collections measured in number of term occurrences per transcript. In the tables, "S.D." stands for standard deviation.

(a) BBC1

Transcript	Documents	Avg. len.	S.D. len.	Max. len.	Min. len
SUB	1,856	3,118	3,921	46,952	28
LIMSI	1,860	3,161	3,355	36,312	29

(b) BBC2

Transcript	Documents	Avg. len.	S.D. len.	Max. len.	Min. len
SUB	3,517	3,327	4,098	33,967	5
LIMSI	3,520	3,383	3,548	39,538	50
NST	3,520	2,460	2,584	21,527	18

Quality of ASR transcripts for the BBC collection

Since SCR effectiveness is affected by recognition errors in the automatic transcripts, it is useful to report recognition rates of the transcribed material that is used for experimentation. In the case of the BBC1 and BBC2 collections, recognition rates cannot be estimated directly as no precise reference transcripts exist of the TV shows. As a point of reference, one can refer to the WER figures published by Lanchantin et al. (2013), briefly summarised in the description of the NST system, as well as those reported at the recent Multi-genre Broadcast (MGB) Challenge (Bell et al., 2015).

In the MGB challenge, participating ASR systems were evaluated over TV shows from the BBC archive that aired between April and May 2008, thus covering the same time-period than the episodes contained in the BBC1 and BBC2 collections used at the SH13, SH14, and SAVA tasks. Variations of the LIMSI and NST systems were evaluated at the MGB challenge, Table 4.3c provides a partial view of these results. The best participating systems at this challenge obtained WERs that varied between 10-50% depending on the genre of the shows being transcribed, with an average WER of 23-28%. Talk shows such as "Daily Politics" could be transcribed with WERs as low as 10.4%, whereas Drama series were found the most challenging, with WERs as high as 50.1%.

Cross-document term count differences among the entries of the subtitle and ASR indices may also provide some indication of the quality of the ASR transcripts. Tables 4.3a and 4.3b show various index similarity metrics, specifically, unique term error rate (UTER), term error rate (TER) (Johnson et al., 1999a), binary index accuracy (BIA), and ranked index accuracy (RIA) (van der Werff and Heeren, 2007) calculated for each automatic transcript index against the subtitle index. In contrast to WER, these index similarity measures disregard word ordering and have been shown to better reflect the potential impact that ASR errors may have on the retrieval effectiveness of SCR applications (van der Werff and Heeren, 2007). Appendix B includes definitions for all these metrics.

As a point of comparison for the figures from Tables 4.3a and 4.3b, van der Werff and Heeren (2007) report an average BIA and RIA of 0.51 and 0.70 respectively for ASR

Table 4.3: Recognition accuracy of BBC transcripts as measured by various index similarity metrics.

(a) BBC1

Transcript	Vocabulary	UTER	TER	BIA	RIA
SUB	59,033	0.00	0.00	1.00	1.00
LIMSI	36,464	0.31	0.93	0.36	0.45

(b) BBC2

Transcript	Vocabulary	UTER	TER	BIA	RIA
SUB	78,953	0.00	0.00	1.00	1.00
LIMSI	40,198	0.30	1.00	0.36	0.44
NST	33,450	0.35	0.92	0.35	0.45

(c) WER[a] reported at the MGB challenge (Bell et al., 2015) for selected shows.

Show	LIMSI	NST
Daily Politics	11.8%	13.6%
Top Gear	26.3%	27.2%
Oliver Twist	50.1%	49.4%
Overall WER[b]	27.5%	28.8%

[a] Figures are for similar systems to those used to transcribe the BBC collections and may not represent the real accuracy of these systems on the BBC1 and BBC2 collections.

[b] Averaged over the full list of shows shown in (Bell et al., 2015).

Table 4.4: Examples of transcripts for the show *Daily Politics*.

Type	Transcription
SUB	*Morning folks welcome to the Daily Politics. What should what can, the world do about Zimbabwe?*
LIMSI	*My morning thoughts folks. Welcome to the Daily Politics politics what should what can the world do about Zimbabwe.*
NST	*EVOLVES WELCOME TO THE DAILY POLITICS WHAT SHOULD WHAT AND THE WORLD DO ABOUT SINBAD WAY*

transcripts used at the TREC SDR track (Garofolo et al., 2000). Compared with these numbers, the BIA and RIA values shown in Tables 4.3a and 4.3b for the BBC transcripts are substantially lower, suggesting that the BBC collection may be more challenging for SCR than that used at the TREC SDR track. The figures from these tables also indicate that transcription quality is similar for BBC1 and BBC2 datasets and that the NST and LIMSI systems attain similar BIA and RIA. It must be pointed out that, unlike NST models, the models used by the LIMSI system were not trained on BBC material. Tables 4.4, 4.5, and 4.6 show examples of transcripts for three shows of the BBC2 collection.

4.1.3 Topics

Topics for the BBC1 and BBC2 collections were collected in different user studies carried out by the SH and SAVA organisers (Aly et al., 2013b; Eskevich et al., 2014, 2015).

Table 4.5: Examples of transcripts for the show *Top Gear*.

Type	**Transcription**
SUB	*We've got a gong for Best Factual Programme, which is astonishing when you think we haven't actually put a fact in the show for the past five years.*
LIMSI	*we, we've got a gong. {fw}. For the best factual program programme which is astonishing. When you think we haven't actually put got a fact in this over the last 5 years.*
NST	*WE ARE WE'VE GOT TO GO ALONG FOR THE BEST FACTUAL PROGRAMME WHICH IS ASTONISHING WHEN YOU THINK WE HAVEN'T ACTUALLY GOT A FACTOR IN THIS OVER LAST FIVE YEARS*

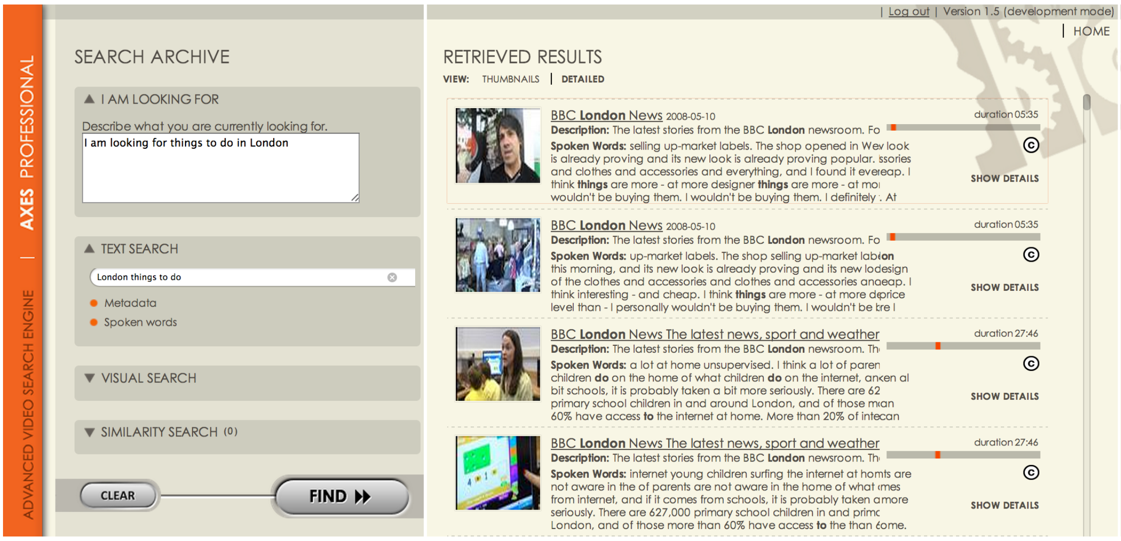
Table 4.6: Examples of transcripts for the show *Oliver Twist*.

Type	**Transcription**
SUB	*Oh, Mr Fagin, let's not muddy the waters with reasons and motives motives. Very good, sir. The issue is clear. He must hang. It's easy enough to get a pauper child hung.*
LIMSI	*Mr. Fagan, let's not muddy the waters as reasons and notice favorites. The issue is clear in this time is sees enough to get a pulpit child Charles hunt Hun Hunt.*
NST	*MR FECKLESS NOT MUDDY THE WATERS AS REASONS AND MOTIVES SHE WAS CLEAR IN THE SAND SUZIE ENOUGH TO GET A PAUPER CHILD HUM*

50 known-item topics were collected for the BBC1 collection, while two groups of ad-hoc topics, 36 in SH14 and 30 in SAVA, were collected for the BBC2 collection. As mentioned previously, known-item topics are those that target a single relevant item from the collection and are generally thought as being formulated by someone who partially recalls the content for which they are searching for. By contrast, ad-hoc topics represent an information need for which there could be more than one item deemed relevant in the collection. The remainder of this section provides details about these two topic sets.

SH13 topics for the BBC1 collection

The known-item topics for the BBC1 dataset were generated in a user study described in (Aly et al., 2013b). The study involved 30 participants aged 16-30 from London, UK, selected as a typical group of "home users" who frequently use search engines and watch TV over the Internet. Participants were set within a home-user search scenario where they were asked to search for BBC material that would be entertaining or interesting for them. The study first required participants to use the AXES video search system (McGuinness et al., 2013) to browse the videos and to get familiar with the contents of the collection. Figure 4.2 shows a screen capture of the system used in the study. Participants were then asked to select a segment of video from the collection that they considered interesting and generate a text query that could be used to re-find the segment if using the search system again. Users had to provide specific starting and ending times for each selected segment by using the UI shown in Figure 4.3. Furthermore, as the AXES system could perform search based on visual concepts, users were also asked to provide an additional



Describe what you are currently looking for.
I am looking for things to do in London

London things to do

BBC **London** News 2008-05-10 duration 05:35
Description: The latest stories from the BBC **London** newsroom. Fo
Spoken Words: selling up-market labels. The shop opened in Wev look is already proving and its new look is already proving popular. ssories and clothes and accessories and everything, and I found it evereap. I think **things** are more - at more designer **things** are more - at moı wouldn't be buying them. I wouldn't be buying them. I definitely : At

BBC **London** News 2008-05-10 duration 05:35
Description: The latest stories from the BBC **London** newsroom. Fo
Spoken Words: up-market labels. The shop selling up-market lab**on** this morning, and its new look is already proving and its new lodesign of the clothes and accessories and clothes and accessories anaeap. I think interesting - and cheap. I think **things** are more - at more deprice level than - I personally wouldn't be buying them. I wouldn't be kre I

BBC **London** News The latest news, sport and weather duration 27:46
Description: The latest stories from the BBC **London** newsroom. Th
Spoken Words: a lot at home unsupervised. I think a lot of paren children **do** on the home of what children **do** on the internet, anken al bit schools, it is probably taken a bit more seriously. There are 62 primary school children in and around London, and of those mıan 60% have access **to** the internet at home. More than 20% of intecan

BBC **London** News The latest news, sport and weather duration 27:46
Description: The latest stories from the BBC **London** newsroom. Th
Spoken Words: internet young children surfing the internet at homts are not aware in the of parents are not aware in the home of what ımes from internet, and if it comes from schools, it is probably taken amore seriously. There are 627,000 primary school children in and primc London, and of those more than 60% have access **to** the than ćome.

Figure 4.2: The search interface of the AXES system used by SH13 organisers in the topic-generation study.

Table 4.7: Example of SH13 known-item topics for the BBC1 collection.

Topic	Query	Visual cues
SH13-10	*new rules for qualified drivers statistics of injuries and casualties on the roads amongst young drivers*	*statistics*
SH13-18	*What does a ball look like when it hits the wall during Squash*	*ball hitting a wall in slow motion*
SH13-19	*how much gas do cows produce*	*cows at a farm*
SH13-31	*rhinos and lions in kenya*	*fields in africa lions wildlife*
SH13-38	*little britain comedy sketch prime minister moustache*	*prime minister moustache*

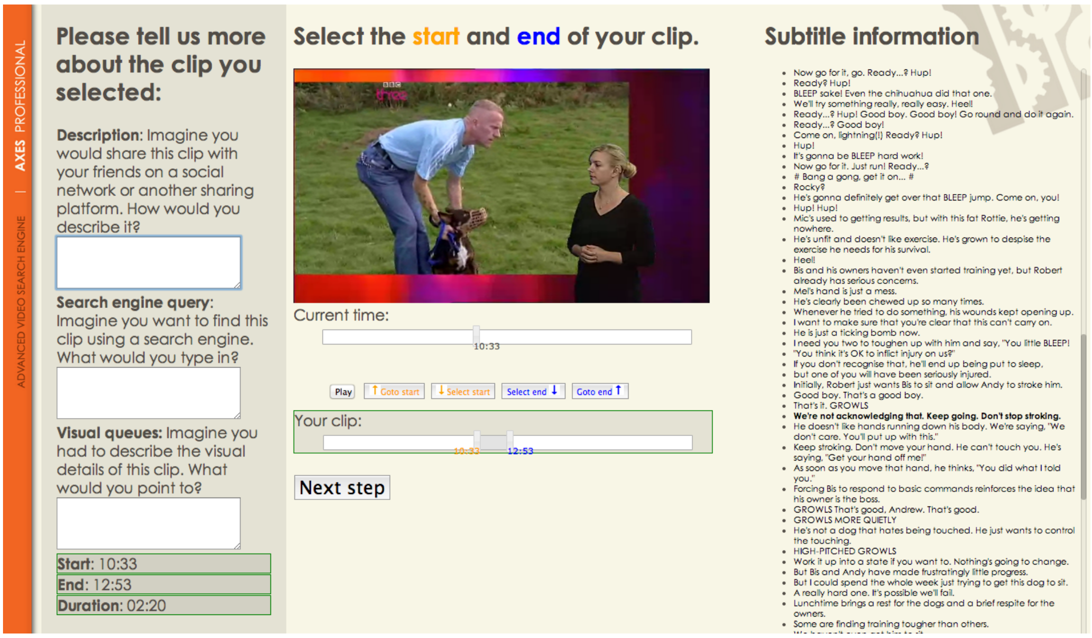
list of visual cues or keywords in order to complement their original queries. 50 topics were collected in total and used for testing in the SH13 task (Eskevich et al., 2013a).

Table 4.7 shows some examples of these topics. In contrast to the topics that are commonly used in TRECVid[6] tasks, those from the SH and SAVA tasks were sought to be multimodal, in the sense that they often target information that could be present in multiple modalities within the content to be searched, including the spoken, visual, or textual (metadata) modalities. As can be seen from the examples, some topics are informational (SH13-10 and SH13-19), while others are more visually oriented (SH13-18 and SH13-31). Also, it is common for users to include names of celebrities and programme's titles (SH13-38) in their queries.

SH14 topics for the BBC2 collection

The ad-hoc queries used at the SH14 task were gathered in a similar user study (Eskevich et al., 2014) as those for SH13. In this study, 28 participants were recruited, with a

[6] https://trecvid.nist.gov/



Please tell us more about the clip you selected:

Description: Imagine you would share this clip with your friends on a social network or another sharing platform. How would you describe it?

Search engine query: Imagine you want to find this clip using a search engine. What would you type in?

Visual queues: Imagine you had to describe the visual details of this clip. What would you point to?

Start: 10:33
End: 12:53
Duration: 02:20

Select the start and end of your clip.

Current time:

Your clip:

Subtitle information

- Now go for it, go. Ready...? Hup!
- Ready? Hup!
- BLEEP sake! Even the chihuahua did that one.
- We'll try something really, really easy. Heel!
- Ready...? Hup! Good boy. Good boy! Go round and do it again.
- Ready...? Good boy!
- Come on, lightning(!) Ready? Hup!
- Hup!
- It's gonna be BLEEP hard work!
- Now go for it. Just run! Ready...?
- # Bang a gong, get it on... #
- Rocky?
- He's gonna definitely get over that BLEEP jump. Come on, you!
- Hup! Hup!
- Mic's used to getting results, but with this fat Rottie, he's getting nowhere.
- He's unfit and doesn't like exercise. He's grown to despise the exercise he needs for his survival.
- Heel!
- Bis and his owners haven't even started training yet, but Robert already has serious concerns.
- Mel's hand is just a mess.
- He's clearly been chewed up so many times.
- Whenever he tried to do something, his wounds kept opening up.
- I want to make sure that you're clear that this can't carry on.
- He is just a ticking bomb now.
- I need you two to toughen up with him and say, "You little BLEEP!
- "You think it's OK to inflict injury on us?"
- If you don't recognise that, he'll end up being put to sleep,
- but one of you will have been seriously injured.
- Initially, Robert just wants Bis to sit and allow Andy to stroke him.
- Good boy. That's a good boy.
- That's it. GROWLS
- **We're not acknowledging that. Keep going. Don't stop stroking.**
- He doesn't like hands running down his body. We're saying, "We don't care. You'll put up with this."
- Keep stroking. Don't move your hand. He can't touch you. He's saying, "Get your hand off me!"
- As soon as you move that hand, he thinks, "You did what I told you."
- Forcing Bis to respond to basic commands reinforces the idea that his owner is the boss.
- GROWLS That's good, Andrew. That's good.
- GROWLS MORE QUIETLY
- He's not a dog that hates being touched. He just wants to control the touching.
- HIGH-PITCHED GROWLS
- Work it up into a state if you want to. Nothing's going to change.
- But Bis and Andy have made frustratingly little progress.
- But I could spend the whole week just trying to get this dog to sit.
- A really hard one. It's possible we'll fail.
- Lunchtime brings a rest for the dogs and a brief respite for the owners.
- Some are finding training tougher than others.

Figure 4.3: The boundary refinement interface of the system used by SH13 organisers in the topic-generation study.

Table 4.8: Example of SH14 ad-hoc topics for the BBC2 collection.

Topic	**Query**	**Description** ***(I am looking for video clips ...)***
SH14-8	*usain bolt*	*... about athletics, for example Usain Bolt.*
SH14-11	*history of the bbc*	*... about the history of the British Broadcast Company (BBC).*
SH14-18	*polar bears*	*... with polar bears in the wild.*
SH14-29	*buckingham palace crowds*	*... that show crowds at Buckingham Palace.*
SH14-35	*world cup goals*	*... that show goals from the FIFA World Cup.*

similar profile than those recruited for the SH13 queries, and were asked to generate a set of ad-hoc topics. For this, participants were first instructed to think of information needs for content they would find interesting and to state them in natural language. Next, they were requested to formulate short keyword-based textual queries for their information needs similar to those they would use in *YouTube*. Finally, they had to enter their queries into the AXES system and find segments of video relevant to their information needs. To ensure the generation of ad-hoc queries, users had to identify at least two relevant video segments per information need within the BBC2 collection. If a query did not trigger enough relevant results, users were instructed to re-formulate and commence a new search. The organisers selected 36 topics to form the test set used at the SH14 task. Examples of these topics are shown in Table 4.8. In contrast to the SH13 queries, the topics from the SH14 are more ambiguous, contain fewer terms, and target broader subjects.

Table 4.9: Length statistics of SH13, SH14, and SAVA queries.

Topics	**Type**	**Number**	**Ave. len.**	**S.D. len.**	**Max len.**	**Min len.**
SH13	Known-item	50	6.2	3.3	19	1
SH14	Ad-hoc	36	2.5	0.7	4	2
SAVA	Ad-hoc	30	5.0	2.5	11	1

The SAVA topics for the BBC2 collection

The SAVA topic set contains 30 ad-hoc topics in total (Eskevich et al., 2015). Two different user-groups took part in the topic-generation study. The first group consisted of media professionals, including journalists, archivists, and researchers, who claimed to frequently use retrieval systems to find reusable video material. The second group was representative of a general-audience, similar to the team recruited in the generation of the SH topics. Participants were asked to explore the BBC2 collection by using the AXES system and to generate search topics, in a similar setup as that used for the collection of the SH14 topics. Some examples of SAVA topics are shown in Table 4.10.

Query processing and query length statistics

In all IR experiments reported in this thesis with the BBC collections, English queries were preprocessed in the same way as the English document transcripts, that is, by using Terrier's text processing pipeline as described in Section 4.1.2, including tokenisation, lowercasing, stopword removal, and stemming. Table 4.9 shows query term statistics for the SH13, SH14, and SAVA topic sets after queries are processed. In terms of length, queries from the SH13 and SAVA sets contain a larger proportion of content bearing terms than queries from the SH14 set.

4.1.4 Relevance assessments

Relevance judgements for the SH13 topics were determined based on the segments of video content that users had selected for generating their known-item topics (Eskevich et al., 2013a). In particular, each segment selected to produce a topic was automatically judged to be relevant to that topic. In addition, query creators were also asked to re-adjust the boundaries of their segments in order to remove from them any content not considered relevant to their information needs. Figure 4.6a shows the number of relevant segments segregated by length. As can be seen in the figure, about 60% of all segments marked as relevant to some SH13 topic are 2 minutes long or less, while the longest segments are approximately 10 minutes long.

Segment-level relevance assessments for the SH14 and SAVA topics were obtained by the benchmark organisers through a pooling procedure (Eskevich et al., 2014, 2015). The pools were formed with the top 10 video segments from each ranked list of results submitted by different participating teams in the search task of the benchmarks. For these tasks, it

Table 4.10: Example of SAVA ad-hoc topics for the BBC2 collection.

Topic	Query	Visual cues	Description
SAVA-13	*jaipur terrorist*	*explosion*	*I am looking for information concerning the Jaipur bombing. Relevant clips contain information about Jaipur, the casualties and the terrorists.*
SAVA-15	*murder* *police* *crime scene* *statistics* *london*	*police,* *crime scene*	*I am looking for clips about murders in London. Relevant clips contain news items about murders, crime scenes and statistics about murder rates in different parts of London.*
SAVA-20	*hill walking* *public footpaths*	*public footpaths,* *hill, walking*	*I am looking for the countryside. Relevant clips contain hill walking and landscapes.*
SAVA-24	*squirrels*	*squirrels*	*I am looking for films that contain squirrels.*
SAVA-41	*burmese* *rangoon* *cyclone*	*cyclone*	*I'm looking for information about the cyclone that hit Burmah. Related links should provide visual and/or spoken information about the disaster.*

was common among participating systems to return overlapping segments in the search results for a query, these are, video clips whose time-spans overlay with those from another clip within the same video. In order to reduce the amount of annotation effort required in assessing these segments, the task organisers pre-processed each ranked list returned by the participants before generating the pools. This procedure consisted of removing the overlapping material from a segment in a ranked list if was found to overlap with some other segment located at higher-positions in the same ranked list. Further, in order to comply with BBC regulations on the use of its material in crowdsourcing experiments, the ending times of the segments were adjusted to restrict their maximum length to 120 seconds.

Binary relevance judgements were then produced for each segment-topic pair by assessors recruited from Amazon Mechanical Turk[7] (Larson et al., 2012). Assessors were presented with a simple user interface showing the description of the topic and with a playback tool to reproduce the video segment to be judged. Figure 4.5 shows a screen capture of this web interface. Also, in order to comply with legal requirements from the BBC, the playback tool had restricted access to the contents of a video. In this respect, each crowd-worker was only permitted to view the contents of the segment to be judged. Besides providing a relevance judgement, assessors had to describe the reasons underlying why each judgement was made, as well as to provide a list of meaningful keywords spoken in the clip. Circa 10,000 segment-topic pairs were assessed against the SH14 topics in the

[7] `http://www.mturk.com`

crowdsourcing study, while 2,300 were assessed against the SAVA topics. In both cases, about 35% of all segments were judged as relevant.

Because participating systems produced different results for a given query, it was still common for a pool of results to contain overlapping segments. Moreover, since these overlaps were not completely removed from the pools of results by the organisers, a substantial amount of video content in the SH14 and SAVA studies was judged by more than one assessor. In particular, about 48% and 12% of all segments judged in the SH14 and SAVA studies respectively were judged by two or more assessors. In this respect, it is useful to measure the inter-annotator agreement among crowd-workers in order to determine the reliability of the assessments produced as part of the SH14 and SAVA studies. A way to achieve this is to calculate the Krippendorff's α coefficient (Krippendorff, 2011)[8].

Based on a Krippendorff's α of 0.41 for the SH14 and 0.43 for the SAVA assessments respectively, the agreement among multiple assessors at judging the relevance of a segment can be considered in the range of "fair" to "moderate" if compared to other values of α reported in the IR literature (Schaer, 2012; Schaer et al., 2016; Verma et al., 2016). This means that while assessors agreed moderately when judging a video segment to be relevant to a query, it was also frequent for them to disagree and to provide inconsistent judgements. Nevertheless, the grade of agreement between assessors in the SH14 and SAVA studies is comparable with that from other relevant assessment studies and can be therefore considered reliable for experimentation.

To remove segment-overlap and solve possible label inconsistencies found in the ground truth, the benchmark organisers decided to take the union of overlapping segments judged in the relevance assessment study. A union was considered to be relevant if contained a single passage judged as relevant by an assessor. Figure 4.4 shows an example of this process, where a common segment was judged by three different assessors as both relevant (green) and non-relevant (red). In the example, a new relevant segment is formed from the union of two relevant segments: one produced by Judge 1 and another one by Judge 3. The irrelevant segment identified by Judge 2 is effectively ignored in the calculation of the union.

Figures 4.6b and 4.6c show the length distribution of relevant segments resulting from the union process described above. Compared to the length distribution of the SH13 ground truth shown in Figure 4.6b, those of SH14 and SAVA are skewed towards segments that are less than 150 seconds long. In the latter case, the extent and boundaries of the relevant segments were largely determined by the retrieval systems that contributed to the pools of results and the length restrictions imposed in the crowdsourcing experiments. In addition, assessors were not asked to correct the boundaries of the segments judged relevant if these were found to contain some leading or trailing irrelevant material. Consequently, while the boundaries of the segments from the SH13 ground truth can be considered more

[8] Note that using other measures of inter-annotator agreement, such as Cohen's or Fleiss' Kappa, would not be appropriate in this case since not all assessors were asked to judge all of the segments from the pools.

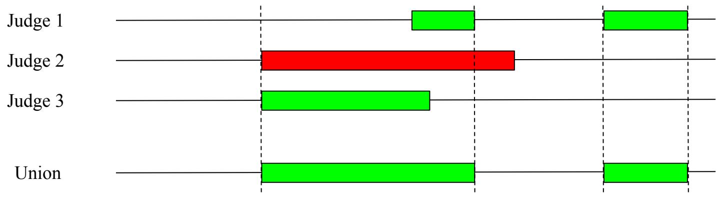
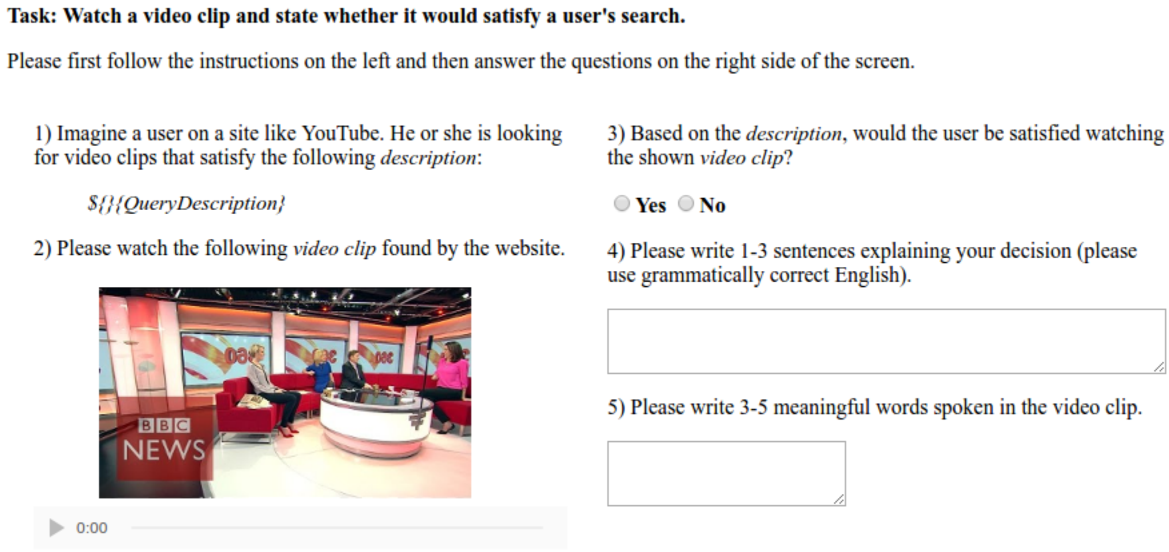


Figure 4.4: Example of video segments judged as relevant (green) and non-relevant (red) by three assessors and their unions into single relevant segments.



Task: Watch a video clip and state whether it would satisfy a user's search.

Please first follow the instructions on the left and then answer the questions on the right side of the screen.

1) Imagine a user on a site like YouTube. He or she is looking for video clips that satisfy the following *description*:

${}{QueryDescription}

2) Please watch the following *video clip* found by the website.

3) Based on the *description*, would the user be satisfied watching the shown *video clip*?

○ **Yes** ○ **No**

4) Please write 1-3 sentences explaining your decision (please use grammatically correct English).

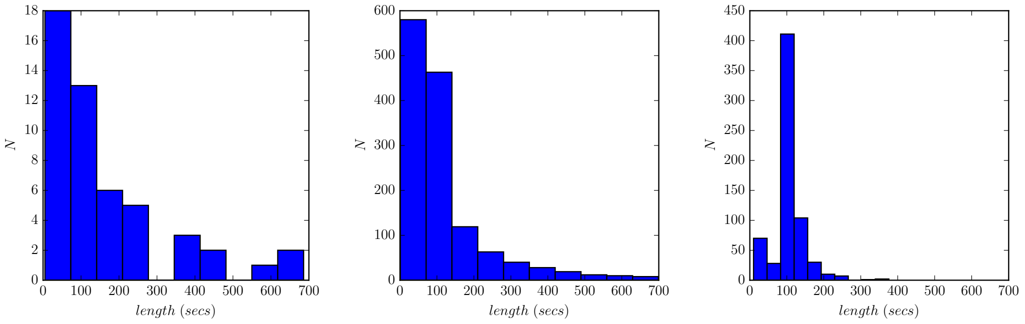
5) Please write 3-5 meaningful words spoken in the video clip.

Figure 4.5: Web interface for collecting relevance assessments in SH14 and SAVA.

reliable, those from the SH14 and SAVA ground truths should be considered sub-optimal in the sense that they may not reflect the best possible boundaries an annotator would have selected for these segments.

4.2 The SDPWS collections of academic presentations

The SDPWS collection contains oral presentations recorded at different editions of the Spoken Document Processing Workshop (SDPWS), an annual scientific meeting organised by the speech processing community in Japan. This collection has been used in several SCR benchmarks, namely the NTCIR-10 SpokenDoc-2 (SD2), the NTCIR-11 SpokenQuery&Doc (SQD1), and the NTCIR-12 SpokenQuery&Doc-2 (SQD2) tasks (Akiba et al., 2013a, 2014, 2016). The remainder of this section describes the details of the speech recordings, transcripts, topics, and relevance assessments that were distributed to the participants of these tasks and used in the experiments described in this thesis.



(a) SH13 topics.

(b) SH14 topics.

(c) SAVA topics.

Figure 4.6: Length distribution of relevant segments in the BBC collections.

4.2.1 Overview

The SDPWS collection contains 114 academic presentations, comprising about 30 hours of speech material. The NTCIR task organisers provided the audio file of each presentation, plus the time boundaries of utterances produced by a voice-activity detection (VAD) tool, and hand-labelled timestamps of slide transitions. In addition, manual and automatic transcripts of the presentations are provided, plus the acoustic and language models that were used to perform speech recognition.

The presentation recordings from the SDPWS collection contain semi-scripted spoken material produced by a single speaker in front of a live audience. In contrast to TV speech, speech produced in these academic talks presents less acoustic variability, as the majority of the presentations were recorded in relatively similar acoustic conditions. Also, since speakers often divert from their planned presentations, their speech can be regarded as more spontaneous than that from broadcast news and other scripted material. This is evidenced by the high frequency with which hesitations, false starts and other spontaneous speech phenomena occur in the talks.

In terms of domain, most talks in the SDPWS collection are about speech processing, information retrieval, machine learning and related topics. Thus, the recordings contain occurrences of highly technical terms with a widespread use of acronyms, scholarly terms, and foreign words in English, as commonly seen in computer science talks.

Two versions of the SDPWS collection were distributed to the participants of the NTCIR tasks. The data distributed in the SD2 task (SDPWS1) (Akiba et al., 2013a) contains 106 presentations corresponding to talks recorded at the first six editions of the SDPWS. In the subsequent SQD1 (Akiba et al., 2014) and SQD2 (Akiba et al., 2014) tasks, this collection was extended with 10 additional talks recorded at the seventh edition of the workshop along with slide change annotations for a subset of 98 presentations. This smaller subset of 98 presentations (SDPWS2) was used at the following SQD1 and SQD2 tasks as the target collection for the search task. Table 4.11 summarises this distinction and provides duration statistics for each collection. Since pauses between utterances were

Table 4.11: Duration statistics of presentation recordings from the SDPWS collection.

		Duration				
Dataset	**Documents**	**Total**	**Avg.**	**S.D.**	**Min**	**Max**
SDPWS1	104	28 hrs 38 min	16 min	2 min	11 min	21 min
SDPWS2	98	26 hrs 46 min				

Table 4.12: Manual and ASR transcripts provided by the NTCIR task organisers for the SDPWS collections.

SpokenQuery&Doc ID	Short ID	SDPWS1	SDPWS2
MANUAL	MAN	✓	✓
K-REF-WORD-MATCH	K-MATCH		✓
REF-WORD-MATCH	MATCH	✓	✓
REF-WORD-UNMATCHLM	UNMATCH-LM	✓	✓
REF-WORD-UNMATCHAMLM	UNMATCH-AMLM		✓

removed from the audio files before calculating the duration estimates, the figures from Table 4.11 do not reflect the actual duration of the SDPWS recordings, which could be up to 20-30% longer if periods of silence were to be included.

4.2.2 Speech collection and transcripts

As part of the data preparation, the organisers of the NTCIR tasks segmented the audio recordings into small speech fragments at pauses longer than 200ms, based on the output of a voice-activity detector (VAD). This process resulted in a list of sequential spoken fragments for each presentation termed inter-pausal units (IPUs) which can be considered as approximations of utterances. The speech collections were then distributed as sequences of IPUs associated with a particular presentation ID. IPUs were released in WAV format with a sampling rate of 16kHz and 16-bit of precision recorded in a single audio channel.

After audio fragmentation, the data creators produced orthographic and automatic transcripts for each IPU. Details of the alternative transcripts available are summarised in Table 4.12. The IDs shown in the first column of the table correspond to those reported in (Akiba et al., 2016). The third and fourth columns indicate the transcript types available for the SDPWS1 (Akiba et al., 2013a) and SDPWS2 (Akiba et al., 2014) collections respectively. Since the full range of ASR transcripts is only available for SDPWS2, the experiments conducted in this thesis are carried out with this version of the collection.

The rest of this section describes the transcripts, ASR systems, and models that the NTCIR organisers used to automatically transcribe the presentation speech from the SDPWS2 collection.

Manual transcripts

The orthographic transcripts of the SDPWS2 talks were produced by hired transcribers who further annotated the transcripts with markers as shown in Table 4.13. These marked

Table 4.13: Annotations of spontaneous speech phenomena in manual transcripts of the SDPWS collection.

Marker	Description
<H>	Non-lexical lengthening of vowel.
<Q>	Non-lexical lengthening of consonant.
<FV>	Vowel with unrecognisable phonemic status.
<息>	Breathing noise.
(笑) / <笑>	Laughter with/without speech.
(泣) / <泣>	Cry with/without speech.
(咳) / <咳>	Cough with/without speech.
(あくび)	Yawn with speech.
<雑音>	Noisy speech.
(L)	Whispery speech.
(D), (D2)	Word-fragment, unfluent speech.
(W)	Reduced, truncated, or incorrect pronunciation.
(?)	Uncertainty in the transcription.
(F)	Filled-pause.
(M)	Meta-linguistic expression.
(O)	Archaic Japanese.
(A)	Use of Latin scripts in transcription.
(K)	Use of Katakana scripts in transcription.
(s)	Slide transition.

the presence of various spontaneous speech phenomena, such as noise, whispery speech, hesitations, and filled-pauses. Transcribers created orthographic transcripts with a mix of *Kanji* (Chinese logographs), *Kana* (Japanese syllabary), and *Romaji* (latin scripts) by following a strict set of guidelines designed to reduce the common phenomenon of variation among transcribers of written Japanese.

Besides possessing a high degree of freedom in word formation, no characters are placed between words to delimit word boundaries in written Japanese. This is the case for the manual transcripts of the SDPWS collection, in which IPUs were transcribed in *Yokogaki* style with words contiguously placed from left to right, one after the other. In order to generate a term-index for a collection of Japanese documents, it is common practice to first split the text into morphemes by using a morphological analyser. The particular tokenisation process applied to the SDPWS2 transcripts is explained in later sections of this chapter.

The slide transition annotations (s) were included in the manual transcripts of the SDPWS2 collection to indicate the times when slide transitions were made by the presenters. In addition to this, transcribers were required to identify groups of consecutive slides that were used to present a single topic or idea in a presentation. These slide groups thus define a list of topical homogeneous segments within a presentation and were used in the SQD1 and SQD2 tasks as pre-defined passages to be retrieved in response to a query. The set of slide group passages provides a ground-truth segmentation of the SDPWS2 collection, and is the main focus of the experiments presented in Chapters 6 and 7 of this thesis.

ASR models and transcripts

The organisers of the NTCIR benchmarks produced various automatic transcripts of the speech material from the SDPWS collection. These were generated with the Julius[9] (Lee and Kawahara, 2009) and Kaldi[10] (Povey et al., 2011) ASR toolkits under different training conditions of language and acoustic models.

The Julius ASR system

The front-end of the Julius-based recogniser was configured to extract 38 cepstral features for every 10ms of speech data (Akiba et al., 2014). This feature vector included 12 MFCCs plus their first and second derivatives as well as the first and second derivatives of the signal energy. The back-end of this system followed a standard GMM-HMM framework, with tri-phone state context-dependent HMMs and 32 gaussian mixtures used for modeling phone-transition probabilities.

Spoken utterances were transcribed by the Julius system in two passes. In the first pass, a 1-best transcription was obtained with a left-to-right (forward) bi-gram language model and a frame-synchronous beam search procedure, while in the second pass a right-to-left (backward) tri-gram language model was used along with a stack-decoding search algorithm. In the second pass, the output from the first pass was used to construct a "word trellis index" which enables the search space of possible recognition hypotheses to be narrowed and the computation time of the second decoding pass to be reduced (Lee et al., 1998).

The Julius ASR system was configured to output the 10-best recognition hypotheses, plus word lattices and confusion networks for each processed IPU. In addition, the system was configured to produce confidence scores for each word in the transcription hypothesis. These confidence estimates are calculated with the algorithm proposed by Lee et al. (2004), which approximates word posterior probabilities based on the likelihoods of partial sentence hypothesis that are generated by the stack-decoding search algorithm.

The Kaldi ASR system

The Kaldi-based recogniser is based on a recipe distributed with the Kaldi toolkit[11] originally created for building ASR models with data from the Corpus of Spontaneous Japanese (CSJ) (Maekawa et al., 2000). The front-end of this system extracts 12 MFCCs, their deltas and delta-deltas, and performs cepstral-mean and variance normalisation (CMVN) per IPU. Subsequently, a number of reduce and transform operations are applied to these feature vectors including: linear discriminant analysis (LDA) (Haeb-Umbach and Ney, 1992), maximum-likelihood linear transform (MLLT), and feature-space MLLT (fMLLT) (Gales, 1998). The resulting 40-dimensional feature vectors per frame are used to

[9] http://julius.osdn.jp

[10] http://kaldi-asr.org

[11] https://github.com/kaldi-asr/kaldi/tree/master/egs/csj

pre-train a deep belief network (DBN) with the contrastive divergence algorithm (Hinton et al., 2006). The weights estimated from this unsupervised training procedure are then used to initialise the weights of a DNN which is later trained to classify speech frames into HMM states with stochastic-gradient descent (SGD) and the cross-entropy objective function (Hinton et al., 2012). Finally, to better model the dependencies that exist between acoustic frames, this DNN is fine-tuned using sequence-discriminative training (Veselỳ et al., 2013) using the state-level minimum Bayes risk (sMBR) criterion. The benchmark organisers used this DNN-HMM based acoustic model along with a tri-gram language model to transcribe the speech content from the SDPWS collection. The same models obtain WERs as low as 9% when transcribing a held-out set of academic presentations from the CSJ. This system also produced word-level confidence scores based on word posterior probabilities calculated from the recognition lattice.

ASR transcripts

Two groups of ASR transcripts were generated by the benchmark organisers using the systems described above: one containing recognised word units; and another containing recognised sub-word (syllables) units. In all of the experiments described in this thesis with the SDPWS collection, only the word-level transcripts were used, hence in the following only these are described.

Table 4.14 summarises the main characteristics of the acoustic and language models that the benchmark organisers used to transcribe the SDPWS collection. They obtained the K-MATCH and MATCH transcripts with acoustic and language models which they previously trained with transcribed speech from the Corpus of Spontaneous Japanese (CSJ) (Maekawa et al., 2000; Maekawa, 2003; Maekawa et al., 2004).

The CSJ is one of the largest collections of academic presentation speech available in the Japanese language. It contains 606 hours of academic talks, derived from two types of presentation speeches: academic presentations (APS) and simulated public presentations (SPS). Much of the material covered by the APS is about topics in computational linguistics, phonetics, phonology, and speech processing, and thus provide an adequate set of training data for recognising the highly technical speech content from the SDPWS collection.

The benchmark organisers obtained the UNMATCH-LM transcripts by using the same acoustic model they used to generate the MATCH transcripts, but with a language model estimated from 75 months of newspaper articles from the Continuous Speech Recognition Consortium (CSRC) corpus (Lee et al., 2002). Lastly, organisers produced a third set of transcripts, UNMATCH-AMLM, with the acoustic and language models that are distributed with the *Julius dictation kit v4.3.1*. The latter models were originally trained with text from the Balanced Corpus of Contemporary Written Japanese (BCCWJ) (Maekawa et al., 2014) and speech data from the Japanese Newspaper Article Sentences (JNAS) (Itou et al., 1999) corpus.

Table 4.14: Details of the ASR models used to automatically transcribe the SDPWS collection.

(a) Language models

ID	**System**	**Dataset**	**Vocab.**	**In-genre?**
K-MATCH	Kaldi	CSJ	29,186	Yes
MATCH	Julius	CSJ	29,186	Yes
UNMATCH-LM	Julius	CSRC	21,322	No
UNMATCH-AMLM	Julius	BCCWJ	64,274	No

(b) Acoustic models

ID	**System**	**Dataset**	**Size**	**In-genre?**
K-MATCH	Kaldi	CSJ (APS)	240 hrs	Yes
MATCH	Julius	CSJ (APS + SPS)	606 hrs	Yes
UNMATCH-LM	Julius	CSJ (APS + SPS)	606 hrs	Yes
UNMATCH-AMLM	Julius	JNAS	86 hrs	No

The K-MATCH and MATCH transcripts differ in the underlying ASR technology and the amount of speech data used to train the acoustic models. While the K-MATCH transcripts were produced with the Kaldi toolkit by using a DNN-HMM framework, those from the MATCH group were generated with the Julius toolkit by using a more conventional GMM-HMM framework. Also, the acoustic models generated with Kaldi were trained with a subset of approximately 240 hours of academic presentation speech from the CSJ, whereas those generated with Julius were trained with all recordings from the CSJ (606 hours). The MATCH, UNMATCH-LM, and UNMATCH-AMLM transcripts were all produced with Julius, but they differ due to differences in the speech and text data used to train the recognition models, with the same framework being applied in this case for both training and decoding of the spoken material.

Processing and indexing of Japanese transcripts

This section describes how we processed and indexed the orthographic and automatic Japanese transcripts from the SDPWS and CSJ collections for the experimental work presented in this thesis.

In order to obtain tokens that can be used as indexing features of documents, we processed the orthographic transcripts of the SDPWS collection with the morphological analyser MeCab[12] v0.996. For this purpose, the Ipadic dictionary v2.7.0 (Asahara and Matsumoto, 2003) was used along with the MeCab analyser. Besides text tokenisation, MeCab also provides the base (root) form, pronunciation form, and POS tag of each identified word. When constructing text indices from SDPWS transcripts, we used the base form of words produced by MeCab as indexing terms.

As an additional pre-processing step for manual transcripts, we handled annotation labels from table 4.13 as follows. Annotations that do not relate to any content-bearing

[12] http://mecab.sourceforge.net/

words, these are those with label codes H, Q, FV, 息, L, F, M, O, s, 雑音, were discarded from the manual transcripts. Words marked with special pronunciations, specifically those marked with 笑, 泣, 咳, あくび, D, D2, ?, K were retained, while their label codes were removed from the text prior to performing morphological parsing, since the presence of annotations labels affects MeCab's output.

Annotations with label codes A and W mark usage of alphabetic characters (borrowed words) and incorrect pronunciations respectively. A annotations, e.g. (A ティーエフアイディーエフ;tf-idf), specify two forms for a word, its alphabetic form and its pronunciation form in Katakana characters, while W annotations, e.g. (W エーキュー;要求), specify a mispronounced word along with its correct pronunciation. In the construction of the IR indices from manual transcripts, we kept the alphabetic forms of words and correct pronunciations as indexing terms while processing the text with MeCab.

While knowing the identity of the words spoken is necessary for creation of a term index, knowing the exact times when words are spoken is required for performing acoustic/prosodic analysis. The orthographic transcripts from the SDPWS collection do not include the start and end times of individual words as these were not originally added by the NTCIR transcribers. Nonetheless, such timestamps can be recovered through forced-alignment by constraining an ASR system, with a given pre-trained acoustic model, to recognise a given sequence of words. In order to get the starting time and duration for each word in the manual transcripts, we first translate its pronunciation form in Katakana characters into its phonemic representation by means of a pronunciation dictionary provided by the NTCIR task organisers. For instance, we used the sequence "m a z u i" as the phonemic representation for the Katakana sequence "マズイ" corresponding to the Kanji "まずい". Furthermore, the alternative pronunciation forms as found in the A or W labels were used for words annotated with these type of special markers. We then performed forced-alignment by using the *Julius Segmentation Kit*[13] v1.0. In this step, word timestamps were obtained for an IPU by feeding this tool with the phonemic translations and the WAV file of the IPU, plus the acoustic model MATCH (see Table 4.14) provided by the task organisers.

In all the experiments conducted with automatic transcripts from the SDPWS collection, we only processed and indexed the text from the 1-best recognition hypothesis of each IPU. The language models used by the NTCIR task organisers to obtain these transcripts (Table 4.14a) had originally been estimated from text tokenised with the ChaSen[14] analyser v2.4.4 and the UniDic dictionary v1.3.9. Further, instead of defining words with their surface text form only, the words in these language models had been defined as the concatenation of their surface and base forms, and POS tags, as produced by ChaSen. Thus, since the ASR systems were constrained to produce text from these language models, the 1-best hypothesis already contained suitable tokens that could be used as indexing

[13] `http://sourceforge.jp/projects/julius/downloads/32570/julius4-segmentation-kit-v1.0.tar.gz`

[14] `http://chasen-legacy.sourceforge.jp`

terms for SCR purposes.

Although the tokens produced by ChaSen could be used as indexing terms for ASR transcripts, other researchers have shown these to be less effective than tokens recognised by MeCab Nanjo et al. (2014). For this reason, we re-tokenised the text from the 1-best ASR hypothesis with MeCab in a two-step process: first, we generated a new string generated by concatenating the surface form of every token present in a 1-best hypothesis, without inserting spaces between the extracted terms; second, we processed this string processed with MeCab to obtain a new possibly different tokenisation result than ChaSen's. Whenever MeCab produced a different tokenisation string than ChaSen's, the timestamp of each word included in the original output of the 1-best hypothesis needed to be re-estimated. To do this, we performed forced-alignment against the new tokenisation string produced by MeCab, by following the same process described before for the manual transcripts.

In addition to using MeCab's tokenisation, previous research has demonstrated that lemmas of nouns and verbs are more effective indexing features in Japanese SCR than character or phone n-grams (Shigeyasu et al., 2009). Therefore, we removed all tokens not tagged as verbs or nouns by MeCab from the manual and ASR transcripts before constructing retrieval indices. Additionally, we removed words contained in a stop word list with 44 frequent prepositions and determiners, in order to discard some function words from the indices that MeCab repeatedly misclassified as nouns or verbs. By filtering text this way, the length of each presentation transcript was reduced to about 50% of its original length.

As the last processing step prior to indexing, we converted simple-width characters into their full-width Unicode equivalent. This step was necessary as it was common for transcribers of the SDPWS presentations to utilise 8-bits (simple-width) and 16-bits (full-width) variants of Latin characters interchangeably when transcribing *Romaji* words in Japanese. Mapping characters to a consistent character set avoid the problem of missing trivial matching instances between terms in the query and the documents.

After processing the text, we used Terrier to generate an index for each transcription type. In this case, Terrier was configured the same as for the indexing of English transcripts, as described in Section 4.1.2, with the difference that stemming was disabled and the English stop word list was replaced by the list of 44 common Japanese words introduced previously.

Table 4.15 presents term statistics obtained from the inverted indices of all available transcripts types for the SDPWS collection, while Table 4.16 reports word recognition rates and index similarity metrics computed against the reference index. Among all ASR transcripts, the K-MATCH and MATCH transcripts present the best recognition quality overall, with K-MATCH transcripts being substantially more accurate despite being produced with acoustic models trained with less speech data. The UNMATCH-LM and UNMATCH-AMLM transcripts present significantly higher error rates in comparison due

Table 4.15: Length statistics in number of terms of processed transcripts from the SDPWS2 collection.

Transcript	Avg. len.	S.D. len.	Max. len.	Min. len.
MAN	1,769.17	276.15	2,424	895
K-MATCH	1,763.09	274.01	2,385	923
MATCH	1,752.15	262.81	2,360	983
UNMATCH-LM	1,922.76	285.21	2,537	1,094
UNMATCH-AMLM	1,594.50	247.70	2,176	842

Table 4.16: Recognition accuracy of presentation transcripts as measured by various index similarity metrics for the SDPWS2 collection.

Transcript	#Terms	WER	UTER	TER	BIA	RIA
MAN	6,230	0%	0	0	1.00	1.00
K-MATCH	6,350	22.0%	0.27	0.48	0.49	0.56
MATCH	6,131	43.7%	0.41	0.82	0.28	0.37
UNMATCH-LM	11,219	67.5%	0.51	1.57	0.10	0.19
UNMATCH-AMLM	14,190	70.5%	0.54	1.46	0.10	0.20

to the higher mismatch in domains that exist between the language models used to generate these transcripts and the academic talks of the SDPWS2 collection. A distinguishable characteristic of UNMATCH-LM transcripts is that they contain almost twice the number of insertion errors compared to the other types of transcripts, this issue is somewhat reflected on the transcript length statistics and the index similarity metrics.

4.2.3 Topics

During the four cycles of the NTCIR SCR benchmarks, different topic sets were collected and released to task participants for each cycle. The first two editions of the benchmark, SD1 and SD2 focused on search queries that were stated as written text, while the subsequent SQD1 and SQD2 did so on queries stated with the spoken word. While the former sets represent a conventional search scenario where users type in their search requests on a keyboard, the latter introduced a novel scheme in which users communicate their queries by using a voiced-enabled (spoken) interface. This change is in line with the increased general research interest in conversational interfaces received in recent years. The remainder of this section describes these topic sets in greater detail.

The SD2 topics for the SDPWS1 collection

The SD2 topic set (Akiba et al., 2013a) contains 120 ad-hoc written queries targeting the SDPWS1 collection. These topics were formulated by six volunteers who took part in a query generation study organised by NTCIR researchers. Each volunteer was asked to create 20 queries based on the content of the articles from the proceedings of the 1st-6th editions of the SDPWS workshop and the orthographic transcripts of their corresponding audio presentations. In particular, query creators were asked to generate some topics

Table 4.17: Examples of SD2 topics for the SDWPS1 collection.

Topic	Query	English translation
SD2-003	サフィックスアレイとはどんなものか。	*What is a suffix array?*
SD2-008	決定木を利用している研究について知りたい。	*I would like to know about research using decision trees.*
SD2-014	音声の韻律情報とはどんなものか。	*What is prosodic information of speech?*
SD2-074	集合知とはどのようなものでどう利用されているのか知りたい	*I would like to know about collective intelligence and how it is used*
SD2-092	重要な文を自動的に求める方法が知りたい	*I want to know how to automatically obtain important sentences*

based on the content of the articles and some others based solely on the content of the transcripts. In total, 80 topics were produced from the workshop proceedings and 40 from the presentation transcripts. Additionally, participants were encouraged to produce topics whose relevant information may be encountered within passages of varying length in a presentation, thus encouraging topic creators to think about information needs that target different levels of content granularity.

Table 4.17 shows five sample topics from the SD2 set. The table shows the original query text in Japanese and its corresponding English translation obtained with *Google translate*. The great majority of queries from this set are purely informational with most queries stated as Wh-questions, as if they were to be input into a question-answering system.

The SQD1 and SQD2 topics for the SDPWS2 collection

The SQD1 and SQD2 topic sets contain 37 and 80 ad-hoc spoken-queries respectively, these were recorded by NTCIR researchers in two independent user studies following a similar methodology. In these studies, each volunteer was asked to select an article from the 1-7th editions of the SDPWS proceedings and to formulate a query based on the content of one of the article's paragraphs. The recording session occurred at a later stage, in which volunteers were not shown the content of the article they had selected to avoid them from uttering verbatim sentences found in the article. Volunteers were given unlimited time to formulate their queries and were not interrupted while in the recording sessions. As result of this, participants tended to produce extremely verbose queries. At the end of a recording session, participants were asked to listen and transcribe the spoken queries they had produced. These manual transcripts were made available by the dataset creators as well as ASR transcripts of the spoken queries that were later generated with the same set of ASR systems described in Section 4.2.2.

Tables 4.18 and 4.19 show examples of spoken queries from the SQD1 and SQD2 sets respectively. The queries from these sets tend to be extremely long, resembling full paragraphs rather than the more typical keyword-based queries. The large majority of

Table 4.18: Examples of SQD1 topics for the SDWPS2 collection.

Topic	Query	English translation
SQD1-05	(F えーと)(F ま)最近その(D に)音声認識が¡息¿結構(F ま)いろんなところで使われるようになってきて¡H¿(F ま)だいぶ精度もいいような気がしているんですけどやっぱりまだ上手く認識されないっていうことが結構あって¡H¿例えばなんか漢字¡H¿が間違っていたりとか¡H¿(F えー)(F ま)それはよくあるんですけど後は(F ま)全然違う単語に(F ま)認識されてしまったりとかして¡H¿(F ま)なんでこんなふうになってしまうのかっていうのがちょっとよくわからないんですけど(F ま)そのように何か誤認識が起きてしまうような(F ま)原因は何であるかというのが知りたいです	*Recently there is quite a lot of speech recognition comes to be used in many places It seems that accuracy seems to be good, but I feel that it is still quite good to not recognise, for example, something like Kanji There are times when it is wrong or something it is common but afterwards it is not recognised at all whether it will be recognised as a different word why it will become like this but why It seems that something misleading seems to happen So I want to know what the cause is*
SQD1-08	論文の中で(F えっと)アライメントについて説明しているところがあると思うんですけど(F えっと)論文の中だと(F えっと)統計的機械翻訳説明のところで(F と)そのアライメントのアライメントについて説明がされているんですけど(F んー)そこの説明のところの¡息¿スライド探して欲しいです(F えっと)その論文の中の例だと(F えっとー)(D と)確か私は本を借りますっていう例文を使って多分説明していたと思うんですけどそこのところの¡息¿(F えっとー)スライドでの説明が聞きたいです	*I think there is a place to explain the alignment in the paper, but in the paper I will explain the alignment of the alignment in statistical machine translation explanation, but in the explanation there I'd like you to find a slide of the paper I certainly believe that I was explaining it probably by using example sentences like borrowing a book as an example in that paper but I would like to hear the explanation on the slide there is*
SQD1-13	ドキュメント中の(D き)(F えー)(D きょ)(F えー)強調発話の検出に(D つか)(F えー)¡息¿(D と)ドキュメント中の強調発話の検出は(F え)どのようなアルゴリズムを使って実装しましたか	*On detection of emphasised speech in a document, what kind of algorithms are used to implement detection of emphasised utterance in the document*

these spoken queries provide a in-depth description of the information needs, and may even include participants' interests and motivations.

Query processing and query length statistics

We processed the written queries and transcripts of the spoken queries in the same way as the presentation transcripts of the SDPWS collection, as described in Section 4.2.2. Recall that this includes the removal of tokens that are not nouns or verbs from each query.

Table 4.20 summarises term statistics of the processed queries. The figures in the table reflect a striking difference between the length of queries stated in written and spoken form. On average, the transcribed spoken queries contain 4 times more terms than the written queries.

Table 4.19: Examples of SQD2 topics for the SDWPS2 collection.

Topic	Query	English translation
SQD2-40	(F えー)学会講演音声¡H¿をリアルタイムで(F えー)字幕を表示するという研究について(F えー)知りたいことがあります(F えー)その先行研究と致しまして(F えー)文を入力の一単位としまして(F えー)その字幕に(F えー)改行を挿入するという研究が(F えー)まず始めにありました(F えー)その研究では(F えー)文を(F えー)入力の一単位として区切っているので(F えーっと)リアルタイムでの表示には(F え)遅延が発生(D してましめ)してしまうという問題がありました(F え)それに対して本研究では¡H¿(F えー)音節を入力の一単位としまして(F えー)¡H¿(F えー)改行の挿入を行うことで(F え)遅延時間を(F えー)短くするように工夫を行った研究がなされています(F えー)それに際しまして(F えー)先行研究文単位での挿入を行う研究と今回の(F えー)文節単位で挿入を行う研究についての結果に関しまして(F えー)確かに文節単位で(F えー)改行の挿入を判断する改行の挿入を行う研究では(F えー)遅延時間が短くなりました(F え)しかしながら(F えー)文節単位で改行の挿入を行うとその改行の挿入が正しいか正しくないかの再現度と精度が(F えー)低下してしまうという問題も(F え)発生しているそうですその(F えー)文単位での(F えー)改行挿入と文節単位での改行挿入を比較しとき(F えー)精度と(F え)再現度はどの程度低下したか(F えー)教えてください	*Society Lecture There is something you want to know about the research that displays subtitles in real time. As a preceding study I will start with a research that inserts a line break in that subtitle as a unit of input In the research, since sentences are delimited as one unit of input, there is a problem that delay occurs in display in real time. In contrast to this, in this research, a syllable is divided into units We are doing research to make the delay time shorter by inserting line breaks In that case we are conducting research inserting in preceding research sentence and research inserting in this phrase unit As regards the result of certainly judging the insertion of a new line in units of clauses The delay time is short in the study inserting line breaks However, if you insert a line break in units of clauses, the insertion of newlines is correct or not There seems to be a problem that reproducibility and precision deteriorate is also occurring It tells to what extent the precision and the reproducibility degrade when comparing line break insertion in sentence unit and line feed insertion in unit of clause*
SQD2-70	(F っと)大学の講義などでは(Fえっと)専門用語の出現が多くまたその専門用語はいくつかの単語を合わせた複合語であることが多い¡H¿と思います¡息¿そこで講義の音声認識の形態素解析におけるキーワード抽出(D2を)¡息¿について¡息¿実験を行った論文¡H¿があったと思うんですがそのちゅー¡息¿そのキーワード抽出方法や結果について教えてください	*In university lectures and others, there are many occurrences of technical terms, and I think that the technical term is often a compound word that combines several words. So I think about experimenting on keyword extraction in morphological analysis of speech recognition of lecture I think that there was a paper that I did. After a while, I will tell you about the keyword extraction method and result*
SQD2-78	(F えー)サブワードを用いた¡H¿音声文書の検索¡H¿において¡H¿(F えー)その検索精度を向上させるために何か有用なものはありますか	*Is there anything useful for improving search accuracy in searching spoken documents using subwords?*

Table 4.20: Length statistics in number of terms for processed queries from the SD2, SQD1, and SQD2 sets.

Topics	Number	Transcript	Ave. len.	S.D. len.	Max len.	Min len.
SD2	120	MAN	6.77	2.62	16	1
SQD1	37	MAN	24.13	11.61	55	8
		MATCH	29.64	15.14	65	9
		UNMATCH-LM	39.10	23.34	108	10
		UNMATCH-AMLM	32.89	19.64	93	7
SQD2	80	MAN	30.77	12.06	67	13
		K-MATCH	30.32	13.67	75	5
		MATCH	36.85	16.65	89	4
		UNMATCH-LM	50.81	28.37	161	8
		UNMATCH-AMLM	42.41	21.09	122	5

Table 4.21: Recognition error rates for the transcripts of the spoken queries from the SQD1 and SQD2 topic sets.

Topics	Transcript	#Terms	WER	UTER	TER	BIA
SQD1	MAN	373	0%	0	0	1.00
	MATCH	490	51.6%	0.31	0.81	0.42
	UNMATCH-LM	871	77.7%	0.44	1.25	0.23
	UNMATCH-AMLM	754	69.1%	0.46	1.10	0.25
SQD2	MAN	714	0%	0	0	1.0
	K-MATCH	766	33.7%	0.24	0.45	0.59
	MATCH	961	49.2%	0.26	0.68	0.47
	UNMATCH-LM	1,763	75.4%	0.38	1.12	0.26
	UNMATCH-AMLM	1,615	66.1%	0.39	0.99	0.29

Table 4.21 reports ASR error rates for the transcription of the spoken queries. The values show that these transcripts present slightly greater error rates than the presentation transcripts indicating that the spoken queries are more difficult to recognise accurately. Similarly to the spoken presentation case, the most accurate transcripts correspond to those obtained with the K-MATCH and MATCH models, while UNMATCH-LM and UNMATCH-AMLM correspond to the noisiest. In contrast, UNMATCH-LM transcripts contain a higher number of errors compared to the UNMATCH-AMLM transcripts in the spoken query case, due to the increased number of insertion errors present in the former.

4.2.4 Relevance assessments

Relevance judgements for the SD2, SQD1, and SQD2 topics were collected by the benchmark organisers from pools of results submitted during the corresponding NTCIR cycle (Akiba et al., 2013a, 2014, 2016). The methodology used by the human assessors to generate the relevance judgements differed slightly between cycles and depended on the specifics of the search tasks for which the assessments were gathered.

The first cycles (Akiba et al., 2011, 2013a) of this benchmark posed an SDR task over the collection of academic presentations, also called "lectures" (LEC) by the task organisers. In these initial cycles, the information units to be ranked were provided in advance to the retrieval systems whose only goal was to produce a sorted ranking of such pre-defined lectures in order of relevance to a query.

Subsequent cycles of the NTCIR benchmarks (Akiba et al., 2014, 2016) evaluated the ability of an SCR system at the task of ranking variable-length passages from the academic presentations. Two variations of passage retrieval tasks were evaluated in these cycles: a slide-group-segment (SGS) task, which required systems to rank a collection of pre-defined segments in order of relevance to a query; and a passage (PAS) task, which required systems to rank arbitrary-sized passages in response to a query.

In the SGS task, systems were only allowed to retrieve results from a pre-defined collection of spoken passages termed "slide-group" segments. A slide-group segment is a span of contiguous utterances (IPUs) produced during the presentation of a group of slides. A sequence of slides form a slide-group when used in a presentation to support the description of a single topic or idea. Thus, a slide-group segment represents a topically homogeneous unit of information. The set of all slide-group segments comprised the collection of units to be ranked used at the SGS task.

The passage (PAS) retrieval task did not impose hard constraints on the size and boundaries of the passages to be returned in response to a query. The goal of the PAS task was to study whether SCR technology was capable of determining the exact location and extent of the relevant content within a presentation. In this task, participating systems were allowed to return passages formed by any number of consecutive utterances (IPUs) from a presentation. Each utterance in this task was considered an atomic and indivisible retrieval unit: systems could group adjacently occurring utterances together but were not allowed to split these into shorter units.

Each search task at the NTCIR benchmarks imposed a different constraint on the type of units to be searched: lectures in the case of LEC, slide-group segments in SGS, and arbitrary passages (constrained by utterances) in PAS. In all cases, relevance assessments were generally carried out at the passage level, either by assessing the relevance of an individual slide-group segment or span of utterances. Relevance assessments at the presentation level were then obtained from those conducted at the most granular levels. For the LEC task, a presentation was deemed as relevant for a query if it contained at least one relevant passage for that query.

Relevance judgments for the SD2 topics were generated from the pool of submissions to the PAS task gathered at this edition of the benchmark. In this case, a fine grained assessment procedure was conducted. This procedure involved the assessment of the IPUs surrounding the top 20 arbitrary variable length passages from each ranked list of submitted results. In these assessment studies, annotators were instructed not only to consider the contents of the passage to determine its relevance status with respect to a query, but also its surrounding context. Thus, a passage was only deemed relevant to a query if there was enough contextual evidence around it to support this fact. For example, for the query *"how can we evaluate the performance of information retrieval?"* an isolated mention of the term *"F-measure"* in a presentation would not be considered relevant if there were not stated previously that the F-measure is in fact a measure of IR effectiveness or if this could not be properly inferred from context.

In the SGS task of the SQD1 and SQD2 benchmarks, slide-group passages were used as retrieval units. For every query in the SQD1 and SQD2 sets, pools of slide-group segments were formed with the top 20 results submitted by each system that participated in the SQD1 and SQD2 cycles. The relevance assessments for a given query in the SGS task were performed at the slide-group level, that is, by assessing the relevance of each slide-group segment included in the pooled results against the pertinent query. Assessors were required to determine the relevance of a slide-group segment based on evidence from: (i) the contents of the query for which the search result was produced; (ii) the contents of the article's paragraph from the SDPWS proceedings that motivated the creation of such a query during the topic generation study; (iii) and the information conveyed by the presenter both in spoken and visual form while presenting the slides associated with the slide-group segment to be assessed. All presentations containing one or more slide-group segments deemed relevant to a query, were given special treatment in the assessment study. These were assessed exhaustively by an assessor, who determined the relevance of every slide-group segment occurring in the presentation.

For the PAS task at the SQD1 benchmark, the boundaries of the passages to be retrieved were unknown to task participants. Because the same set of queries was used for both SGS and PAS tasks, the relevance assessments for PAS task were initially based on those performed for the slide-group segments. A second stage of fine-grained assessment was then conducted that looked at the relevance of each individual IPU within the boundaries of a slide-group segment. In addition, these assessments were extended to cover the IPUs occurring before and after every slide-group segment deemed relevant in order to precisely determine the true extents of the relevant content within the presentation. As a result of this more granular assessment process, relevant slide-group segments can be characterised by their constituent and surrounding relevant IPUs. Note that the boundaries of a relevant slide-group segment may not necessarily align with the start and end of their relevant IPUs. Occasionally, relevant IPUs associated to a relevant slide-group segment extent beyond the boundaries of the segment, reaching the relevant IPUs of the following slide-group segment. Conversely, the relevant IPUs associated with a relevant slide-group segment may only represent a small fraction of all IPUs included in that segment.

In all relevance assessments conducted at the NTCIR SCR benchmarks, three relevance levels were annotated: full (R), partial (P), and no relevance (I). However, details of IPUs assessed as non-relevant were not provided to task participants and are thus not available for the SDPWS2 queries. Moreover, relevance assessments were not conducted at every level of granularity for some of the topic sets. In particular, IPU-level assessments were carried out for the SD2, and SQD1 topics but not for the SQD2 topics. Also, slide-group-level assessments are only available for the SQD1 and SQD2 topics. Table 4.22 shows a

Table 4.22: Availability of relevance assessments (ground truth) for NTCIR topics according to two levels of assessment granularity: slide-group-segments (SGS) and arbitrary passages (PAS).

Topics / Task	PAS	SGS
SD2	✓	
SQD1	✓	✓
SQD2		✓

summary of the ground truth and assessments made available for each topic set.

4.3 Summary

This chapter described two spoken collections in detail, the BBC collection of English TV broadcast, and the SDPWS collections of Japanese presentations. These contain all of the elements required for SCR experimentation and thus allow for the study and comparison of the effectiveness of different retrieval approaches over the same sets of spoken documents, transcripts, queries, and relevance assessments.

The BBC collection contains the recordings of 5,843 TV shows split into two subsets: the BBC1 (1,860) and the BBC2 (3,520). These collections were used at the different cycles of the MediaEval SH tasks. Both subtitles as well as ASR transcripts of the audio material from the TV shows are available. Three topic sets were gathered by the task organisers at the SH13, SH14, and SAVA tasks. The SH13 contains 50 known-item topics, while the SH14 and SAVA contain 36 ad 30 ad-hoc topics respectively. Relevance assessments for these topics were collected via crowd-sourcing experiments with video clips submitted by the participants at the SH13, SH14, and SAVA tasks.

The SDPWS collection used for the different cycles of the NTCIR SpokenDoc benchmarks contains 114 academic presentations from which a subset of 98 presentations (SDWPS2) was used in the experiments described in this thesis. Three sets of ad-hoc topics are also available: SD2, SQD1, and SQD2. While the SD2 set contains 120 typed queries, the SQD1 and SQD2 contain 36 and 80 spoken queries respectively. Manual transcripts as well as ASR transcripts produced by the Julius and Kaldi recognition systems with models of varying quality are available for both spoken presentations and spoken queries. Relevance assessments for the SD2, SQD1, and SQD2 topics were collected by task organisers for a lecture (LEC) retrieval task, a slide-group segment (SGS) retrieval task, and a passage (PAS) retrieval task, and are available for various levels of passage granularity.

The subsequent chapters of this thesis present the experimental work carried out in this PhD with the BBC and SDPWS collections towards seeking answers for the research questions described in Section 1.2.

Chapter 5

Prosodic-based Term Weighting

This chapter describes a series of SCR experiments and analysis conducted with the BBC and SDPWS collections to study the potential utility of prosodic information for improving lexical-based SCR methods. Within the broad range of possible applications of prosodic information in SCR, this investigation focuses on one particular aspect: the use of acoustic prominence as complementary information to term distribution statistics for estimating the weights of topically significant terms occurring in spoken documents and passages.

If it is true that the most prominent words are those that best describe the topic in discourse, an SCR system could potentially exploit this fact to generate better estimates of the importance, or weight, that a term is given in a particular portion of speech. In other words, terms that are made prominent or emphasised by a speaker could be considered more representative of the topic of the content, and hence given increased emphasis in the SCR process. The use of these acoustically enhanced term weights could be then used to rank spoken documents more effectively in order of relevance to the user's query.

In previous work, Silipo and Crestani (2000) studied the extent to which a word's grade of prominence relates to its grade of informativeness by observing the correlation between acoustic scores, derived from manual annotations of syllable stress, and BM25 weights. Guinaudeau and Hirschberg (2011) and Chen et al. (2001) took a step forward in this line of research and attempted to combine a word's prominence score, automatically derived from the speech signal, with a standard TF-IDF score to obtain an enhanced vector representation for documents that could improve their retrieval. These more recent studies reported mixed results; while the acoustically enhanced term weights were shown to be helpful for topic tracking (Guinaudeau and Hirschberg, 2011), they were not useful in a SDR setting (Chen et al., 2001).

The experiments described in this chapter continue where the previous investigations by Silipo and Crestani, Guinaudeau and Hirschberg, and Chen et al. left off, extending them in several ways. First, a series of retrieval experiments are presented that explore whether a similar technique to those proposed by Guinaudeau and Hirschberg (2011), and Chen et al. (2001), hereafter GH and CWL respectively, can be effective in terms of improving upon a lexical-based SCR system. Second, to better understand the relationship between acoustic prominence, informative words, and relevant content, similar analysis to that carried out by Silipo and Crestani (2000) was conducted with speech data from the BBC and SDPWS collections. Finally, the utility of acoustic information was further investigated by training a state-of-the-art learning-to-rank approach to re-rank documents based on term-acoustic information.

This chapter is organised as follows. Section 5.1 describes the acoustic features explored and how these were combined into a prominence score that captures the extent to which a spoken word "stands out" in context. Section 5.2 describes the GH and CWL approaches for integrating prominence scores into existing ranking models, while Section 5.3 presents retrieval experiments conducted with these in spoken document and passage retrieval settings. Data analysis and learning-to-rank experiments are described next in Section 5.4. Finally, Section 5.5 summarises our findings.

5.1 Prominence score computation

The first step towards studying the utility of acoustic prominence for improving existing term weighting schemes is to obtain a representative set of features that reflect the grade of salience of each word spoken in a test collection. Recall from Section 3.3.1 that the grade of salience by which a spoken word is perceived by listeners is mainly influenced by three acoustic correlates of speech prosody: duration, fundamental frequency (F_0), and loudness. This investigation follows previous research (Silipo and Crestani, 2000; Crestani, 2001; Chen et al., 2001; Guinaudeau and Hirschberg, 2011) and explores ways to define and combine these set of acoustic correlates into a numeric score, called a "prominence score" that could be used to study the usefulness of prosodic information for term weight calculation.

This section describes in detail how acoustic correlates of duration, F_0, and loudness were extracted from a speech signal, and how prominence scores were then calculated from them for each spoken word in the BBC and SDPWS collections.

5.1.1 Extraction of low-level descriptors

For each audio file in the speech collections, contours of loudness and F_0 were obtained by using the Open-Source Media Interpretation by Large-space feature Extraction (OpenSMILE)[1] v2.0 toolkit (Eyben et al., 2013). This toolkit provides implementations of standard signal-processing algorithms, including procedures for extracting loudness, and F_0 contours from speech waveforms. In addition, this release of OpenSMILE includes configuration files that define many of the feature-extraction workflows used at the different editions of the Interspeech Computational Paralinguistics ChallengE (ComParE) (Schuller et al., 2017). In order to extract loudness and F_0 correlates with OpenSMILE, we defined a simpler

[1] `http://opensmile.sourceforge.net`

configuration file based on one of the existing configurations provided with the software. What follows is a description of the features extracted with this custom configuration.

While speech is a time-varying signal, speech sounds produced by humans can be considered to remain stationary for short periods of time of about 10-30 ms. Since most signal processing methods assume that the signal under analysis is invariant with respect to time, short-time analysis is commonly performed to analyse the characteristics of speech waveforms. To capture the time-dynamics of the F_0 and loudness correlates, short-time analysis is performed with OpenSMILE by grouping samples into overlapping frames of 50ms length with 40ms of overlap or, equivalently, 10ms of time-shift.

A value of loudness was then calculated for a frame with the OpenSMILE component *cIntensity*. This component obtains an approximation of the loudness as perceived by a human-listener based on a simplified sub-band auditory model (Kießling, 1997). Specifically, this correlate of loudness is calculated as shown in Equation 5.1,

$$E_l = \left(\frac{I}{I_0}\right)^{0.3} \tag{5.1}$$

where I is the signal intensity and I_0 is the reference intensity defined as $I_0 = 10^{-6}$ (Kießling, 1997, 156–157 pp).

For a time-discrete signal $x(n)$ with $n = 0, \ldots, N-1$ representing the speech samples of a frame, OpenSMILE calculates the intensity (I) by first applying a Hamming window function (Young et al., 2002) to $x(n)$ and then computing its normalised energy (E_n) using Equation 5.2.

$$E_n = \frac{1}{N}\sum_{n=0}^{N-1} x^2(n) \tag{5.2}$$

In addition to the approximation of perceived loudness described here, the root-mean squared (RMS) energy (E_{rms}) was extracted for a frame by means of the *cEnergy* component as shown in Equation 5.3.

$$E_{rms} = \sqrt{\frac{1}{N}\sum_{n=0}^{N-1} x^2(n)}, \tag{5.3}$$

The purpose of extracting this second correlate of signal magnitude was to experiment with the same descriptor of signal energy as used by Guinaudeau and Hirschberg (2011) and Chen et al. (2001).

In order to obtain a value of the fundamental frequency (F_0) for a frame, the signal was first passed through a Gaussian filter (Eyben, 2016) by means of the OpenSMILE component *cWindower*. Subsequently, the Fast-Fourier Transformation (FFT) was applied through the *cTransformFFT* and *cFFTmagphase* components to obtain magnitudes and phase values for each frequency band. Finally, the *cPitchACF* component was used to produce a value of F_0 and a probability of voicing (p_v) for the frame. This pitch detection algorithm (Eyben, 2016) uses an auto-correlation method to compute p_v, while estimates F_0 by locating prominent peaks in the signal's Cepstrum (Bogert et al., 1963). As the F_0 values produced by the algorithm can be inaccurate for unvoiced regions of speech, frames with p_v below 0.55 were considered voiceless and assigned a F_0 of 0.

The process described above produced a single value of loudness (E_l), RMS energy (E_{rms}), and F_0 for each frame in a waveform. The series of values corresponding to all frames from an utterance can then be arranged sequentially to form contours of $E_l(n)$, $E_{rms}(n)$, and $F_0(n)$. In order to eliminate possible errors which may occur due to noise perturbations during estimation of the acoustic descriptors, these contours were smoothed by using the *cContourSmoother* component of OpenSMILE with a moving average window of size 3. Background music is yet another factor which may introduce some noise to the extracted features. In the experiments conducted in this thesis, no attempts were made to adjust features for speech regions containing background music. Note that this effect could have affected features extracted for the BBC recordings only. Since SDPWS recordings do not contain background music, features for this dataset were free from such errors.

Figure 5.1 depicts smoothed $E_l(n)$, $E_{rms}(n)$, and $F_0(n)$ contours for an utterance from the BBC collection. The top of the figure shows the utterance's waveform, ASR transcript, and spectrogram. Prominent words in this utterance are those that have increased loudness and pitch values relative to other words. For the utterance in Figure 5.1, the words *MORTAGE*, *GOVERNMENT* and *SO* stand-out in terms of their $E_{rms}(n)$ values. In terms of $F_0(n)$, the top prominent words are *MORTAGE*, *SHOULD*, and *SO*, while for $E_l(n)$ these are *MORTAGE*, *CRISIS*, and *SO*.

5.1.2 Speaker-based standardisation, time-alignment, and word durations

The spoken material from the BBC and SDPWS collections was produced by a large number of speakers in different acoustic conditions and by using different recording devices. On top of this, it is well known that the characteristics of the speech can differ greatly among speakers, due to differences in accent, style, gender, social class, age group, etc. It is therefore paramount to standardise acoustic features appropriately in order to control for these type of variations in the data and to enable fair comparison of feature values across speakers. For this purpose, the loudness, energy, and F_0 contours were standardised based on statistics computed from the regions of speech that are believed to be produced by a single speaker. In this process, each value from a contour $C(n)$ produced by speaker s was replaced by its standard score (Z-score) as shown in Equation 5.4,

$$C_z(n) = \frac{C(n) - \mu_s}{\sigma_s} \tag{5.4}$$

where μ_s and σ_s are the mean and standard deviations calculated from all values in the contour $C(n)$ believed to be produced by speaker s.

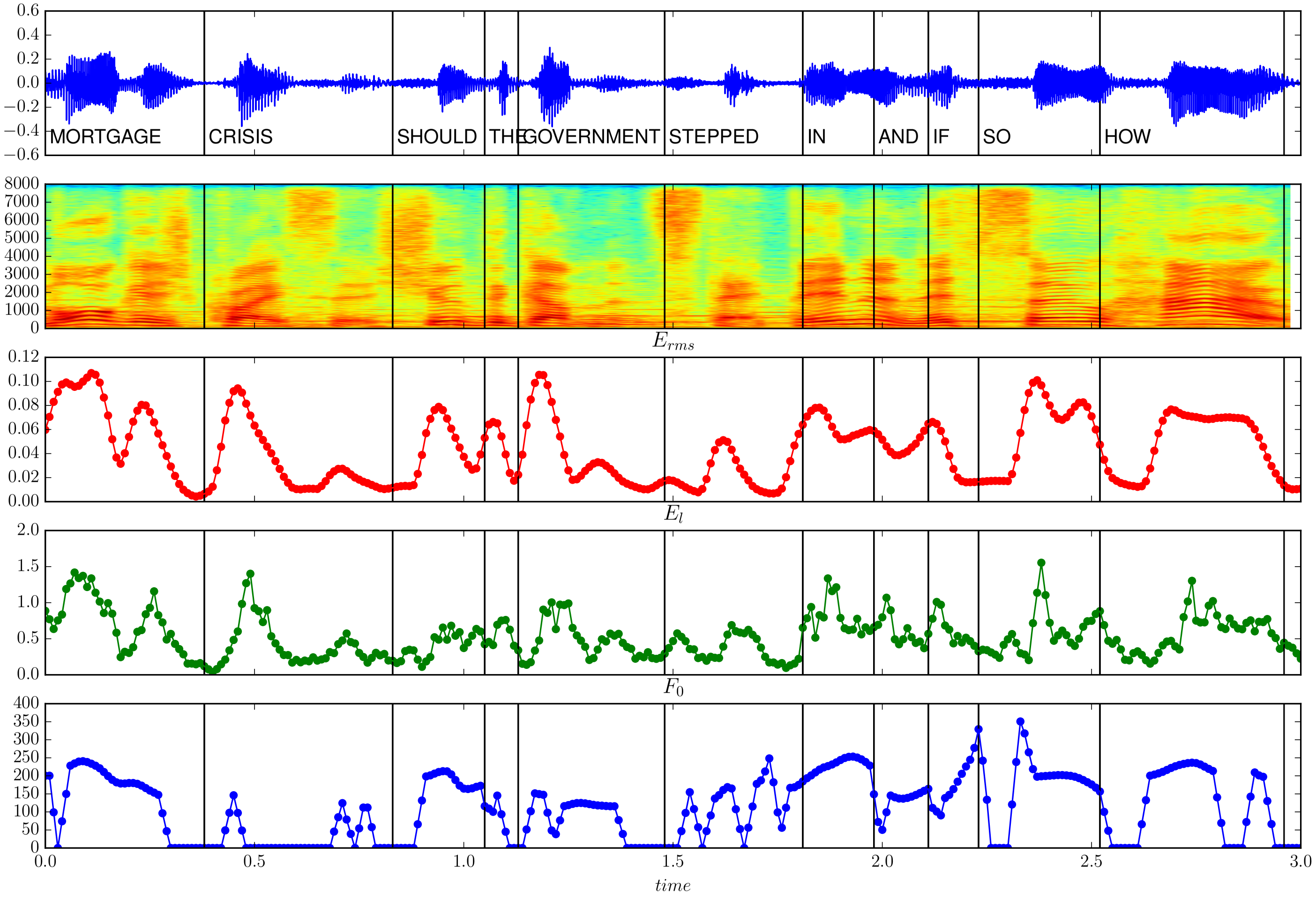


Figure 5.1: Acoustic features extracted with OpenSMILE for an utterance from the BBC collection. The figure shows, from top to bottom: the waveform, spectogram, RMS energy ($E_{rms}(n)$), loudness ($E_l(n)$), and fundamental frequency ($F_0(n)$) contours. The corresponding ASR transcript (NST system) is shown below the waveform along with vertical lines marking word boundaries.

Speaker-based standardisation is only possible when information about "who" spoke "when" is available. In the BBC collection, this is the case for the LIMSI and NST transcripts. In this case, every speech segment is associated with a particular speaker ID, originally produced during the clustering and speaker segmentation processes implemented by these ASR systems. Since the segmentations produced by these systems differ, standardisation was applied to each transcript type separately, based on the speaker diarisation information available in each case. For the transcripts from the SDPWS collection, detailed speaker information is not available nor necessary, since this dataset only contains monologues produced in relatively uniform acoustic conditions. In this case, the feature contours were speaker-standardised by assuming only one speaker per presentation.

Besides information about individual speakers, the majority of the ASR transcripts in the BBC and SDPWS collections contain predicted word timestamps which indicate the starting and duration times of each recognised word. This timing information was used to align words against the feature contours, as shown in the example of Figure 5.1. In this manner, every spoken word was associated with a sequence of values from each feature contour, corresponding to the time when the word is most likely to have been uttered in a speech file.

Besides aligning values of energy, loudness, and fundamental frequency to words, the duration of each word was also extracted from the ASR transcripts as this is also considered an important feature of acoustic prominence. In order to control for variations in speaking rate, the duration estimates were speaker-standardised as was done with the feature contours. Furthermore, extreme durations were frequently assigned by the ASR systems to special words such as long numeric expressions or URLs, and symbols representing filled pauses (e.g. "eh", "mm", "emm") which would sometimes expand to non-speech regions in the output of the ASRs. To avoid considering outliers in the estimation of duration statistics, a practical consideration was taken regarding words assigned extreme incorrect durations by the ASR with only those shorter than 2 seconds being kept in the transcripts.

In the case of subtitles and manual transcripts from the BBC and SDPWS collections, word timestamps were not initially available. In the experiments reported in this thesis, forced alignment was applied to the manual transcripts of the SDPWS collection to obtain timestamp information, as described in Section 4.2.2. Forced alignment was only performed over the SDPWS collection since this process requires accurate transcriptions and an acoustic model to be available, neither of which are available for the BBC material[2]. After obtaining word timestamps for the manual transcripts of the SDPWS collection, the contours and duration features were aligned to each word in the transcripts and words assigned durations greater than 2 seconds were discarded in order to avoid outliers.

[2] Although we could have used some of the pre-trained ASR models for English that are available online, a large number of subtitle files from the BBC collection have erroneous segment-level time-stamps (in some cases captions are more than 10 seconds off w.r.t. the audio track).

5.1.3 Combining low-level descriptors into prominence scores

The previous section described how speaker-standardised duration (D) and contours of loudness ($E_l(n)$), energy ($E_{rms}(n)$), and F_0 ($F_0(n)$) were assigned to individual words in the BBC and SDPWS collections. This section explains how prominence scores were computed from the previously described set of word-level descriptors for these collections.

The standardised contours $E_l(n)$, $E_{rms}(n)$, and $F_0(n)$ contain multiple data points per every occurrence of a word in the speech transcripts. In the Guinaudeau and Hirschberg (2011) (GH), and Chen et al. (2001) (CWL) methods, these contours are aggregated in order to produce a single value of loudness, energy, and F_0 for a word occurrence. For this purpose, GH and CWL experimented with different functions to compute an aggregate score for a contour $C(n)$, particularly: its maximum ($C^{\vee}$), minimum ($C^{\wedge}$), mean (C^{μ}), and standard deviation (C^{σ}). There are therefore 12 possible values that can be derived from these feature aggregations applied to the $E_l(n)$, $E_{rms}(n)$, and $F_0(n)$ contours.

Ultimately, a prominence score for a word should reflect how noticeable or salient that word is, given its acoustic realisation, relative to other words spoken elsewhere. To define such a score, it is necessary to have a reference point or value that could serve as a base for comparing the absolute magnitudes of the word's acoustic descriptors. For instance, the feature values of a word could be compared against those from another word spoken in the same utterance or elsewhere by the same speaker. As a result of the speaker-based standardisation process described in the previous section, any value from a standardised contour $C_{\mathcal{Z}}(n)$ will reflect the difference of its original value ($C(n)$) with respect to the speaker's mean (μ_s), measured in numbers of speaker-dependent standard deviations (σ_s). Thus, based on this fact, any data point in $C_{\mathcal{Z}}(n)$ or derived from it along the boundaries of a spoken word, can in principle be used as the prominence score for it.

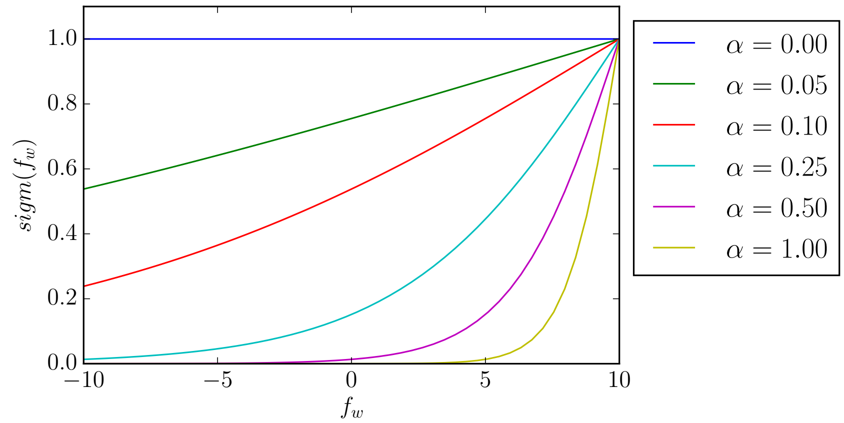
Instead of using a speaker-based point of reference, prominence scores may be defined relative to other values, such as those calculated from all words appearing in the utterance containing the word or the associated document transcript. In the general case, features can be standardised or normalised by considering feature statistics calculated for a restricted set of words $\mathcal{W}$.

In the original implementation of the CWL method, the feature values assigned to a word $w \in \mathcal{W}$ were normalised in the range $[0, 1]$ using a sigmoid function as shown in Equation 5.5,

$$sigm(f_w) = \frac{2}{1 + \exp(-\alpha\ (f_w - \mathcal{W}_f^{\vee}))}, \qquad \mathcal{W}_f^{\vee} = \max_{w \in \mathcal{W}} f_w \tag{5.5}$$

where $\alpha \leq 0$ controls the function's slope, f_w denotes a feature value assigned to w, and $\mathcal{W}$ the set of words from which the maximum feature value ($\mathcal{W}_f^{\vee}$) is calculated. Figure 5.2 shows the shape of the $sigm$ function for different values of α when $\mathcal{W}_f^{\vee} = 10$. The function saturates at 1.0 when given the maximum value of f_w as input and asymptotically decreases towards 0 for decreasingly lower inputs. In the original implementation of CWL,

Figure 5.2: The *sigm* normalisation function (Equation 5.5) for different values of α.



$\mathcal{W}$ was set to all words found in a document transcript. Alternatively, the feature values associated with a word can be range-normalised by using Equation 5.6

$$range(f_w) = \frac{f_w - \mathcal{W}_f^{\wedge}}{\mathcal{W}_f^{\vee} - \mathcal{W}_f^{\wedge}}, \qquad \mathcal{W}_f^{\wedge} = \min_{w \in \mathcal{W}} f_w \tag{5.6}$$

which maps feature values linearly in the $[0, 1]$ interval.

When deciding on a normalisation approach to be applied to prominence scores, it is important to consider the implications of choosing a particular set of words ($\mathcal{W}$) to be used in the calculation of the normalisation statistics. Ideally, $\mathcal{W}$ should contain sufficient words to allow good estimates of the true values of the extremes, means, and standard deviations to be obtained for each feature, so that normalised values can be reliably compared across different sets and documents. In all experiments reported in this thesis with the GH and CWL methods, features were first standardised per speaker, as explained in the previous section, and subsequently normalised in the range $[0, 1]$ using the sigmoid or range functions based on the maximum and minimum values obtained across all words in the collection. While speaker-based standardisation is required to control for speaker-dependent variations, this normalisation process was applied only for the purpose of mapping Z-scores into the more convenient range of values $[0, 1]$.

At this stage in the derivation process, a prominence score for every word occurrence can be defined by using any of the normalised word's features in isolation or in combination. For instance, in the original implementation of the CWL method, the energy and duration features $sigm(E^{\mu}_{rms})$ and $sigm(D)$ were combined with a geometric mean to form the final prominence score for a word occurrence (Chen et al., 2001). In the original GH method, the energy features $range(E^{\mu}_{rms})$ and $range(E^{\vee}_{rms})$ were multiplied against the pitch derived features $range(F^{\mu}_0)$ and $range(F^{\vee}_0)$ respectively to obtain combined scores (Guinaudeau and Hirschberg, 2011). None of these authors provided a clear justification as to why these functions might produce prominence scores that are effective for the underlying retrieval task on which they were evaluated. Nonetheless, it should be mentioned that, as opposed to an arithmetic mean, an geometric mean assigns equal importance to all features involved in its calculation, irrespective of their differences in scale. In a geometric mean, an increment of any of the features by x% always produces the same fixed increment on the final value of the mean. Note that this property also applies to the product-based combination adopted in the GH method, as the geometric mean is the root of the product of features.

5.2 Prominence score integration

Once a prominence score has been calculated for every word occurrence in the collection, the next step is to incorporate it into a retrieval model. This section describes the retrieval models and the different strategies explored in this thesis that seek to integrate prominence scores into the computation of relevance scores for document and passages. Similar to our prominence scores, these models and integration approaches are inspired by those originally proposed in the GH and CWL methods.

5.2.1 General integration approach

In the GH and CWL methods, prominence scores were used within a vector-space model (VSM) for IR. Recall from Section 2.1.2 that a VSM represents documents and queries by vectors in a vector space, where the significance of a term for a given document (query) is expressed as the magnitude or weight assigned to this term's dimension in the vector representation of the document (query), and is estimated by the product between the term's within-document frequency (TF) and inverse document frequency (IDF).

The general approach adopted in the GH and CWL methods to exploit prominence information is based on the assumption that significant terms in a spoken document, i.e. those that best characterise the topic of the document, are those whose occurrences are prominent to a greater extend. Thus, the basic integration approach in the GH and CWL methods simply increases the weights of terms in a document's vector representation that are deemed highly prominent in the document. In the original implementation of GH and CWL, this was achieved by combining prominence scores with TF-IDF scores.

Note that within-document and collection term frequencies are properties of a term, i.e., properties that can be attributed to a stem, lemma or other type of indexing feature uniquely assigned to a document. By contrast, prominence scores are attributes of each individual occurrence of a term and, as such, there may be multiple such scores associated with a given term-document pair. Therefore, any attempt to combine prominence with TF-IDF scores must first decide how to aggregate the multiple scores of a term-document pair into a single value for use in the computation of a query-document matching score. The GH and CWL approaches differ in this aspect as explained in the following sections.

5.2.2 GH's integration approach

In the GH approach, all prominence scores from separate occurrences of a term in a document are aggregated into a single score by computing their mean or by retaining their maximum value (Guinaudeau and Hirschberg, 2011). Formally, the final prominence score for a term i in a document can be given by Equation 5.7 or Equation 5.8

$$ps_{\mu}(i) = \frac{1}{tf_i} \sum_{k} ps(k,i), \tag{5.7}$$

$$ps_{\vee}(i) = \max_{k}\, ps(k,i) \tag{5.8}$$

where $ps(k,i)$ is the prominence score associated with the kth occurrence of term i in the document, and the sum and max shown in the equations range over all occurrences of this term in this document.

The previous possible definitions of the prominence score of a term emphasise two different interpretations of the desirable features of the value a term's prominence score should have for a given document. While $ps_{\mu}(i)$ emphasises that a term's overall prominence score should be high whenever the term is spoken prominently several times in the document, $ps_{\vee}(i)$ supports the interpretation that the overall score should be high whenever any of the occurrences is spoken prominently. Because it is unlikely that speakers will emphasise every single occurrence of the same term they utter, the maximum aggregation seems a more appropriate approach a priori. Furthermore, the prosody with which a term is mentioned will tend to vary across the document depending on factors such as the syntactic role that the word plays in its utterance, or whether it introduces "new" or previously "given" information, or if it corresponds to the first or subsequent mention of the term in the document (Hirschberg, 2002).

Given a definition of $ps(i)$, Guinaudeau and Hirschberg (2011) calculate the weight of the term i for a document $w_{GH}(i)$ using the function in Equation 5.9,

$$w_{GH}(i) = \frac{\theta_{ir}\; w(i) + \theta_{ps}\; ps(i)}{\theta_{ir} + \theta_{ps}} \tag{5.9}$$

where $ps(i)$ is either $ps_{\mu}(i)$ or $ps_{\vee}(i)$, θ_{ir} and θ_{ps} are tuning parameters, and $w(i)$ is a TF-IDF score. The function assigns increased weights to terms that are deemed highly significant not only based on its TF-IDF score but also on its prominence score. Thus, terms that are both highly representative of the document and whose occurrences are acoustically prominent in the document will be assigned greater weight values.

Note that as the parameters θ_{ir} and θ_{ps} in Equation 5.9 are the same for every term and document in the collection, they can be removed from the denominator, and the resulting sum can be replaced with a linear combination as shown in Equation 5.10,

$$w_{GH}(i) \overset{rank}{=} \delta\; w(i) + (1-\delta)\; ps(i), \tag{5.10}$$

where $0 \leq \delta \leq 1$ determines the relative importance that is given to TF-IDF or prominence scores. This alternative is preferred over Equation 5.9 since it has only one free parameter.

In the original implementation of GH, the TF-IDF score $w(i)$ from Equation 5.9 was calculated based on a term-weighting scheme described in (Lecorvé et al., 2008). This weighting-scheme is defined as shown in Equation 5.11.

$$w_{LE}(i) = \frac{\frac{tf_i}{docl}}{\max_{i \in d} \frac{tf_i}{docl}} \log \frac{N}{n_i} = \frac{tf_i}{\max_{i \in d} tf_i} \log \frac{N}{n_i} \tag{5.11}$$

In the topic tracking experiments conducted by Guinaudeau and Hirschberg (2011), the collection size N and document-frequencies n_i from Equation 5.11 were calculated with a corpus of news articles, different than the document collection on which such term weights were later used for their topic tracking experiments.

In all experiments with the GH approach reported in this thesis, the TF-IDF based weights $w(i)$ from Equation 5.10 are computed using the Okapi BM25 function (Equation 2.9). Adopting BM25 weighting also implies that the final relevance score assigned to a document d for a query q is calculated following the probabilistic approach, that is, by using the scoring function shown in Equation 5.12

$$S_{GH}(q,d) = \sum_{i \in q,d} w_{GH}(i) \tag{5.12}$$

instead of the cosine distance measure as implemented in the original VSM-based approach (Section 2.1.2).

Even though the Okapi BM25 and the VSM weighting functions are similar, in the sense that they both compute an addition of weights for coincident terms in the query and document, there are various reasons for preferring Okapi BM25 weighting over the VSM. First, Okapi BM25 has been shown to perform better than VSM in ad-hoc retrieval tasks (Robertson et al., 1994; Buckley et al., 1994). Second, while the weighting scheme proposed by the VSM is mostly based on heuristics, the Okapi BM25 function emerged as an approximation of a well founded theoretical model. As such, the concepts underlying the Okapi BM25 model provide a more useful framework that can serve for the interpretation of retrieval functions.

5.2.3 CWL's integration approach

In the GH approach, the prominence scores from the occurrences of a term in a document are first aggregated into a single score, $ps(i)$, and then combined with the term's TF-IDF weight via Equation 5.10. Note that this equation combines the $ps(i)$ and TF-IDF scores externally, by treating TF-IDF weights and prominence scores as independent sources of evidence that contribute to the value of the term's weight.

As opposed to combining $ps(i)$ and TF-IDF weights externally, in the CWL method, the prominence scores associated with each occurrence of a term in a document are integrated within the calculation of the TF-IDF weights. More specifically, in the CWL method the occurrence-level prominence scores of a term are summed to produce an alternative estimate of the number of times that this term appears in the document. This is more formally shown in Equation 5.13.

$$ps_{\Sigma_0}(i) = \sum_k ps(k,i) \tag{5.13}$$

In the original implementation of CWL, the occurrence scores $ps(k,i)$ from Equation 5.13 are normalised between 0 and 1 by using the *sigm* normalisation function (Equation 5.5).

The summation from Equation 5.13 is subsequently used to compute the TF component of a term's TF-IDF weight. In the CWL method, this is done with the function shown in Equation 5.14.

$$w_{CWL_0}(i) = (1 + \log\ ps_{\Sigma_0}(i))\ \log \frac{N}{n_i} \tag{5.14}$$

An obvious issue with Equation 5.14 is that it can output negative values when $ps_{\Sigma_0}(i) < 1$, which is likely to occur for very infrequent terms. In particular, for terms that only appear once in the document to be scored, the summation $ps_{\Sigma_0}(i)$ from Equation 5.14 becomes $ps(1,i)$, which is likely to be less than 1 if such a prominence score has been normalised between 1 and 0. In the experiments described in this thesis, negative term weights are avoided in the relevance score calculation by using the alternative definition for $ps_{\Sigma_0}(i)$ shown in Equation 5.15.

$$ps_{\Sigma}(i) = \begin{cases} 0 & \text{if} \quad tf_i = 0 \\ 1 + ps_{\Sigma_0}(i) & \text{otherwise} \end{cases} \tag{5.15}$$

Instead of computing term-document weights as in Equation 5.14, the experiments conducted in this thesis are carried out using the Okapi BM25 function (Equation 2.9). The resulting adaptation of BM25 with integrated prominence scores is then shown in Equation 5.16,

$$w_{CWL}(i) = \frac{(k_1 + 1)\ ps_{\Sigma}(i)}{ps_{\Sigma}(i) + k_1\ (1 - b + b\ \frac{docl}{avel})} \quad \frac{(k_3 + 1)\ qf_i}{k_3 + qf_i} \quad cfw(i) \tag{5.16}$$

where the variables k_1, b, k_3, qf_i, *docl*, and *avel* take the same values than in the original BM25 formulation (Equation 2.9). The overall effect of using the quantity $ps_{\Sigma}(i)$ instead of the original frequency counts of the term (tf_i) in Equation 5.16 is to produce term weights that are sensitive to the prominence scores associated with this term in the document. Thus, a term will acquire a high weight value if its associated sum of prominence scores is high.

Recall from Section 2.1.2 that in Equation 2.9 *docl* is the length of the document to be scored, equal to the total number of term-occurrences in the document, while *avel* is the average document length in the collection. In our adaptation of the CWL approach, the values of *docl* and *avel* are still estimated based on the original term counts from each document rather than on a sum of prominence scores. The reason why *docl* and *avel* can be still calculated from the original terms counts is that the ratio *docl*/*avel* will remain approximately the same in either case[3].

Note also that if *sigm* is used to normalise the prominence scores $ps(k, i)$ in Equation 5.15, then the α parameter in *sigm* can be altered to increase or reduce the emphasis that is given to extreme prominence scores. As can be seen from the plots in Figure 5.2, *sigm* becomes approximately constant when α approaches zero and so $ps_{\Sigma_0}(i) \approx tf_i$. Therefore, as α approaches zero the weights computed by the CWL function (Equation 5.16) will approximate those computed by the original BM25 function.

In the experiments reported in this thesis, the final ranking of documents for a query q for the CWL approach is computed with the ranking function shown in Equation 5.17.

$$S_{CWL}(q, d) = \sum_{i \in q,d} w_{CWL}(i) \tag{5.17}$$

Based on this definition, the function $S_{CWL}(q, d)$ will produce greater scores for documents containing a high number of terms with a relatively high summation of prominence scores across occurrences.

5.2.4 A rough interpretation of GH and CWL under the PRF

The approaches presented in Sections 5.2.3 and 5.2.2 seek to incorporate additional occurrence features into the Okapi BM25 retrieval function. While in GH the prominence scores are externally combined with BM25 scores in a linear fashion, in CWL the sum of prominence scores of a term are used in the internal calculation of its BM25 score.

At first glance, these integration approaches may seem impromptu. In fact, no theoretical justifications exist in the literature as to why the scoring functions described in Sections 5.2.3 and 5.2.2 are appropriate for integrating prominence information into a retrieval function, beyond perhaps the intuitive interpretations underlying the application of these integration approaches in the context of a VSM. Under the VSM interpretation, the greater the incidence weight of a term in the representation of a document, the more this document is considered to be about the topic induced by this term. Thus, in the context of a VSM, increasing the incidence weight of a term proportionally to its inferred grade of prominence seems at least intuitively reasonable.

An alternative interpretation of the GH method can be given from the viewpoint of the probabilistic relevance framework (PRF), previously described in Section 2.1.2. Consider the following representation for a document d, $\vec{d} = \langle (d_1, f_1), \ldots, (d_M, f_M) \rangle$ where each d_i is a discrete random variable representing the frequency of a term i in d and each f_i is a

[3] In fact, in the BBC1 collection, the Pearson's r correlation between ratios *docl*/*avel* based on tf_i and $ps_{\Sigma}(i)$ is greater than .999 for documents and .987 for passages.

continuous random variable representing a feature value associated to term i in d. Assume, further, that the variables d_i and f_i are conditionally independent given relevant and non-relevant documents. Starting from the probabilistic ranking principle in Equation 2.3, one can obtain the approximation shown in Equation 5.18.

$$\begin{aligned}
\frac{P(rel|\vec{d},\vec{q})}{P(\overline{rel}|\vec{d},\vec{q})} &= \sum_{i\in q} \log \frac{P(d_i = tf_i, f_i < x_i|rel)}{P(d_i = tf_i, f_i < x_i|\overline{rel})} \\
&\overset{rank}{=} \sum_{i\in q} \log \frac{P(d_i = tf_i|rel)}{P(d_i = tf_i|\overline{rel})} \frac{P(f_i < x_i|rel)}{P(f_i < x_i|\overline{rel})} \\
&= \sum_{i\in q} \log \frac{P(d_i = tf_i|rel)}{P(d_i = tf_i|\overline{rel})} + \log \frac{P(f_i < x_i|rel)}{P(f_i < x_i|\overline{rel})} \\
&\approx \sum_{i\in q,d} w_{BM25}(i) + \sum_{i\in q} ps(i). \qquad (5.18)
\end{aligned}$$

Thus, if the additional term-level features to be incorporated are assumed to be independent from the frequencies by which terms occur in relevant and non-relevant documents, the form of the resulting retrieval function closely resembles that of the GH function (Equation 5.12). Under this interpretation, it can be said that the GH function makes a strong assumption about the prominence score of a term, namely that this score is independent of the number of times that the term appears in a document. This is a potential limitation of the approach since previous research suggests that the prosody of a word is affected by its frequency and predictability (Hirschberg, 2002; Wagner and Watson, 2010).

In the CWL approach, the term frequency counts tf_i are replaced by the quantity $ps_\Sigma(i)$, defined in Equation 5.15. By re-defining the random variables d_i to represent the quantity $ps_\Sigma(i)$ in the document representation, the first steps in the derivation of the 2-Poisson model can be applied to this representation to obtain a result equivalent to that presented in Equation 2.8. However, in this case the term incidence variables d_i are continuous rather than discrete and cannot be strictly assumed Poisson. A possible alternative is to use the Gamma-based approximation $E_i(x) = \lambda_{E_i}^x \; e^{-\lambda_{E_i}} \, \Gamma(x+1)^{-1}$ (Ilienko, 2013) and its analogous for $\overline{E}_i(x)$ in Equation 2.8, and maintain similar Poisson distributional assumptions. Although providing a formal derivation of the CWL formula is beyond the scope of this thesis, it can be argued that a similar approximation to Equation 2.8, in which tf_i is replaced by $ps_\Sigma(i)$, would be also appropriate under these conditions. By including factors for query term frequencies and document length normalisation, the resulting approximation would match the form of the CWL function (Equation 5.16).

An aspect of the CWL approach that is worth noting based on this re-interpretation is that the model does not make explicit use of term frequency counts. This is by design, since the term frequency counts have been removed from the document representation in place of a summation of prominence scores. Although not explicitly modelled, term frequencies are still considered, since the quantity $ps_\Sigma(i)$ is a sum over "tf_i" occurrences of the term. Also, the sum $ps_\Sigma(i)$ is likely to be correlated with tf_i in practice. Despite

Table 5.1: Retrieval tasks, collections, topics, and transcript types in which the GH and CWL functions were evaluated.

Task	**Collection**	**Topics**	**Transcript types**	
			Queries	**Documents**
SDR, SPR	BBC1	SH13	MAN	LIMSI
	BBC2	SH14, SAVA	MAN	LIMSI, NST
	SDPWS2	SD2, SQD1, SQD2	MAN	MAN

this, using distorted term frequencies may negatively impact the ranking effectiveness of the CWL function since the frequency by which a term appears in a document is normally a useful feature for distinguishing between relevant and non-relevant documents.

5.3 Experiments with heuristic retrieval functions

This section presents a series of retrieval experiments conducted over the BBC and SDPWS collections with the GH and CWL ranking functions. These experiments aim to assess whether the GH and CWL functions can benefit from utilising prominence scores, in addition to the standard lexical-based estimates of TF and IDF. The section begins by defining the retrieval tasks, it then describes the test collections and evaluation measures used, and continues by presenting the experiments conducted.

5.3.1 Tasks and test collections

The effectiveness of the GH and CWL functions was studied in two SCR tasks that differ in terms of the type of unit to be retrieved:

- A spoken document retrieval (SDR) task which focuses on the ranking of programme IDs in the case of the BBC1 or BBC2 collections, or presentation IDs in the case of the SDPWS2 collection.

- A spoken passage retrieval (SPR) task which consists of generating a ranking of pre-defined non-overlapping passages extracted from the documents of the BBC1, BBC2, or SDPWS2 collections.

These experiments used the SH13, SH14, and SAVA topic sets for the BBC1 and BBC2 collections, and the SD2, SQD1, and SQD2 topic sets for the SDPWS2 collection. For the SQD1 and SQD2 topic sets, which contain spoken queries, experiments were conducted with the manual (MAN) transcripts of the speech queries only, whereas for the spoken documents, experiments were carried out with the transcript types shown in Table 5.1. The table also summarises all tasks and test collections in which the GH and CWL functions were evaluated. In total, for each task, the retrieval functions were evaluated across eight different test conditions.

Table 5.2: Length statistics of segmented transcripts from the BBC1, BBC2, and SDPWS2 collections.

Collection	Transcript	Segmentation	Passages	Avg. len.	S.D. len.	Max. len.
BBC1	LIMSI	Fixed-length	52,957	102.4	42.8	227
BBC2	LIMSI	Fixed-length	104,500	105.1	42.6	335
BBC2	NST	Fixed-length	105,188	81.6	34.6	191
SDPWS2	MAN	Slide-groups	2,328	74.8	67.6	757

The main motivation for preferring manual over automatic transcripts in the experiments with the SDPWS collection is to isolate the potential effects that the use of prominence information may introduce in the SCR process from external factors caused by ASR errors. As described in Section 3.3, previous research has shown that prominent words with extreme prosodic realisations are more likely to be misrecognised by the ASR (Goldwater et al., 2010). Exploring the quality of SCR methods over error-free transcripts enable us to control for ASR error effects plus additionally to assess how these acoustically-enhanced retrieval methods would perform under ideal conditions. Since reference transcripts with precise word time-stamps are not available for the BBC data, experiments are carried out with ASR transcripts only. Because a large number of BBC recordings are completely out-of-sync (more than 10 seconds shift) with respect to the ASR transcripts and the audio track, forced and flexible alignment techniques would be difficult to apply to obtain word timestamps for the BBC subtitles. For this reason, experiments with prosodic-based techniques on the BBC data were only conducted over ASR transcripts.

The pre-defined passages used in the SPR task were obtained based on different segmentation strategies for the BBC and SDPWS collections respectively. The transcripts from the BBC1 and BBC2 collections were split into non-overlapping segments of 90 seconds length via a conventional sliding-window approach. In this case, the SPR task was re-stated as that of producing a ranking of pre-defined triplets $(id, start, end)$, where id is a programme ID, and $start$ and end indicate the passage's starting and ending times within the TV-programme. The transcripts from the SDPWS2 collection were split based on their associated slide-group segments (SGS), described in more detailed in Section 4.2.2. Thus, the SPR task with the SDPWS collection consisted of producing an ordering of pre-defined slide-group segment IDs.

Table 5.2 provides general statistics about each segmented collection, while statistics about the unsegmented (document) collections were previously summarised in Tables 4.2a, 4.2b, and 4.15. In comparison with their unsegmented counter-parts, the segmented collections contain about 20-30 times more retrieval units (passages).

Evaluation measures

In the SDR task, the quality of a ranking of documents produced for a query was measured in terms of mean average precision (MAP). MAP was also used for evaluation of passage retrieval effectiveness in the SPR task with the SDPWS2 collection. Recall, however, that the organisers of the NTCIR SD2 task did not originally produce relevance assessments for slide-group segments but rather for arbitrary spans of consecutive IPUs. For the purpose of conducting slide-group retrieval experiments with the SD2 topics, the relevance judgements for slide-group segments for a particular topic were inferred based on the relevance status of the IPUs they contain. In particular, a slide-group segment was deemed relevant to a topic whenever one or more of the IPUs falling within the boundaries of that segment were marked as relevant to the topic in the relevance assessments.

Recall from Section 4.1.4, that the relevance assessments for the SH13, SH14, and SAVA topics were originally produced for passages submitted by different SCR systems. Because these systems may have segmented the transcripts of the BBC collections differently, there may not be a 1:1 correspondence between the segments $(id, start, end)$ produced by the windowing segmentation approach described previously and the segments included in the relevance assessments for these topics. Consequently, standard MAP cannot be used to evaluate SPR rankings produced for the SH13, SH14, and SAVA topics.

In order to measure SPR effectiveness for the SH13, SH14, and SAVA topics, a simple extension of AP called "overlap AP" (oAP) (Aly et al., 2013a) was used in which retrieved segments are deemed relevant if they overlap with any region marked as such in the relevant assessments. Note however that there can be multiple segments overlapping with a single relevant region in a ranked list of results. In the original formulation of oAP (Aly et al., 2013a), multiple results overlapping a relevant passage found in the rank list are counted multiple times. By contrast, in the experiments in this thesis, an alternative version of this measure is used where only the top-ranked passage overlapping a relevant one is considered as relevant, to avoid accounting for duplicate relevant results in the calculation of oAP.

Formally, for a ranking $s_1, \ldots, s_N$ of triplets produced for a query and a set $r_1, \ldots, r_R$ of triplets known to be relevant to that query, oAP can be calculated as

$$oAP = \frac{1}{R} \sum_{k=1}^{N} oP[k] \; o_k, \qquad\qquad oP[k] = \frac{1}{k} \sum_{i=1}^{k} o_i,$$

where

$$o_k = \begin{cases} 1 & \text{if} \quad \exists j \leq R : \big\langle \, over(s_k, r_j) \;\wedge\; \forall i < k : \neg over(s_i, r_j) \, \big\rangle \\ 0 & \text{otherwise,} \end{cases}$$

and

$$\begin{aligned} over(s, r) \equiv id(s) = id(r) & \\ \wedge \; \Big(start(s) \leq start(r) \leq end(s) & \\ \vee \; start(r) \leq start(s) \leq end(r) \Big). & \end{aligned}$$

Given the above definition, overlap MAP (oMAP) is defined as the average of oAP across a set of queries.

Baseline results and parameter estimation

To determine the potential utility of the GH and CWL retrieval methods, their effectiveness was compared against that achieved by the standard text-based Okapi BM25 function (Equation 2.9), which does not utilise prominence scores in the estimation of term weights.

Before any experiment can be run with the BM25-based ranking functions, the parameters, b, k_1, and k_3, need to be set to specific values. Based on extensive experimentation in the context of the TREC evaluation campaigns, it is generally recommended to set $b = 0.75$ and $k_1 = 1.2$, while k_3 can be set to zero if queries are known to be short or to a positive value otherwise (Robertson et al., 2009). Nonetheless, adjusting these parameters appropriately for a particular task and test collection can often provide increased retrieval effectiveness in comparison to using the recommended settings, particularly if the latter have been estimated for tasks and collections that are different to the ones being tested (Chowdhury et al., 2002). For this reason, experiments were first carried out with the text-based BM25 function to determine good performing parameter settings for each task and test collection.

Existing approaches to optimising multiple parameters of retrieval models can be classified into two broad categories. Those that try to maximise retrieval effectiveness metrics that are defined over the ranks of the relevant documents (Taylor et al., 2006), such as MAP or NDCG, and those that try to optimise alternative objective functions, commonly designed to correlate well with rank-dependent metrics and to permit, at the same time, the application of gradient-descent methods (Burges et al., 2005).

For tuning of BM25 parameters, a general optimisation method was implemented in the experiments of this thesis which seeks to maximise MAP directly on a given set of queries. This method can be considered a more efficient alternative to exhaustive search since it selectively explores different regions in the search space that seem more likely to contain a global or local optima. The particular optimisation method implemented belongs to the family of unconstrained line search optimisation methods (Luenberger and Ye, 1984), and has been already used to optimise BM25 parameters in previous research (Taylor et al., 2006). The details of this algorithm are presented in Appendix D.

Tables 5.3a and 5.3b show the retrieval effectiveness obtained with alternative and recommended parameter settings for the SDR and SPR tasks respectively when the k_3 parameter is set to zero. In these tables, bold figures and * symbols mark respectively significant ($p < 0.05$) and highly significant ($p < 0.01$) differences based on paired t-tests. These results show that the effectiveness of BM25 varies widely across test conditions and that BM25 performs better when using these alternative settings. In addition, while optimal parameters remain consistent for different transcripts (LIMSI and NST), parameters vary more widely across topic sets. The largest differences are observed between written

Table 5.3: Comparison between Okapi BM25 with TREC's recommended parameter settings ($b = .75$ and $k_1 = 1.2$) and alternative settings (best).

(a) SDR task.

Topics	**Transcript**	**BM25 (best)**			**BM25**
		b	k_1	MAP	MAP
SH13	LIMSI	.47	3.15	.546	.517
SH14	LIMSI	.20	6.40	**.418**	.380
SH14	NST	.26	4.85	**.465**	.427
SAVA	LIMSI	.30	5.65	**.386***	.335
SAVA	NST	.26	10.0	**.383**	.338
SD2	MAN	.66	3.10	.719	.711
SQD1	MAN	.50	4.42	.718	.640
SQD2	MAN	.69	6.16	**.668***	.587

(b) SPR task.

Topics	**Transcript**	**BM25 (best)**			**BM25**
		b	k_1	(o)MAP	(o)MAP
SH13	LIMSI	.80	1.11	.316	.305
SH14	LIMSI	.63	1.04	**.337**	.328
SH14	NST	.65	0.98	**.330***	.322
SAVA	LIMSI	.57	0.74	.304	.292
SAVA	NST	.59	0.75	.242	.237
SD2	MAN	.10	0.73	**.451**	.423
SQD1	MAN	.33	3.28	.241	.210
SQD2	MAN	.75	5.65	**.258***	.227

(SD2) and spoken queries (SQD1 and SQD2), and may be due to differences in length. For SQD1 and SQD2 queries, which contain a greater number of terms than SD2 queries, within-document term frequencies may become increasingly useful for retrieval as they may help distinguish which terms in the query are more discriminative of relevance.

Since in most cases significant improvements can be obtained with the alternative settings shown in the tables, these were then used in all experiments reported hereafter with BM25 derived functions, including those conducted with the GH and CWL functions. The results from Tables 5.3a and 5.3b Retrieval effectiveness

5.3.2 Comparison between GH, CWL, and Okapi BM25

This section presents the results of experiments that compare the effectiveness of the GH and CWL methods with Okapi BM25.

Prominence scores considered

Table 5.4 summarises the possible variations of prominence scores that were explored with the GH and CWL integration approaches. A particular prominence score is derived by applying any of the functions shown in each individual cell of the table in a left-to-right fashion. In the table, the $\circ$ symbol denotes function composition. For instance, the

Table 5.4: Summary of the prominence score derivations and integration approaches explored in the experiments with prominence scores. For each integration method GH and CWL, $ps(i)$ specifies different alternatives for how occurrence-level scores were aggregated into a term-level score. Similarly, $ps(k, i)$ indicates how feature contours were aggregated into an occurrence-level score.

Integration	$ps(i)$	$ps(k, i)$	$C_z(n)$
GH	$\vee, \mu$	$range \circ \{D, \vee, \wedge, \mu, \sigma\}$	E_{rms}, E_l, F_0
CWL	Σ	$sigm \circ \{D, \vee, \wedge, \mu, \sigma\}$	

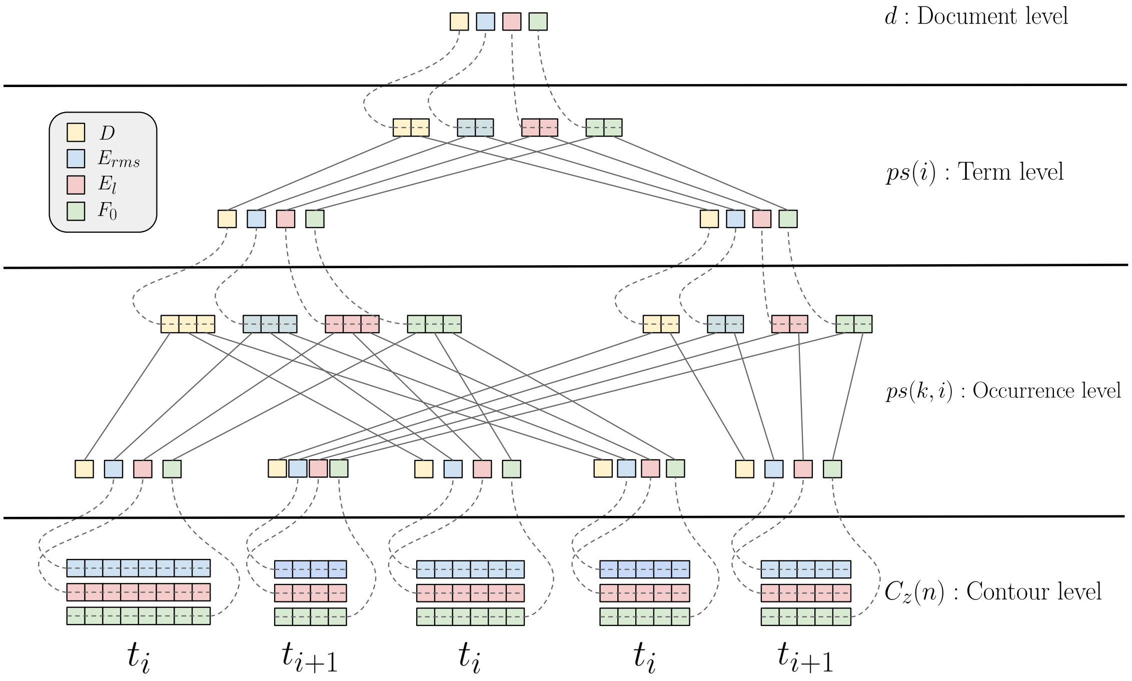
score that results from taking the maximum F_0 score across all occurrences of a term in a document, when each occurrence score is defined as the ranged-normalised minimum value of its speaker-standardised $F_0(n)$ contour, is written as $\vee \circ range \circ \wedge \circ F_0$ or, in short, $\vee \circ range(F_0^{\wedge})$. In what follows, a "base" feature will refer to any feature that can be derived from the family of feature contours E_{rms}, E_l, and F_0, or D.

Figure 5.3 illustrates how the aggregation process of prominence scores was carried out at the contour, occurrence, and term levels. The example from the figure is for a hypothetical document containing three occurrences of a term t_i and two occurrences of second term t_{i+1}, appearing in different positions within the document. At the contour level, the individual occurrences of t_i and t_{i+1} have associated feature vectors $E_{rms}(n)$, $E_l(n)$, and $F_0(n)$, represented in the diagram by arrays of blue, red, and green boxes respectively. At the occurrence level, each of these contour vectors is mapped onto a single prominence score $ps(k, t_i)$ for $k = 1, 2, 3$ and $ps(k, t_{i+1})$ for $k = 1, 2$ by applying an aggregation function ($\vee$, $\wedge$, μ, or σ). In the diagram, the aggregation process is depicted by dashed lines running across an array of values. At the term level, prominence scores for individual occurrences are grouped by feature type and term and then aggregated via $\vee$, μ, or $\sum$. For instance, the three scores derived from F_0 (green boxes) for term t_i are first gathered into a single array of three values and then mapped onto a single F_0 score for term t_i for the document. The aggregation stage at the top of the figure depicts how document-level scores can be obtained from term-level scores. The latter were not used in the experiments reported in this section but in those reported later in Section 5.4.

While term-level scores derived from different base features could be additionally combined to form more complex features, evaluating every possible feature combination via the GH or CWL integrations would not be practical nor would it facilitate the analysis of the performance of individual features. For this reason, the experiments presented in this section evaluate the effectiveness of prominence scores derived from a single base feature.

Tables 5.5a and 5.5b show general statistics about the occurrence-level features obtained for the BBC and SDPWS collections respectively, while Figures 5.4a and 5.4b show how feature values are distributed in these collections. Most features follow a normal distribution with distinctive spread depending on the nature of the feature as well as how it was computed. Features calculated with a max ($\vee$) for each occurrence have the greater spread (increased variance), while those calculated with a mean (μ) and min ($\wedge$) have

Figure 5.3: Visualisation of the multi-level aggregation approach used to calculate prominence scores of terms. The example shows three occurrences of term t_i and two of t_{i+1} in a document d. Dashed lines represent an aggregation function being applied to an array of data points, while continuous lines represent a "copy" operation.



smaller variance. In contrast to the features calculated for the SDPWS collection (Japanese), those calculated for the BBC collection (English) tend to have greater means and ranges (difference between max and min), which suggest these features vary more widely in the English broadcast TV data than in the Japanese monologues.

Table 5.6 shows how the average value of each feature varies across TV show genres in the BBC2 collection. Sports, Quiz, and News shows are among those in which speakers speak generally louder than average (high $E^{\vee}_{rms}$ and $E^{\vee}_{l}$), whilst speakers tend to use lower volumes in Soap opera, Drama, and Children shows. The figures also indicate that News content is characterised by high variations in prosody, as E^{σ}_{rms}, E^{σ}_{l}, and F^{σ}_{0} are greater for this genres. Speech encountered in shows for children and comedy shows are characterised for containing words with shorter duration than on average.

Results of experiments with the GH function

For the experiments with the GH integration approach, there were a total of 26 possible single feature derivations of $ps(i)$ to be tested for the 8 test collections in the SDR and SPR tasks, therefore 416 possible conditions to be evaluated. Tables 5.7a and 5.7b show the best results obtained with the GH integration approach for the SDR and SPR tasks respectively. The results shown for each test condition are for the best performing prominence score

Table 5.5: General statistics (mean, standard deviation, max, min, and 25, 50, 75 percentiles) of occurrence level prominence scores (features) for words in:

(a) the BBC2 collection (11 million words)

Feature	**Mean**	**Std**	**Min**	**25%**	**50%**	**75%**	**Max**
E_l^{μ}	0.02	0.50	-2.38	-0.33	-0.02	0.32	5.12
E_l^{σ}	0.83	0.27	0.01	0.65	0.82	1.00	3.48
$E_l^{\vee}$	1.79	0.94	-2.10	1.15	1.78	2.41	10.07
$E_l^{\wedge}$	-1.24	0.36	-3.67	-1.47	-1.29	-1.07	2.94
E_{rms}^{μ}	0.04	0.61	-3.21	-0.40	-0.02	0.41	7.60
E_{rms}^{σ}	0.73	0.31	0.00	0.50	0.72	0.95	4.30
$E_{rms}^{\vee}$	1.21	0.90	-2.99	0.60	1.23	1.81	10.99
$E_{rms}^{\wedge}$	-1.08	0.51	-5.56	-1.39	-1.19	-0.89	4.53
F_0^{μ}	0.02	0.59	-3.06	-0.40	-0.02	0.39	6.21
F_0^{σ}	0.74	0.32	0.00	0.56	0.77	0.94	3.84
$F_0^{\vee}$	1.03	0.70	-3.06	0.63	0.95	1.36	10.98
$F_0^{\wedge}$	-1.06	0.58	-8.64	-1.38	-1.21	-0.98	5.31
D	0.60	1.01	-2.78	-0.12	0.44	1.14	15.29

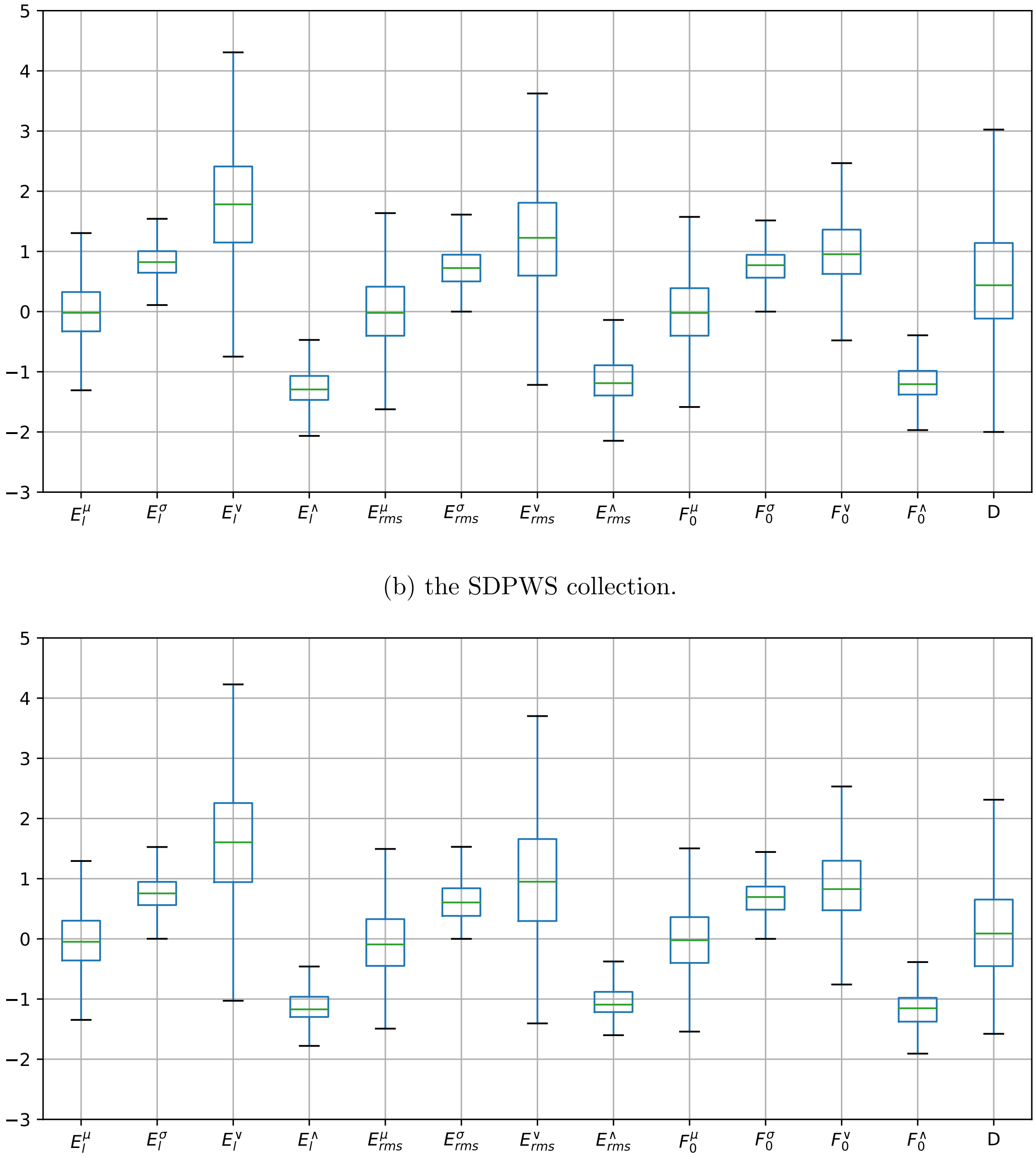
(b) the SDPWS collection (200 thousands words)

Feature	**Mean**	**Std**	**Min**	**25%**	**50%**	**75%**	**Max**
E_l^{μ}	0.00	0.51	-1.49	-0.36	-0.05	0.30	3.75
E_l^{σ}	0.76	0.29	0.00	0.56	0.76	0.95	3.77
$E_l^{\vee}$	1.62	1.02	-1.42	0.94	1.60	2.26	12.62
$E_l^{\wedge}$	-1.07	0.37	-1.78	-1.30	-1.17	-0.96	3.21
E_{rms}^{μ}	-0.02	0.61	-1.49	-0.45	-0.09	0.33	5.81
E_{rms}^{σ}	0.63	0.35	0.00	0.38	0.60	0.84	9.22
$E_{rms}^{\vee}$	1.04	1.07	-1.40	0.30	0.95	1.66	30.42
$E_{rms}^{\wedge}$	-0.96	0.43	-1.60	-1.22	-1.09	-0.88	4.11
F_0^{μ}	-0.01	0.59	-2.08	-0.40	-0.02	0.36	4.40
F_0^{σ}	0.67	0.36	0.00	0.49	0.70	0.87	3.02
$F_0^{\vee}$	0.97	0.95	-2.08	0.48	0.83	1.30	7.13
$F_0^{\wedge}$	-1.03	0.60	-2.08	-1.38	-1.15	-0.98	4.28
D	0.23	1.02	-1.58	-0.45	0.09	0.65	10.01

Table 5.6: Mean values of occurrence-level features in the BBC2 collection for different TV genres.

Genre	E_l^{μ}	E_l^{σ}	$E_l^{\vee}$	$E_l^{\wedge}$	E_{rms}^{μ}	E_{rms}^{σ}	$E_{rms}^{\vee}$	$E_{rms}^{\wedge}$	F_0^{μ}	F_0^{σ}	$F_0^{\vee}$	$F_0^{\wedge}$	D
Chat	0.02	0.82	1.75	-1.21	0.03	0.72	1.15	-1.07	0.01	0.68	0.93	-1.01	0.61
Children	0.02	0.84	1.74	-1.29	0.03	0.75	1.16	-1.16	0.02	0.72	0.98	-1.10	0.52
Comedy	0.02	0.82	1.71	-1.20	0.04	0.72	1.18	-1.04	0.03	0.69	0.95	-1.00	0.53
Documentary	0.03	0.83	1.83	-1.25	0.04	0.73	1.23	-1.09	0.01	0.73	1.08	-1.05	0.62
Drama	0.02	0.80	1.71	-1.22	0.04	0.67	1.10	-1.01	0.01	0.68	1.03	-0.88	0.56
Music	0.02	0.82	1.79	-1.23	0.04	0.71	1.18	-1.06	0.00	0.73	1.08	-1.01	0.61
News	0.02	0.86	1.88	-1.25	0.04	0.78	1.33	-1.12	0.02	0.78	1.04	-1.16	0.64
Quiz	0.03	0.86	1.90	-1.23	0.06	0.78	1.32	-1.11	0.03	0.76	1.04	-1.10	0.58
Reality	0.01	0.80	1.69	-1.19	0.03	0.69	1.12	-1.03	0.01	0.70	0.98	-0.99	0.57
Soap opera	0.03	0.77	1.62	-1.13	0.04	0.63	1.03	-0.90	0.02	0.68	0.99	-0.88	0.53
Sports	0.03	0.85	1.82	-1.25	0.06	0.75	1.24	-1.08	0.04	0.77	1.04	-1.04	0.60

Figure 5.4: Distribution of occurrence level prominence scores (features) for words in:



(a) the BBC2 collection.

(b) the SDPWS collection.

Table 5.7: Comparison of retrieval effectiveness between the best instantiation of GH and Okapi BM25.

(a) SDR task.

Topics	**Transcript**	**GH (best)**					**BM25**
		$ps(i)$	$ps(k,i)$	$C_z(n)$	δ	MAP	MAP
SH13	LIMSI	μ	σ	E_l	0.12	.572	.546
SH14	LIMSI	$\vee$	σ	F_0	0.22	.423	.418
SH14	NST	$\vee$	$\wedge$	E_l	0.31	**.469**	.465
SAVA	LIMSI	$\vee$	σ	F_0	0.09	.391	.386
SAVA	NST	$\vee$	μ	E_l	0.40	.384	.383
SD2	MAN	$\vee$	$\wedge$	E_{rms}	0.26	.722	.719
SQD1	MAN	$\vee$	D	-	0.21	.724	.718
SQD2	MAN	$\vee$	σ	E_{rms}	0.07	.687	.668

(b) SPR task.

Topics	**Transcript**	**GH (best)**					**BM25**
		$ps(i)$	$ps(k,i)$	$C_z(n)$	δ	(o)MAP	(o)MAP
SH13	LIMSI	$\vee$	σ	E_{rms}	0.11	.330	.316
SH14	LIMSI	μ	$\vee$	E_{rms}	0.69	.337	.337
SH14	NST	$\vee$	σ	E_l	0.75	.330	.330
SAVA	LIMSI	$\vee$	$\vee$	F_0	1.00	.304	.304
SAVA	NST	$\vee$	σ	F_0	0.76	.242	.242
SD2	MAN	$\vee$	D	-	0.69	.451	.451
SQD1	MAN	$\vee$	$\vee$	F_0	1.00	.241	.241
SQD2	MAN	$\vee$	D	-	0.88	.258	.258

found among the 26 possible derivations of scores from base features. Also, these results are for optimised values of the δ parameter, which are also depicted in the tables. Recall that the δ parameter in the GH function (Equation 5.10) controls the amount of influence that prominence scores have on the final weight of a term. Values of δ close to 0 signify major contribution from prominence scores and minor contribution from lexical scores.

Based on the MAP scores from Table 5.7a, it can be seen that the GH method provided, in the best case scenario, only minor, mostly non-significant improvements in document retrieval effectiveness over the BM25 baseline. In the SPR task however, differences in MAP were generally minuscule and in no case significant, meaning that the GH retrieval function could not outperform the BM25 baseline, even if using the best possible combination of features and δ values. In addition, the fact that the best values for δ are generally greater in the SPR results than in the SDR results suggests that prominence scores are potentially more effective when used in the latter task, which involves the ranking of larger retrieval units in which all occurrences of a term in a document are considered when computing its aggregated prominence score $ps(i)$.

Results of experiments with the CWL function

In the case of the CWL integration method, there were 13 possible variations of $ps(i)$ and therefore 208 experimental conditions to be evaluated. Tables 5.8a and 5.8b show

Table 5.8: Comparison of retrieval effectiveness between the best instantiation of CWL and Okapi BM25.

(a) SDR task.

Topics	**Transcript**	**CWL (best)**						**BM25**
		$ps(i)$	$ps(k,i)$	$C_z(n)$	k_1	α	MAP	MAP
SH13	LIMSI	Σ	σ	E_l	0.30	1.00	.562	.546
SH14	LIMSI	Σ	σ	F_0	6.34	0.05	.419	.418
SH14	NST	Σ	μ	F_0	4.88	0.07	.465	.465
SAVA	LIMSI	Σ	$\vee$	E_{rms}	1.05	0.18	.392	.386
SAVA	NST	Σ	$\wedge$	F_0	9.50	0.05	.385	.383
SD2	MAN	Σ	σ	E_l	0.35	1.00	.721	.719
SQD1	MAN	Σ	σ	F_0	0.78	1.00	.738	.718
SQD2	MAN	Σ	μ	E_{rms}	0.04	0.95	.686	.668

(b) SPR task.

Topics	**Transcript**	**CWL (best)**						**BM25**
		$ps(i)$	$ps(k,i)$	$C_z(n)$	k_1	α	(o)MAP	(o)MAP
SH13	LIMSI	Σ	σ	E_{rms}	0.08	1.00	.336	.315
SH14	LIMSI	Σ	$\vee$	F_0	1.04	0.00	.337	.337
SH14	NST	Σ	$\vee$	F_0	0.98	0.00	.330	.330
SAVA	LIMSI	Σ	$\vee$	F_0	0.74	0.00	.304	.304
SAVA	NST	Σ	σ	E_l	0.40	0.45	.243	.242
SD2	MAN	Σ	μ	E_{rms}	0.22	0.33	**.458**	.451
SQD1	MAN	Σ	$\vee$	E_l	0.75	0.21	**.256**	.241
SQD2	MAN	Σ	σ	F_0	2.55	0.48	.263	.258

the results obtained by the instantiation of CWL that achieved the highest MAP score considering all possible $ps(i)$ instantiations and values for the k_1 and α parameters. The results obtained in the SDR task suggest once again that the use of prominence scores can only provide minor improvements in retrieval effectiveness over the BM25 baseline.

Compared to GH, the CWL approach obtained slight improvements over BM25 in the SPR experiments with the Japanese collections (last three rows in Table 5.8b). However, these improvements may be attributed to chance as the observed differences are no longer statistically significant at 95% confidence according to a t-test if the experiments are repeated with minor variations of the α parameter in the order of 0.01. In particular, for $\alpha = 0.32$ and $\alpha = 0.34$ and the SD2 topic set, the CWL function achieves MAP scores of 0.4559 and 0.4557 respectively, with p-values of 0.07 and 0.14 based on paired t-tests. The fact that the best value of α was zero in most of the SPR runs with the BBC collection indicates that the alternative prominence-based estimates of term frequency provide no benefit over the original frequency estimates in these evaluation conditions.

5.3.3 Comparison between acoustic and randomised scores

The results presented so far cast doubt upon the utility of prominence scores as defined previously, based on simple aggregations of a basic set of acoustic features. In order to evaluate whether these scores are meaningful, experiments were conducted to compare the effectiveness of the GH retrieval function with prominence scores defined: (i) randomly; or (ii) based on any of the acoustic scores from Table 5.4.

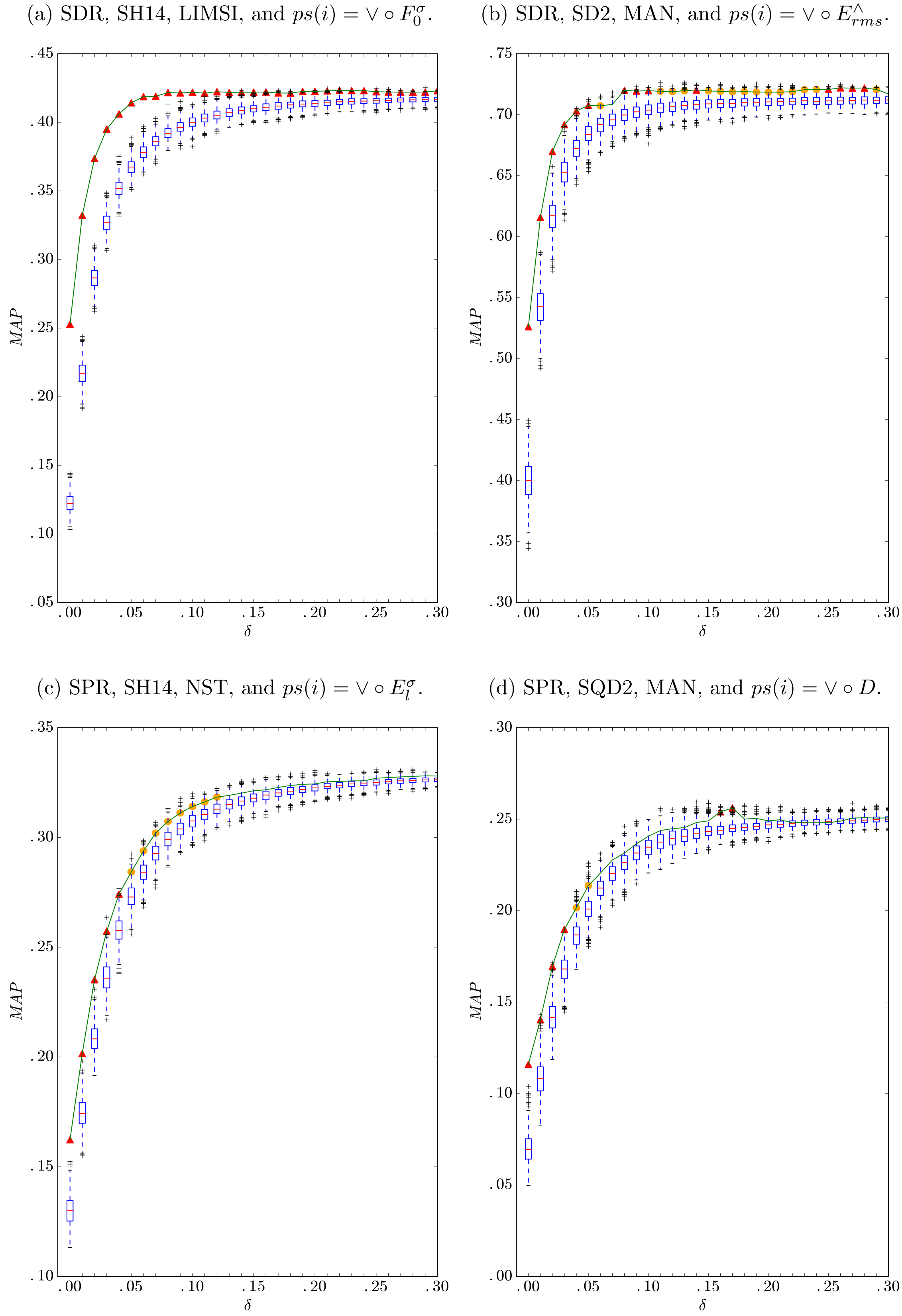
Recall that the δ parameter in the GH function (Equation 5.10) controls the extent to which the score $ps(i)$ affects the overall weight estimation of a term. To make a fair comparison between random and acoustic scores, it is important to ensure that they both produce the same degree of impact on a term's overall weight when used in Equation 5.10. Thus, besides using the same value for δ, a fair comparison also requires random scores to be similar to the acoustically-motivated ones in terms of scale and distribution. Note further that the scores from Table 5.4 may be distributed differently across acoustic features, despite these having been normalised to values between 0 and 1, and that these distributions could possibly vary across languages and word classes. To account for these factors, the random scores used in the following experiments were generated for a particular instance of $ps(i)$, collection C, and query terms $\mathcal{Q}$ from a topic set, based on a random permutation of the acoustic scores $ps(i)$ that are assigned to any term $i \in \mathcal{Q}$ appearing in any document $d \in C$. That is, the random score $rs(i)$ assigned to term i in document d was uniformly sampled from the set $\{ps(k) : k \in \mathcal{Q} \cap d' \ \wedge \ d' \in C\}$.

A permutation experiment evaluated the GH function with 1000 random permutations of acoustic scores for a fixed δ. The resulting distribution of MAP scores was then used to calculate a p-value equal to the proportion of MAP scores from the distribution that were greater than the MAP score obtained with the original (non-random) assignment of the acoustic scores.

Figure 5.5 shows the results of the permutation experiments for four representative conditions of tasks, test collections, and $ps(i)$ derivations. The plots showcase how MAP scores vary as a function of δ, with green lines showing the effectiveness of GH when using an acoustically-motivated score, and each box plot showing the distribution of MAP scores obtained from using the random permutations. Orange circles and red triangles in the plots mark points at which the estimated p-values are less than 0.05 and 0.01 respectively.

Two important observations can be made from these results. First, as expected, the effectiveness of GH degrades with decreasing δ as the influence of the randomised scores increases in the estimation of the term weights. Note however that this degradation is not evident until δ is small enough, since the weights produced by the BM25 function are on a larger scale than those derived from the acoustic features, which range between 0 and 1. Second, although retrieval effectiveness also decays when the non-randomised acoustic scores are used (green lines), these still provide substantially better results than if using the randomised scores, especially for very small values of δ.

The case when $\delta = 0$ deserves a special mention since it corresponds to the instantiation of GH that assigns term weights solely based on prominence scores and which, consequently, ranks documents (passages) according to the sum of their query terms' prominence scores. The plots in Figure 5.5 show that the non-random assignment of



(a) SDR, SH14, LIMSI, and $ps(i) = \vee \circ F_0^\sigma$.

(b) SDR, SD2, MAN, and $ps(i) = \vee \circ E_{rms}^\wedge$.

(c) SPR, SH14, NST, and $ps(i) = \vee \circ E_l^\sigma$.

(d) SPR, SQD2, MAN, and $ps(i) = \vee \circ D$.

Figure 5.5: Effectiveness of GH with acoustic scores (green lines) and random scores (box plots) for the experimental conditions shown in rows 2 (a) and 6 (b) of Table 5.7a and 3 (c) and 8 (d) of Table 5.7b.

acoustic scores performs significantly better than a random assignment when $\delta = 0$. This last observation is important since it suggests that prominence scores derived from acoustic features may be able to capture, to some extent, information about terms that is useful for ranking spoken documents (passages) in order of relevance to a query, similar to the kind of information that is captured by TF-IDF estimates used in the Okapi BM25 ranking function.

5.3.4 Comparison between acoustic scores and other weighting schemes

In the experiment from Section 5.3.3, terms that matched the query were randomly assigned acoustic-derived prominence scores and their effectiveness compared against non-randomised prominence scores with the GH retrieval function. This comparison focused on small values of δ, since these best demonstrate the potential impact that prominence scores can have on the final ranking of documents and passages. Particularly when δ is zero, the GH ranking function (Equation 5.10) becomes $\sum_{i \in q,d} ps(i)$ and produces relevance scores for documents (passages) exclusively based on a sum of prominence scores, without making use of the term's TF-IDF scores.

Compared to using randomised scores, a more effective yet trivial weighting scheme consists of assigning each term a unit weight, i.e. $w(i) = 1$ whenever $tf_i > 0$ or $w(i) = 0$ otherwise. The document scoring function that results from adopting this scheme is known as coordinate matching (CM), and ranks documents (passages) according to the number of query terms they contain, thus essentially considering all terms equally important in the ranking process. If it is true that the acoustic-based prominence scores can provide useful information about the relative importance that terms should be given in the scoring process, then they should be, at the very least, more effective than unit weights.

To test this hypothesis, experiments were conducted that compare the effectiveness achieved by using acoustic scores in the GH function when $\delta = 0$ against that achieved when using CM weighting. Inbetween CM and Okapi BM25, two intermediate weighting schemes are also worth considering in this analysis: the binary independence model (BIM) in which terms are only differentiated by their IDF scores (Equation 2.7) and a "TF-only" (TFO) model which differentiates terms across documents by considering their within-document frequencies but not their document frequencies (Equation 2.9 with $ctf(i) = 1$).

Tables 5.9a and 5.9b show the results obtained with the GH, CM, BIM, and TFO retrieval functions in the SDR and SRP tasks respectively. Similarly to the results reported earlier with the GH method, the results shown are for the acoustic features that performed best in each test collection, with the difference that δ was set to zero in the GH function. In these tables, bold fonts and the * symbol respectively mark significant ($p < 0.05$) and highly significant ($p < 0.01$) differences with respect to the MAP scores obtained by GH.

As can be seen from the results for the SDR task shown in Table 5.9a, the rankings based on prominence scores (GH) are consistently better than those achieved using CM. Furthermore, the acoustic-based weights are frequently more effective in the SDR task

Table 5.9: Comparison of retrieval effectiveness between the GH, CM, BIM, and TFO functions, when $\delta = 0$ and the best derivation for $ps(i)$ is used in GH.

(a) SDR task.

Topics	**Transcript**	**GH (best when $\delta = 0$)**				**CM**	**BIM**	**TFO**
		$ps(i)$	$ps(k,i)$	$C_z(n)$	MAP	MAP	MAP	MAP
SH13	LIMSI	$\vee$	$\vee$	E_l	.358	**.219***	**.239***	**.496**
SH14	LIMSI	$\vee$	$\vee$	F_0	.254	**.145***	**.158***	**.403***
SH14	NST	$\vee$	$\vee$	F_0	.248	**.143***	**.153***	**.439***
SAVA	LIMSI	$\vee$	σ	E_l	.216	**.157***	.183	**.362***
SAVA	NST	$\vee$	$\vee$	E_l	.216	**.133***	**.154**	**.330***
SD2	MAN	$\vee$	D	-	.644	**.512***	.621	.651
SQD1	MAN	$\vee$	D	-	.574	**.426***	.494	**.664**
SQD2	MAN	$\vee$	σ	E_l	.579	**.435***	**.475***	**.675***

(b) SPR task.

Topics	**Transcript**	**GH (best when $\delta = 0$)**				**CM**	**BIM**	**TFO**
		$ps(i)$	$ps(k,i)$	$C_z(n)$	MAP	MAP	MAP	MAP
SH13	LIMSI	$\vee$	$\wedge$	F_0	.210	.184	.242	.249
SH14	LIMSI	$\vee$	$\vee$	F_0	.201	**.169***	.205	**.292***
SH14	NST	$\vee$	$\wedge$	F_0	.183	.163	.191	**.289***
SAVA	LIMSI	$\vee$	$\wedge$	E_{rms}	.159	.133	**.196**	**.244***
SAVA	NST	$\vee$	$\wedge$	E_{rms}	.139	.123	**.172***	**.203***
SD2	MAN	$\vee$	μ	F_0	.274	**.254**	**.415***	.285
SQD1	MAN	μ	$\vee$	E_l	.145	**.093**	.158	.118
SQD2	MAN	μ	D	-	.117	.101	**.157***	.142

than those estimated with the BIM. Note, however, that the BB1, BBC2, and SDPWS2 collections only contain 1860, 3520, and 98 documents respectively, which makes them relatively small compared to most traditional test collections used in IR research. In these circumstances, any document-frequency derived score is likely to be poorly estimated, which may explain why the BIM performed similarly to CM in the SDR task.

The results obtained for the SPR experiments shown in Table 5.9b, indicate that the prominence scores were less effective in this task than in the SDR task, as the effectiveness of the GH method was in general closer to that achieved by CM and lower than that obtained by BIM. These differences may be explained by the following hypotheses:

(i) The weights produced by CM are relatively more effective at ranking passages than full-documents given that there will likely be fewer candidate passages in a collection than documents containing all (or most terms) from the query.

(ii) The weights produced by BIM are more effective in the SPR task than in the SDR task because document frequency estimates will be more reliable when calculated from a larger collection containing significantly more retrieval units.

(iii) The weights based on prominence scores (GH) are more effective in the SDR task than in the SPR task since some of the acoustic features used for this purpose are more meaningful when aggregated over all spoken occurrences of the same term found

Table 5.10: Relative deterioration in retrieval effectiveness for the SDR and SPR tasks when using the simpler weighting schemes CM, BIM, and TFO, instead of BM25.

Topics	**Transcript**	**CM**			**BIM**			**TFO**		
		SDR	**SPR**	**diff**	**SDR**	**SPR**	**diff**	**SDR**	**SPR**	**diff**
SH13	LIMSI	62%	44%	18%	58%	27%	31%	13%	24%	-11%
SH14	LIMSI	66%	50%	16%	63%	39%	24%	5%	13%	-9%
SH14	NST	69%	51%	19%	67%	42%	25%	6%	13%	-6%
SAVA	LIMSI	60%	56%	4%	53%	35%	18%	7%	20%	-12%
SAVA	NST	65%	49%	16%	60%	29%	31%	14%	16%	-2%
SD2	MAN	29%	44%	-15%	14%	8%	6%	10%	37%	-27%
SQD1	MAN	41%	61%	-20%	32%	34%	-3%	8%	51%	-43%
SQD2	MAN	37%	61%	-24%	31%	39%	-9%	2%	45%	-43%

in a document, than when aggregated over a limited number of such occurrences appearing in a passage.

With respect to (i) and (ii), Table 5.10 shows the relative decrease in MAP when CM, BIM, and TFO are used instead of BM25 in the SDR and SPR tasks. For instance, in the SH13-LIMSI condition (row 1 in the table), CM underperforms BM25 by $.546 - .217 = .321$ points absolute, which corresponds to a 62% loss in MAP relative to the .546 figure obtained by BM25. For the experiments with the BBC collections (rows 1-5 in the table), the MAP values indicate that the performance gap between CM and BM25 is 14% greater on average for the SDR task than in the SPR task, whereas between BIM and BM25 the gap is on average 25% greater for the SDR task than in the SPR task. While these results seem to suggest that claims (i) and (ii) hold, the differences in the last three rows in the table (rows 6-8) show a different trend and indicate that (i) and (ii) are not always true. More importantly, if the MAP scores of CM and BIM shown in Table 5.9a were to be adjusted (increased) to account for the observed cross-task differences from Table 5.10, the MAP values of GH would still be higher than those of CM, but lower than those of BIM for rows 1-6, and substantially higher than both for rows 7-8. Therefore, while the observed differences between GH and BIM are probably due to (ii), it is unlikely that the differences between GH and CM in the SDR task can be attributed only to (i).

In order to validate (iii), the results obtained with the GH function with $\delta = 0$ were grouped by feature configuration and then compared against those obtained with CM for every test collection. Tables 5.11a and 5.11b depict the number of test collections on which GH obtained significant improvements ($p < 0.05$) over CM for every derivation of $ps(i)$. For instance, when prominence scores were derived as $ps(i) = \vee$ and $ps(k,i) = E_l^{\vee}$, the GH ranking function obtained significantly higher MAP scores for all test collections for the SDR task (8/8), while it did so in 6 test collections (6/8) when the scores were derived as $ps(i) = \vee$ and $ps(k,i) = E_l^{\wedge}$. The following observations can be made based on the results from Tables 5.11a and 5.11b:

(I) The weights based on prominence scores (acoustic features) are consistently more effective than the use of uniform weights (CM) for the SDR task.

Table 5.11: Number of test collections on which the GH retrieval function (when $\delta = 0$) is significantly more effective than CM ($p < 0.05$) for prominence scores derived from:

(a) Loudness (E_l), energy (E_{rms}), and fundamental frequency (F_0).

$ps(i)$	$ps(k,i)$	SDR			SPR		
		E_l	E_{rms}	F_0	E_l	E_{rms}	F_0
$\vee$	$\vee$	8	7	8	1	0	1
	μ	8	7	8	0	0	3
	σ	8	7	6	0	0	0
	$\wedge$	6	4	5	0	0	1
μ	$\vee$	1	1	0	1	0	0
	μ	2	1	2	0	0	0
	σ	1	1	0	1	0	0
	$\wedge$	0	0	0	0	0	0

(b) Duration (D).

$ps(i)$	SDR	SPR
$\vee$	8	1
μ	0	1

(II) Acoustic-derived weights were consistently more effective when defined as the maximum prominence score ($ps(i) = \vee$) among all occurrences of a term in a document. While when defined as the mean ($ps(i) = \mu$) over all occurrence scores they were less effective.

(III) In the SDR task, effective weights can be derived from every "base" feature: E_l, E_{rms}, F_0, and D. This is true irrespective of which aggregation function is applied in the calculation of an occurrence's prominence score ($ps(k,i)$), although for occurrence-level scores the maximum ($\vee$), mean (μ), and standard deviations (σ) are frequently more effective than the minimum ($\wedge$).

Overall, the previous observations suggest that terms that are significant or informative from an IR perspective tend to be those that are spoken prominently in a particular mention within the entire document, rather than those spoken prominently on average. Thus, the maximum ($\vee$) prominence score across all mentions of a term in a document, or equivalently, the score assigned to the term's most prominent mention, seems to be a better descriptor of a term's level of significance. Furthermore, the acoustic-derived weights are not as useful for ranking passages as they are for ranking documents, meaning that a term's significance level may not necessarily be signalled in mentions that occur within a relevant passage but in those located elsewhere within the container document. Finally, from observation (III) it follows that effective term weights can be derived from multiple sources of acoustic information (duration, pitch, and loudness), and that a combination of features may provide additional improvements in retrieval effectiveness.

Document-level versus passage-level aggregations

The fact that acoustically-derived term weights are more effective in the SDR task suggests that acoustic features should be aggregated at the level of documents rather than passages. It remains a question though whether utilising these document-level aggregates can result in improved effectiveness in the SPR task.

To test this hypothesis, the effectiveness of the GH function (when $\delta = 0$) with these two aggregation approaches was compared in the SPR task. Table 5.12a and 5.12b summarise the results of such comparisons. In particular, the tables report, for each feature, the number of test conditions on which using the feature resulted in significant improvements when aggregated at the level of documents instead of passages ($doc > pas$) and vice versa ($doc < pas$).

In most cases, there were no significant differences between using document and passage level aggregates ($p \geq .05, doc = pas$). However, for conditions in which such differences were seen to be significant, the document-level aggregates resulted in increased effectiveness more often than their passage-level counterparts. This is particularly evident for $ps(i) = D$ and $ps(k,i) = \sigma$ (last 4 rows in the tables), and $ps(i) = \vee$, $ps(k,i) = \vee$ (first 3 rows in Table 5.12a).

Overall, the figures suggest that document-level aggregates generally provide more effective term weights for the SPR task than passage-level aggregates. This is also evidenced by the fact that document-level aggregated features help close the performance gap between the GH, BIM and TFO functions in the SPR task. The latter effect can be seen by comparing the results from Tables 5.13 and 5.9b.

CWL versus simple weighting schemes

The previous experiments shed light on the meaningfulness of prominence scores that are aggregated via maximum or mean scores across a term's occurrences. In the case of the CWL integration approach, occurrence-level scores are summed instead. This results in different scores than those obtained via max or mean aggregations. In particular, since the summation is applied across the term's occurrences, its resulting value will be correlated with the term's within-document frequency. Thus, even though the within-document frequency of a term is not explicitly used in the calculation of the term's CWL weight (Equation 5.16), the weight produced by this function will still capture much of the same information that the original term-frequency count can capture about the importance of this term.

If the summation of prominence scores for a term does truly provide stronger evidence of its significance compared to that obtained from using within-document term frequencies, then the former should provide greater retrieval effectiveness than the latter. In order to establish a more direct comparison between the summation of prominence scores and the original within-document term frequencies, the CWL retrieval function was compared against TFO by, in this case, setting the IDF factor in CWL (Equation 5.16) to 1 ($ctf(i) = 1$). Tables 5.14a and 5.14b depict the results of this experiment.

As can be seen from the results, the CWL weights perform similarly to TFO's in the large majority of the test conditions. This means that using a sum of prominence scores as an estimate of a term's within-document frequency, as implemented by the CWL function, performs at best as well as if using the original term frequency values. The only exceptions

Table 5.12: Comparison between document-level and passage-level aggregated features in the SPR task. Columns 3-5 of each table show the number of test collections (evaluation conditions) on which the GH function: (i) is equally effective when using document-level and passage-level features ($doc = pas,\ p \geq .05$); (ii) obtains significantly higher MAP when using document-level instead of passage-level features ($doc > pas,\ p < .05$). (iii) obtains significantly lower MAP when using document-level instead of passage-level features ($doc < pas,\ p < .05$).

(a) $ps(i) = \vee$

$ps(k,i)$	$C_Z(n)$	$doc = pas$	$doc > pas$	$doc < pas$
$\vee$	E_l	7	1	0
	E_{rms}	6	2	0
	F_0	7	1	0
$\wedge$	E_l	8	0	0
	E_{rms}	7	1	0
	F_0	8	0	0
μ	E_l	8	0	0
	E_{rms}	8	0	0
	F_0	8	0	0
σ	E_l	7	1	0
	E_{rms}	4	4	0
	F_0	4	4	0
D		5	2	1
Total	(104)	87	16	1

(b) $ps(i) = \mu$

$ps(k,i)$	$C_Z(n)$	$doc = pas$	$doc > pas$	$doc < pas$
$\vee$	E_l	5	2	1
	E_{rms}	6	2	0
	F_0	4	2	2
$\wedge$	E_l	5	1	2
	E_{rms}	3	2	3
	F_0	5	1	2
μ	E_l	4	2	2
	E_{rms}	4	2	2
	F_0	5	1	2
σ	E_l	6	2	0
	E_{rms}	6	2	0
	F_0	4	4	0
D		6	2	0
Total	(104)	63	25	16

Table 5.13: Comparison of retrieval effectiveness between the GH with document-level aggregates, CM, BIM, and TFO functions in the SPR task. Results for GH were obtained with $\delta = 0$, document-level feature aggregations, and with the derivation of $ps(i)$ that provides the highest MAP in each test condition.

Topics	**Transcript**	**GH (best when $\delta = 0$)**				**CM**	**BIM**	**TFO**
		$ps(i)$	$ps(k,i)$	$C_z(n)$	MAP	MAP	MAP	MAP
SH13	LIMSI	$\vee$	μ	E_{rms}	.200	.184	.242	.249
SH14	LIMSI	$\vee$	$\vee$	F_0	.206	**.169***	.205	**.292***
SH14	NST	$\vee$	$\vee$	E_{rms}	.194	**.163**	.191	**.289***
SAVA	LIMSI	$\vee$	$\wedge$	E_{rms}	.165	**.133**	.196	**.244***
SAVA	NST	$\vee$	$\wedge$	E_{rms}	.133	.123	**.172***	**.203***
SD2	MAN	μ	D		.304	**.254***	**.415***	.285
SQD1	MAN	μ	D		.146	**.093***	.158	.118
SQD2	MAN	$\vee$	D		.129	**.101**	**.157**	.142

occur in experiments with SD2 and SQD1 topics. However, this was only the case for the first and second instantiations of $ps(i)$ while, for the remaining 11 instantiations, the MAP values were not significantly different from those obtained by TFO.

5.3.5 Experiments with feature combinations

The experiments described in Sections 5.3.2, 5.3.3, and 5.3.4 with the GH and CWL approaches explored the potential effectiveness of using prominence scores derived from a single "base" feature of loudness (E_l), energy (E_{rms}), fundamental frequency (F_0), or duration (D). Because prominent words are likely to be realised by a combination of such features, instead of any of them in isolation, it is worth investigating whether prominence scores defined through feature combinations may result in term weights that are more effective at characterising significant terms from non-significant ones. In fact, the results from Table 5.11a, suggest that weights that are more effective than uniform weights can be derived independently from different base features. Thus, a prominence score based on a combination of acoustic features, either derived from multiple base features or by applying different aggregation functions over the same base feature, may produce improved term weights. This section reports on experiments carried out with prominence scores defined through such feature combinations.

Inner and outer combinations

Recall from Section 5.3.2 that a term's prominence score was derived in a simple multi-stage aggregation process. First, a feature-contour $C_z(n)$ associated with the kth occurrence of the term was aggregated into an occurrence score $ps(k,i) = \oplus_n C_z(n)$ via an aggregation function $\oplus_n$ across element indices $n = 1, 2, \ldots$. Second, these scores were aggregated across occurrences to obtain a term score $ps(i) = \oplus_k ps(k,i)$. This process was illustrated in Figure 5.3, while the possible aggregation functions explored were summarised in Table 5.4.

Table 5.14: Comparison of retrieval effectiveness between the CWL and TFO functions when $ctf(i) = 1$ and the best derivation for $ps(i)$ is used in CWL.

(a) SDR task.

Topics	**Transcript**	**CWL (best when $ctf(i) = 1$)**				**TFO**
		$ps(i)$	$ps(k, i)$	$C_z(n)$	MAP	MAP
SH13	LIMSI	Σ	σ	F_0	.514	.496
SH14	LIMSI	Σ	μ	F_0	.404	.403
SH14	NST	Σ	$\vee$	F_0	.439	.439
SAVA	LIMSI	Σ	$\vee$	F_0	.362	.362
SAVA	NST	Σ	D	-	.327	**.330***
SD2	MAN	Σ	σ	E_l	**.664**	.651
SQD1	MAN	Σ	D	-	.679	.664
SQD2	MAN	Σ	$\vee$	E_l	.683	.675

(b) SPR task.

Topics	**Transcript**	**CWL (best when $ctf(i) = 1$)**				**TFO**
		$ps(i)$	$ps(k, i)$	$C_z(n)$	MAP	MAP
SH13	LIMSI	Σ	$\vee$	F_0	.254	.249
SH14	LIMSI	Σ	$\vee$	F_0	.292	.292
SH14	NST	Σ	$\vee$	F_0	.289	.289
SAVA	LIMSI	Σ	$\vee$	F_0	.244	.244
SAVA	NST	Σ	σ	F_0	.205	.203
SD2	MAN	Σ	$\vee$	E_l	**.290**	.285
SQD1	MAN	Σ	D	-	**.130**	.118
SQD2	MAN	Σ	σ	F_0	.145	.142

Given a set of occurrence-level features $\mathcal{F} = \{f_1, f_2, \ldots\}$ each assigned to every occurrence of a term in a document, so that $f_h(i, k)$ denotes the value that feature $f_h \in \mathcal{F}$ acquires for occurrence k of term i. A prominence score can then be obtained for term i based on a combination of its associated $f_h(i, k)$ values. In order to obtain a single prominence score for term i, the values $f_h(i, k)$ need to be aggregated along the k (occurrences) and h (features) dimensions. Different scores may result, depending on which dimension is aggregated first. In the experiments from this section, two combinations approaches are explored:

- an "inner" combination (IC), in which features are first combined within occurrences (along h) and then across occurrences (along k), this is

$$ps_{IC}(i) = \oplus_k \left(\oplus_h \left[f_h(i, k)\right]\right);$$

- and an "outer" combination (OC), in which features from different occurrences are grouped by feature type and then combined within groups, as follows

$$ps_{OC}(i) = \oplus_h \left(\oplus_k \left[f_h(i, k)\right]\right).$$

The experiments from this section study the impact of using $\oplus_k = \vee$ (max) for aggregating the occurrence features, used “base” features $F_0^{\vee}$, E_{rms}^{μ}, $E_{rms}^{\vee}$, etc, as the set $\mathcal{F}$, and combined features from $\mathcal{F}$ with an arithmetic mean $\oplus_h = \frac{1}{|\mathcal{F}|} \sum_h$. The reason for not considering other aggregation functions is that prominence scores derived by max-aggregates performed best in the experiments with the GH function described in Section 5.3.4.

Figures 5.6a and 5.6a show an example of how prominence scores are calculated in the OC and IC approaches. In OC, the final prominence score of term t_i, represented as the top-left grey-shaded square in the figure, is the maximum value among 3 occurrence scores, where each occurrence score is calculated as the average of the features derived from D, El, E_{rms}, and F_0 for that occurrence. In IC, the prominence score of term t_i is the average of term-level features derived from D, El, E_{rms}, and F_0, where each term-level feature is the maximum for a specific feature among the three occurrences of term t_i.

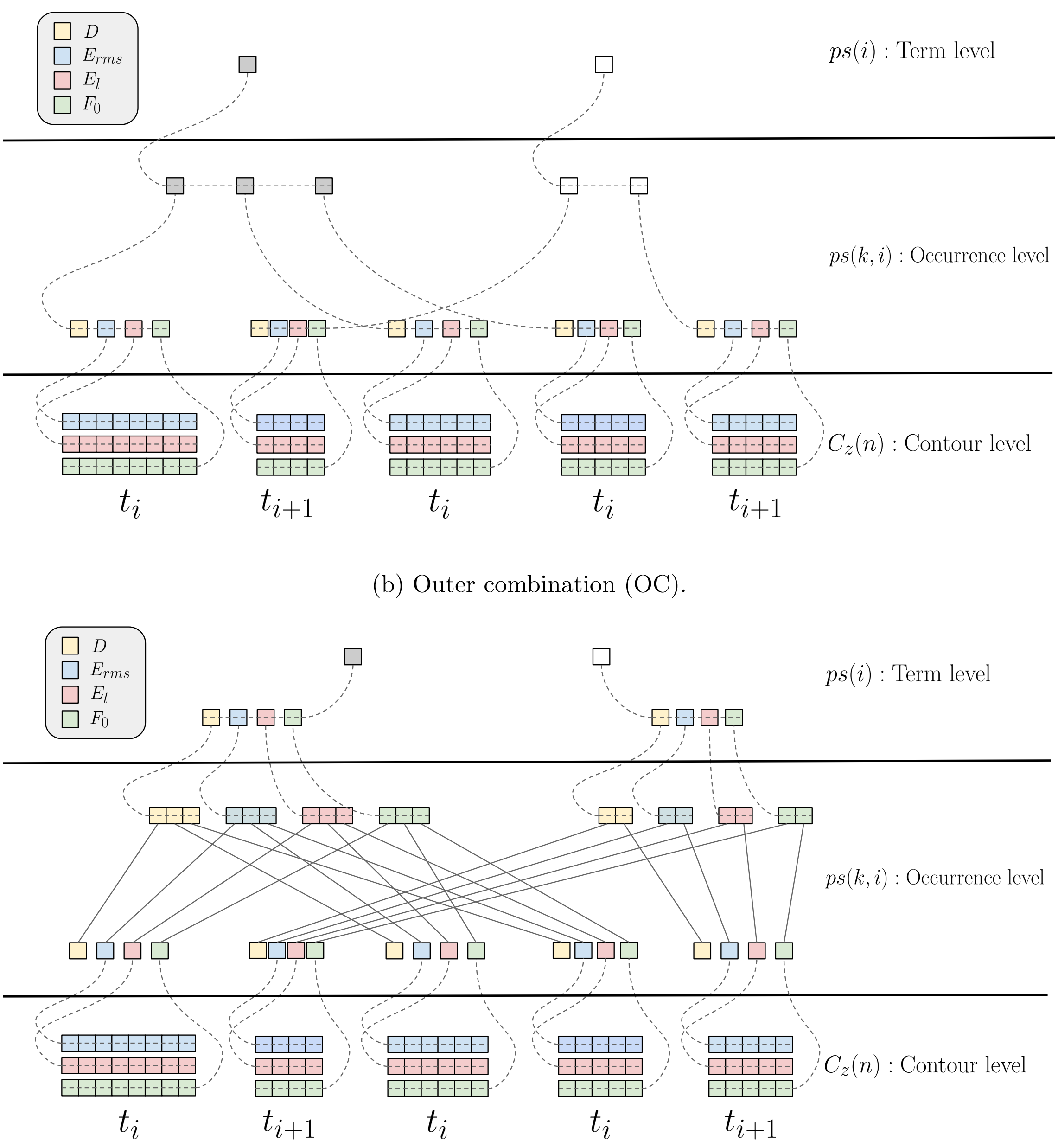
Under the set-up described above, the prominence score of a term under the IC approach is given by the features of the term's single most prominent occurrence, where the grade of prominence of an occurrence in this case is estimated as the average value of its features. By contrast, the combined prominence score of a term according to the OC approach may be determined by the feature values associated with multiple, possibly different, occurrences of the term, each of which may be deemed salient with respect to a specific feature type in isolation. For instance, if loudness and duration were to be combined with the IC approach, a term prominence score would be formed using the average of the loudness and duration values associated with the occurrence of this term that is both deemed the loudest and longest in the document. However, in the OC approach, the same term would acquire a score formed by the average of the loudness value of its loudest occurrence and the duration value of its longest occurrence.

Comparison between inner and outer combinations

Since it is not clear which combination approach would provide the most effective term weights, retrieval experiments were conducted to compare the effectiveness of the GH retrieval function (Equation 5.10) when using prominence scores defined by the IC and OC approaches. Based on the list of available features displayed in Table 5.4, there are 13 distinct occurrence features $ps(k, i)$ to be considered as possible candidates to be combined, and therefore $2^{13} = 8192$ possible ways of grouping these into different feature subsets. Since the total number of possible subsets is not prohibitively large, experiments were conducted with all possible feature subsets. A retrieval experiment consisted of using the function GH with $\delta = 0$ for retrieval, with prominence scores calculated either by IC or OC for a given feature subset. This resulted in 8192 MAP scores for each IC and OC approach.

The box plots shown in Figures 5.7 and 5.8 compare the distribution of MAP scores obtained for the IC and OC approaches across all test conditions over the SH13, SH14, SAVA, SD2, SQD, and SQD2 topic sets. In particular, Figure 5.7 shows results for the SDR task, while Figure 5.8 does so for the SPR task. In each sub-figure, the left and right

Figure 5.6: An example of how prominence scores are calculated in the inner and outer combinations approaches for two terms t_i and t_{i+1} with 3 and 2 occurrences respectively.



(a) Inner combination (IC)

(b) Outer combination (OC).

box plots show respectively the distribution of MAP scores obtained for the IC or OC approach, while the horizontal dashed line depicts the MAP score attained by GH when using prominence scores derived from the single best-performing “base” feature.

Two important observations can be made from these results. First, prominence scores produced by outer combination (OC) performed generally better than those produced by inner combination (IC). This was the case for all test conditions evaluated in the SDR task, although the differences between OC and IC were larger in the experiments with the BBC (SH13, SH14, and SAVA) collection than the SDWPS (SD2, SQD, and SQD2) collection. The plots for the SPR task show a similar trend, with OC outperforming IC in the majority of the test conditions. Overall, these results seem to suggest that multiple acoustic features may not concurrently signal the same spoken occurrence of a term as significant. Instead, the importance status of a term may be signalled in various of its occurrences across the document by means of a diverse set of acoustic features.

The second observation arises from comparing runs that used multiple features (box plots) against those that used a single feature (dashed lines). In most of the test conditions, the majority of MAP scores obtained through feature combinations are above the best single feature line in the plots, meaning that more effective term weights can be derived from multiple acoustic features.

Tables 5.15a and 5.15b presents a more detailed comparison between prominence scores derived from multiple features, via the OC approach, and those derived from a single feature, in the SDR and SPR tasks. In particular, the tables report the MAP scores of the best performing subset of features (the highest extreme points in the box plots) against those obtained with a single feature (dashed lines in the box plots). A tick symbol in a cell indicates if a particular feature was present in the best performing feature subset found for each test condition. Among the individual features considered, duration (D) and minimum F_0 ($\wedge$) were generally present in the best subset of features across the majority of the test collections. Excepting these, no other feature was frequently included in the best feature subset. A possible cause for this may be the existence highly correlated features which could result in several equally performing feature subsets.

The results from Tables 5.15a and 5.15b show that it is possible to obtain more effective term weights if multiple acoustic features are used for prominence score calculations. Furthermore, weights derived through feature combinations were more frequently effective in the SDR than in the SPR task. The latter provides supporting evidence for the observation that the acoustic features explored in this thesis tend to be more useful for retrieval purposes when aggregated from longer excerpts of spoken content containing a higher number of query term occurrences (documents) than from short excerpts (passages).

Comparison with Okapi BM25

The results from Tables 5.15 show that spoken documents can be ranked more effectively if terms are weighted based on prominence scores calculated on combinations of multiple

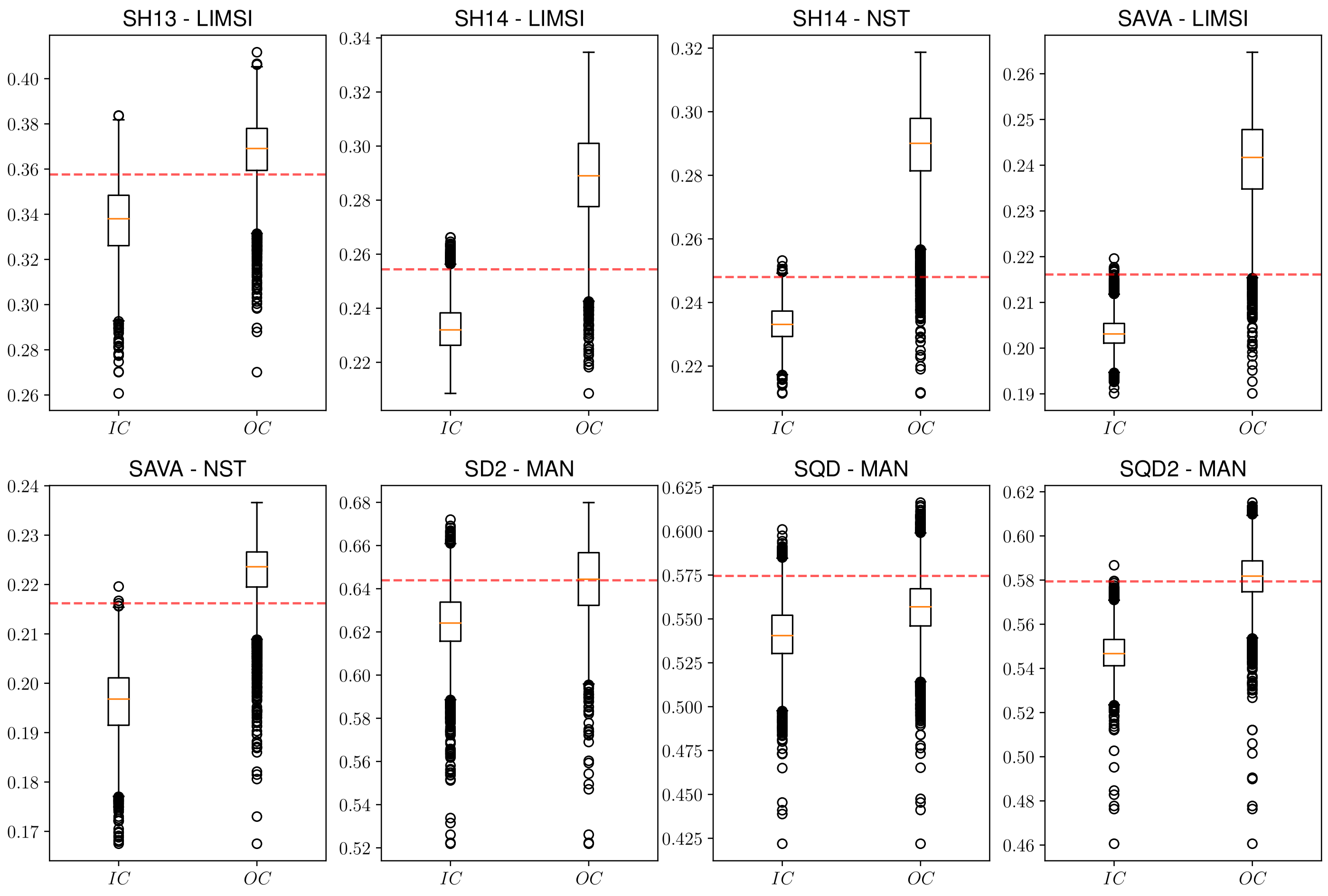


Figure 5.7: Distribution of MAP scores obtained with feature combination approaches in the SDR task. The left and right box plots in each plot show the results for the inner combination (IC) and outer combination (OC) approaches, while the dashed lines show the performance of using the single-best "base" feature.

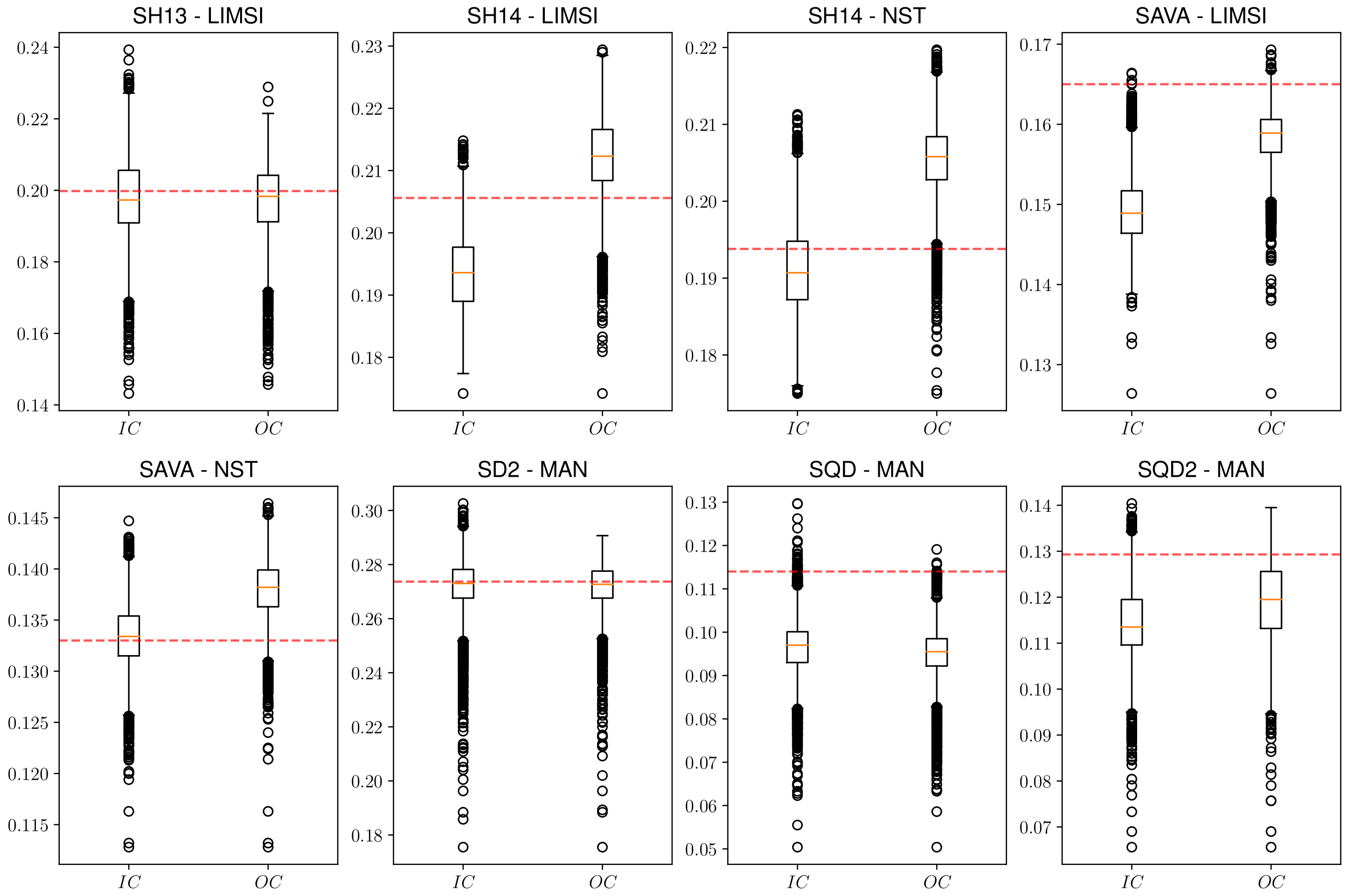


Figure 5.8: Distribution of MAP scores obtained with feature combination approaches in the SPR task. The left and right box plots in each plot show the results for the inner combination (IC) and outer combination (OC) approaches, while the dashed lines show the effectiveness of using the single-best "base" feature.

Table 5.15: Retrieval effectiveness of the GH function (with $\delta = 0$) when using the best subset of features for computing $ps_{\vee}(i)$ versus using the single-best feature.

(a) SDR task.

Topics	**Trans.**	**GH (best feature subset when $\delta = 0$)**														**GH**
		F_0				E_l				E_{rms}				D		**single**
		$\vee$	$\wedge$	μ	σ	$\vee$	$\wedge$	μ	σ	$\vee$	$\wedge$	μ	σ		MAP	MAP
SH13	LIMSI		✓		✓					✓		✓		✓	.412	.358
SH14	LIMSI		✓		✓						✓		✓	✓	**.335***	.254
SH14	NST		✓		✓				✓		✓			✓	**.319***	.248
SAVA	LIMSI		✓		✓				✓		✓	✓	✓	✓	**.265**	.216
SAVA	NST	✓	✓				✓		✓	✓			✓		**.237**	.216
SD2	MAN		✓	✓		✓					✓			✓	**.680***	.644
SQD1	MAN						✓	✓	✓		✓		✓	✓	.616	.574
SQD2	MAN		✓		✓		✓		✓	✓				✓	**.615**	.579

(b) SPR task.

Topics	**Trans.**	**GH (best feature subset when $\delta = 0$)**														**GH**
		F_0				E_l				E_{rms}				D		**single**
		$\vee$	$\wedge$	μ	σ	$\vee$	$\wedge$	μ	σ	$\vee$	$\wedge$	μ	σ		MAP	MAP
SH13	LIMSI	✓								✓				✓	.229	.200
SH14	LIMSI	✓									✓		✓	✓	**.229**	.206
SH14	NST				✓	✓					✓		✓	✓	**.220***	.194
SAVA	LIMSI		✓		✓	✓					✓				.169	.165
SAVA	NST	✓	✓		✓	✓	✓		✓	✓	✓				.146	.133
SD2	MAN			✓				✓		✓			✓	✓	.291	.274
SQD1	MAN												✓	✓	.119	.114
SQD2	MAN				✓	✓								✓	.139	.129

acoustic features. Despite this, it remains unclear whether these improved scores can complement lexical based term weights to improve the quality of the standard BM25 function. To investigate this situation, the experiments from Section 5.3.2 were repeated, but this time using the best combination of acoustic features found in each evaluation condition to derive the prominence scores in the GH retrieval function. Here, the best feature combination refers to the most effective feature subset found in the SDR and SPR tasks, and correspond to the experiments whose results are reported in Tables 5.15a and 5.15b.

Tables 5.16a and 5.16b compare the retrieval effectiveness of the Okapi BM25 and the GH retrieval functions when the best combination of features for $\delta = 0$ is used in GH to compute the prominence scores, and the best value of δ is used in each evaluation condition. The results in these tables indicate that utilising the improved prominence scores in combination with lexical BM25 scores does not provide any additional benefits to retrieval, as the quality of the rankings produced by the GH function is at most as high (not significantly different) as those produced by Okapi BM25.

The plots from Figure 5.9 provide an alternative view of the results and show how the MAP values of GH vary for increasing values of δ. By looking at the plots, it is evident that integrating prominence scores into lexical based term weights is generally detrimental for retrieval performance. Although the prominence scores derived from multiple acoustic features demonstrated increased effectiveness over single feature scores (Table 5.15), utilising these improved weights in an BM25 setting does not necessarily result in an enhanced retrieval model.

A note on parameter optimisation and over-fitting

The majority of the experiments described in Section 5.3 involved finding optimal values for various retrieval function parameters, and searching for optimal subsets of features to be combined to form the prominence scores. In this context, optimal features or parameters refer to those model configurations with which the model or retrieval function achieves maximum MAP when evaluated on a particular query set. These optimal parameter or feature configurations were tested on the exact same query sets from which they were obtained. As a consequence, there is a potential risk of having over-fitted the parameters or features selected in many of the experiments presented in this section. Despite this, the conclusions and observations drawn from these experiments are believed to hold. This claim is supported by the following arguments:

- The same parameter values for b, k_1, and k_3, which had initially been optimised with BM25 for each evaluation condition were also used in the acoustically-enhanced BM25 functions. Therefore, the acoustically enhanced functions have always been evaluated with respect to their ability to improve upon a well-tuned and possibly over-fitted BM25 function.

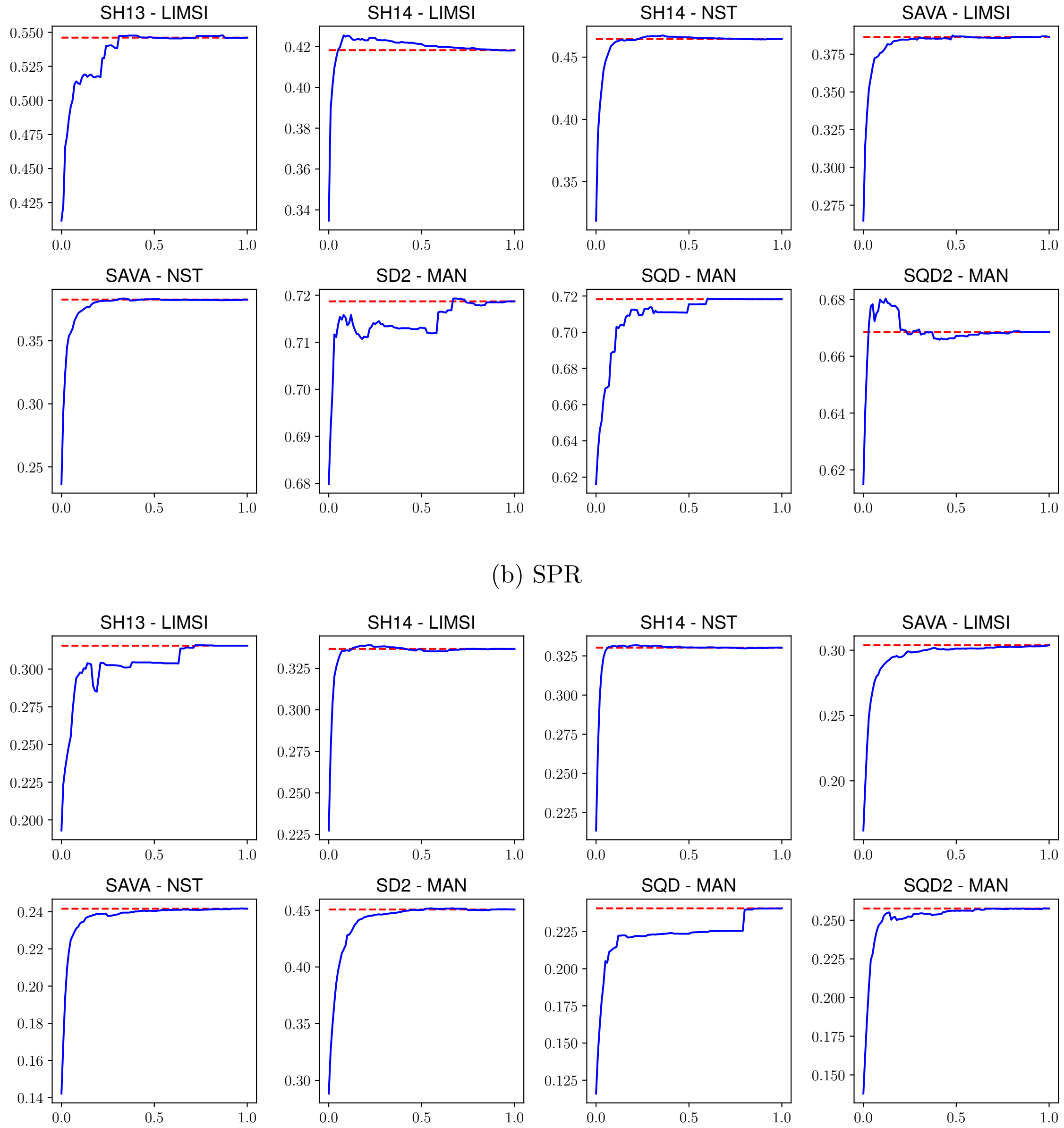
Table 5.16: Comparison between the GH function and Okapi BM25 when using the best value of δ and feature subset for computing $ps_{\vee}(i)$. In GH, the best feature subset in each test condition is the one that maximises MAP when features are combined via OC and $\delta = 0$.

(a) SDR task.

		GH (best feature subset when $\delta = 0$)															**BM25**
Topics	**Trans.**	F_0				E_l				E_{rms}				D			
		$\vee$	$\wedge$	μ	σ	$\vee$	$\wedge$	μ	σ	$\vee$	$\wedge$	μ	σ		δ	MAP	MAP
SH13	LIMSI		✓		✓					✓		✓		✓	0.87	.548	.546
SH14	LIMSI		✓		✓						✓		✓	✓	0.08	.425	.418
SH14	NST		✓		✓				✓		✓			✓	0.36	.467	.464
SAVA	LIMSI		✓		✓				✓		✓	✓	✓	✓	0.48	.387	.386
SAVA	NST	✓	✓				✓		✓	✓			✓		0.33	.383	.383
SD2	MAN		✓	✓		✓					✓			✓	0.71	.719	.719
SQD1	MAN						✓	✓	✓		✓		✓	✓	0.60	.719	.718
SQD2	MAN		✓		✓		✓		✓	✓				✓	0.12	.680	.669

(b) SPR task.

		GH (best feature subset when $\delta = 0$)															**BM25**
Topics	**Trans.**	F_0				E_l				E_{rms}				D			
		$\vee$	$\wedge$	μ	σ	$\vee$	$\wedge$	μ	σ	$\vee$	$\wedge$	μ	σ		δ	MAP	MAP
SH13	LIMSI		✓											✓	0.76	.316	.315
SH14	LIMSI	✓	✓		✓				✓					✓	0.23	.339	.337
SH14	NST		✓		✓				✓					✓	0.22	.332	.330
SAVA	LIMSI		✓		✓		✓		✓		✓		✓	✓	1.00	.304	.304
SAVA	NST	✓	✓				✓				✓		✓		0.97	.242	.242
SD2	MAN			✓									✓	✓	0.55	.452	.451
SQD1	MAN					✓							✓	✓	1.00	.241	.241
SQD2	MAN												✓	✓	0.86	.258	.258


(a) SDR

(b) SPR

Figure 5.9: Effectiveness of Okapi BM25 (red horizontal lines) and GH for increasing values of δ (blue lines) when the best combination of features is used in GH to calculate prominent scores.

- Only a very small number of extra parameters were optimised for the acoustically enhanced functions. In the GH function, this corresponded to the δ parameter set to control the contribution of prominence scores in the term weights. In the CWL function, this was the case for the k_1 and α parameters, where the latter was set to adjust the importance given to highly prominent term occurrences.

- In experiments involving search of feature subsets, results were reported for all possible feature subsets and depicted with box plots to illustrate the distribution of MAP scores obtained. Furthermore, conclusions were drawn by observing differences across such score distributions instead of the extreme values of such distributions.

- Despite having used over-fitted parameter values and best performing feature subsets, the acoustically enhanced functions did not outperform the BM25 baseline significantly.

This last observation is particularly important since it emphasises the fact that the acoustically enhanced retrieval models could not demonstrate a substantial increase in retrieval effectiveness even when using over-fitted parameters and features.

5.3.6 Summary of experiments with heuristic functions

This section described a series of retrieval experiments with two ranking functions inspired by work from Guinaudeau and Hirschberg (2011), and Chen et al. (2001), which combine a term's BM25 weight with a prominence score extracted from multiple mentions of the term in a spoken document or passage. A variety of methods to derive suitable prominence scores for terms were explored based on simple aggregations of a small set of speaker-standardised low-level descriptors of pitch, loudness, energy, and duration. Two alternative approaches were then described in detail: GH, which combines prominence and BM25 scores externally via linear interpolation; and CWL, which updates the within-document term-frequency estimates to reflect the accumulated prominence scores associated with this term's occurrences in a document or passage.

The results of retrieval experiments conducted with these methods on a diverse set of test collections, topics, and relevance assessments show that none of the proposed acoustically-enhanced functions provide consistent significant improvements in retrieval effectiveness over a standard lexical-based BM25 function. Further experimentation with the GH function provided some insight on the type of information that prosodic prominence may encode about terms, as results indicated that documents can be ranked more effectively in order of relevance to a query if the relative importance assigned to their terms is based upon prominence scores rather than on randomised or uniform weights.

Furthermore, the effectiveness of term weights based on prominence information varied depending on the retrieval task, integration method, and acoustic features used. In this respect, weights given by the estimated prominence of the most prominent occurrence of a term in a document ($ps_{\vee}(i)$) demonstrated some utility in the SDR task, while those based on other aggregation functions, like the mean implemented in GH ($ps_{\mu}(i)$) or the sum implemented in CWL ($ps_{\Sigma}(i)$), did not provide clear benefits in retrieval effectiveness over uniform weights. While in the SDR task useful weights could be derived from almost any acoustic source, this did not occur in the SPR task, where the retrieval units are shorter and contain fewer query-term occurrences across which the acoustic features can be aggregated. In this regard, acoustic features aggregated from across complete documents tend to perform better in general than those aggregated within passages, for both SDR and SPR tasks.

Additional experiments investigated the value of prominence scores defined through averages of multiple acoustic features. Two combination approaches were explored in this case. An inner combination (IC), which combined features at the occurrence-level, and an outer combination (OC), which combined features at the term-level. Experiments with these two approaches showed that OC outperformed IC in general, and that both approaches resulted in prominence scores that provide increased retrieval effectiveness compared to using a single acoustic feature. Further comparisons between GH and Okapi BM25 indicated that the former could not outperform the latter in the SDR or SPR tasks, even when using the enhanced prominence scores derived from feature combinations.

Overall, the experimental results collected suggest that words spoken with relatively extreme values of pitch, loudness, and duration do not provide additional complementary information about the topical significance of a word beyond what can be inferred based on its TF-IDF estimates. Although estimates of acoustic prominence can to some extent capture information about terms that is useful for ranking documents in order of relevance to a query, such benefits fade away when these estimates are used within a well-tuned retrieval function.

5.4 Experiments with statistical methods

The retrieval experiments presented in Section 5.3 evaluated two simple approaches that attempt to exploit aggregated acoustic descriptors of the speech prosody of words to improve the ranking of relevant spoken documents and passages. In this line of work, the approaches adopted for computing the prominence scores of terms from acoustic features, as well as for incorporating these into a ranking function, were intuitively reasonable albeit ad-hoc in nature. On one hand, the prominence scores examined were hand-designed and computed extempore by calculating aggregated statistics over low-level feature contours. On the other hand, the integration approaches explored were based on heuristics and, beyond the analysis presented in Section 5.2.4, lacked of a solid theoretical justification.

The prominence scores examined in Section 5.3 were based on aggregations of low-level contours, subsequently aggregated by term and document classes and finally averaged across different families of features into a unique score for a term-document pair. Although simple, this method based on linear combinations of multiple aggregated values presents some obvious limitations. First, much of the information conveyed by the acoustic features is inherently lost in each aggregation step if only averages, standard deviations, and extreme values are retained. Second, using a conventional arithmetic mean for combining different sources of acoustic information effectively assigns equal importance to every feature considered, and is unlikely to take full advantage of any useful inter-dependencies that may exist between features.

In order to cope with some of these limitations as well as to gain further insights about any relationship that may exist between informative terms and their acoustic realisation, this section reports statistical analysis over the acoustic data and describes experiments carried out with machine learning techniques. First, an analysis of the correlation between term-level acoustic features and BM25 weights is presented. Next, experiments with statistical classifiers are described in which binary classifiers were trained with acoustic features to distinguish between terms occurring in relevant and non-relevant spoken documents. Finally, retrieval experiments are presented using a learning-to-rank approach trained with document-aggregated acoustic features to improve upon an initial Okapi BM25 ranking.

5.4.1 Correlation and regression analysis

Recall the study conducted by Silipo and Crestani (2000), which investigated the correlation between manually assigned acoustic scores of words and their BM25 weights in the 2 hours OGI corpus of telephone conversations. In a series of histograms, the authors observed that a high proportion of words that were given high (low) average acoustic scores were likely to have high (low) BM25 scores.

The goal of this section is to extend Silipo and Crestani's analysis to the BBC and SDPWS datasets, which are substantially larger and more varied than the OGI corpus. Without loss of generality, the following analysis was only conducted with speech data and LIMSI transcripts from the BBC1 collection, and with speech data and manual transcripts from the SDPWS2 collection. Two analyses were carried out for this purpose: (i) a correlation analysis of acoustic features and BM25 scores based on Spearman's rank-order correlation coefficients; and (ii) a regression analysis in which a linear model is fitted with the acoustic descriptors of a term to predict its BM25 score. Finally, a set of histograms with similar characteristics than those reported in (Silipo and Crestani, 2000; Crestani, 2001) were plotted based on the predictions made by these regression models.

The experimental analysis described in this section has some important differences with respect to the study conducted by Silipo and Crestani:

1. The acoustic scores utilised by Silipo and Crestani (2000) were derived from manual annotations of acoustic stress produced by trained linguists, while all results reported in this thesis were obtained with the automatically extracted acoustic features described in Section 5.1.

2. While all words including stopwords were considered in the analysis of Silipo and Crestani, the analysis presented in the following section was restricted to indexing terms only, i.e. those terms included in the search indices of the BBC and SDPWS's collections. Consequently, the following analysis excludes stopwords, parts-of-speech other than verbs and nouns in the case of the SDPWS collection, and recognised words with unreliable time-stamps, as per previous descriptions in Sections 4.1.2, 4.2.2 and 5.1.2.

3. Lastly, Silipo and Crestani (2000) calculated term frequency and document frequency statistics by treating every short story in the OGI collection as a single "document". Contrary to this, the analysis reported hereafter calculates frequency statistics based on the full contents of an episode in the case of the BBC1 collection and a lecture in the case of the SDPWS2 collection, both of which are substantially longer than the stories within the OGI corpus.

Correlation analysis

Two statistics commonly used to estimate the strength of the relationship between two variables are the Pearson's product moment correlation (ρ) and the Spearman's rank-order correlation (ρ_s). Since one of the goals of this thesis is to determine whether the prosody of words can be used effectively to improve the quality of term weighting schemes in SCR systems, the interest is on understanding the impact that certain features may have if used in the generation of document rankings. In this context, the Spearman's correlation coefficient seems to be more appropriate, since it can measure monotonic order-preserving associations between the variables under study.

Spearman's correlation coefficients were calculated for features extracted for each term-document pair in the BBC1 and SDPWS2 collections. Recall that the former contains 1860 spoken documents, 34,849 terms, and 1,945,746 unique term-document pairs, while the latter has 98 documents, 6,223 terms, and 38,891 unique term-document pairs. For every term-document pair, a set of acoustic features was calculated as described in Section 5.3.2, by aggregating each occurrence associated value $E_l^{\vee}$, $E_l^{\wedge}$, E_l^{μ}, ..., across all occurrences of such term in each document. BM25, BIM, and TFO scores were subsequently computed for every term-document pair based on the Okapi BM25 ranking function. Finally, a correlation coefficient was then calculated between these scores and every acoustic feature. Besides the set of aggregates max ($\vee$), mean (μ), and sum (Σ) used in the experiments from Section 5.3 for calculating term-level prominence scores, the following analysis also considers min-aggregates ($\wedge$). Additionally, in this analysis, no subsequent range or sigmoid normalisation was applied to the occurrence features, in contrast to normalisation functions applied in the experiments with heuristic retrieval functions described in Section 5.3.

Table 5.17 shows the Spearman's correlation coefficients of term-level acoustic features and BM25 weights. Various observations can be made based on these results:

(i) Averaging features across occurrences ($ps(i) = \mu$) results in scores that are substantially less strongly correlated with IR scores compared to those based on other aggregation functions.

(ii) Taking the maximum value across occurrences ($ps(i) = \vee$) generally results in scores that are positively correlated with TFO and negatively correlated with BIM scores. Consequently, taking the minimum value ($ps(i) = \wedge$) provides scores that follow the inverse correlation directions w.r.t TFO and BIM than those obtained via max-aggregation.

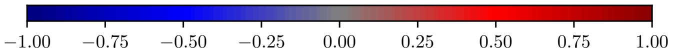
(iii) The addition of occurrence features ($ps(i) = \Sigma$) generally provides term scores that are highly correlated with IR scores. The directions of such associations depend on how contour features are aggregated at the level of individual term occurrences. The direction is positive w.r.t. BM25 and TFO, and negative w.r.t. BIM when $ps(k, i) \in \{\vee, \sigma\}$. For $ps(k, i) = \wedge$, the acoustic scores are strongly correlated with the IR scores but in the inverse directions than those obtained with $ps(k, i) = \vee$.

(iv) Duration related features, $ps(k, i) = D$, exhibit distinctive correlation patterns compared to the rest of the acoustic features. In particular, terms which occur rarely in the collection (high BIM) tend to be lengthened on average (high average duration, $ps(k, i) = \mu$).

(v) With the exception of the groups $ps(i) = \mu$ and $ps(k, i) = D$, the rest of the correlation coefficients calculated against BIM present similar magnitudes and directions across collections. Despite these similarities, the acoustic features are generally more strongly correlated with IR scores in the Japanese (SDPWS2) collection than in the English (BBC1) collection. Furthermore, since BM25 scores are calculated as the multiplication of BIM and TFO scores, the cross-collection differences that can be observed in Table 5.17 between the coefficients calculated for BM25 can be explained by those of BM25 against BIM and TFO. In particular, while $\rho_s(BM25, BIM) = 0.86$ and $\rho_s(BM25, TFO) = -0.04$ for the BBC1 data, the same coefficients for the SDPWS2 data are $\rho_s(BM25, BIM) = 0.36$ and $\rho_s(BM25, TFO) = 0.63$.

The correlation patterns observed in Table 5.17 make evident that some of the statistics used for aggregating occurrence-level features are strongly affected by the sample size across which such aggregates are calculated. By definition, the sum $ps(i) = \Sigma$ ranges across the within-document occurrences of a term. Therefore, the scores obtained via such summation are inherently correlated with the within-document frequency of the term in that document and, consequently, positively correlated with TFO and inversely correlated with BIM. The $ps(i) \in \{\vee, \wedge\}$ aggregation functions suffer from a similar bias. Recall that $ps(i) \in \{\vee, \wedge\}$ functions select, respectively, the maximum and minimum values across

Table 5.17: Spearman's ρ_s rank-order correlation coefficients of term features against BM25, BIM, and TFO scores. The coefficient values are coloured based on the reference scale



(a) BBC1

$ps(i)$	$ps(k,i)$	$C_Z(n)$	BM25	BIM	TFO
Σ	$\vee$	F_0	-0.128	-0.292	0.480
Σ	$\vee$	E_{rms}	-0.109	-0.259	0.449
Σ	$\vee$	E_l	-0.125	-0.298	0.508
Σ	$\wedge$	F_0	0.117	0.295	-0.512
Σ	$\wedge$	E_{rms}	0.119	0.307	-0.530
Σ	$\wedge$	E_l	0.134	0.337	-0.572
Σ	μ	F_0	-0.049	-0.028	-0.005
Σ	μ	E_{rms}	-0.036	-0.019	-0.003
Σ	μ	E_l	-0.032	-0.017	-0.004
Σ	σ	F_0	-0.135	-0.312	0.520
Σ	σ	E_{rms}	-0.138	-0.320	0.532
Σ	σ	E_l	-0.150	-0.342	0.561
Σ	D	-	0.119	0.008	0.228
$\vee$	$\vee$	F_0	-0.084	-0.176	0.282
$\vee$	$\vee$	E_{rms}	-0.068	-0.151	0.261
$\vee$	$\vee$	E_l	-0.063	-0.152	0.273
$\vee$	$\wedge$	F_0	-0.141	-0.211	0.238
$\vee$	$\wedge$	E_{rms}	-0.161	-0.225	0.251
$\vee$	$\wedge$	E_l	-0.170	-0.236	0.262
$\vee$	μ	F_0	-0.135	-0.207	0.267
$\vee$	μ	E_{rms}	-0.122	-0.194	0.259
$\vee$	μ	E_l	-0.123	-0.199	0.266
$\vee$	σ	F_0	-0.075	-0.155	0.251
$\vee$	σ	E_{rms}	-0.073	-0.157	0.259
$\vee$	σ	E_l	-0.085	-0.172	0.275
$\vee$	D	-	0.096	0.010	0.190
$\wedge$	$\vee$	F_0	0.133	0.228	-0.302
$\wedge$	$\vee$	E_{rms}	0.147	0.239	-0.292
$\wedge$	$\vee$	E_l	0.153	0.244	-0.296
$\wedge$	$\wedge$	F_0	0.075	0.137	-0.213
$\wedge$	$\wedge$	E_{rms}	0.045	0.139	-0.256
$\wedge$	$\wedge$	E_l	0.038	0.139	-0.273
$\wedge$	μ	F_0	0.083	0.173	-0.258
$\wedge$	μ	E_{rms}	0.096	0.187	-0.270
$\wedge$	μ	E_l	0.094	0.183	-0.266
$\wedge$	σ	F_0	0.146	0.235	-0.292
$\wedge$	σ	E_{rms}	0.147	0.232	-0.281
$\wedge$	σ	E_l	0.142	0.233	-0.294
$\wedge$	D	-	0.269	0.327	-0.263
μ	$\vee$	F_0	0.023	0.016	0.011
μ	$\vee$	E_{rms}	0.050	0.055	-0.020
μ	$\vee$	E_l	0.056	0.056	-0.014
μ	$\wedge$	F_0	-0.072	-0.089	0.068
μ	$\wedge$	E_{rms}	-0.091	-0.085	0.039
μ	$\wedge$	E_l	-0.097	-0.085	0.028
μ	μ	F_0	-0.042	-0.035	0.022
μ	μ	E_{rms}	-0.029	-0.023	0.018
μ	μ	E_l	-0.028	-0.025	0.019
μ	σ	F_0	0.054	0.064	-0.044
μ	σ	E_{rms}	0.042	0.040	-0.006
μ	σ	E_l	0.031	0.030	-0.002
μ	D	-	0.204	0.179	-0.024

(b) SDPWS2

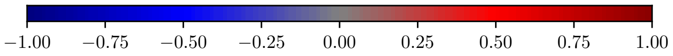
$ps(i)$	$ps(k,i)$	$C_Z(n)$	BM25	BIM	TFO
Σ	$\vee$	F_0	0.525	-0.316	0.744
Σ	$\vee$	E_{rms}	0.487	-0.223	0.636
Σ	$\vee$	E_l	0.569	-0.338	0.797
Σ	$\wedge$	F_0	-0.531	0.330	-0.758
Σ	$\wedge$	E_{rms}	-0.555	0.355	-0.797
Σ	$\wedge$	E_l	-0.587	0.384	-0.853
Σ	μ	F_0	0.068	0.039	0.049
Σ	μ	E_{rms}	0.032	0.081	-0.020
Σ	μ	E_l	0.044	0.044	0.014
Σ	σ	F_0	0.568	-0.365	0.818
Σ	σ	E_{rms}	0.580	-0.351	0.818
Σ	σ	E_l	0.593	-0.392	0.863
Σ	D	-	0.276	0.184	0.130
$\vee$	$\vee$	F_0	0.355	-0.185	0.481
$\vee$	$\vee$	E_{rms}	0.349	-0.125	0.427
$\vee$	$\vee$	E_l	0.364	-0.140	0.452
$\vee$	$\wedge$	F_0	0.126	-0.216	0.281
$\vee$	$\wedge$	E_{rms}	0.199	-0.280	0.406
$\vee$	$\wedge$	E_l	0.241	-0.310	0.463
$\vee$	μ	F_0	0.321	-0.230	0.483
$\vee$	μ	E_{rms}	0.294	-0.196	0.428
$\vee$	μ	E_l	0.296	-0.211	0.438
$\vee$	σ	F_0	0.336	-0.164	0.444
$\vee$	σ	E_{rms}	0.338	-0.114	0.407
$\vee$	σ	E_l	0.351	-0.157	0.452
$\vee$	D	-	0.357	0.065	0.282
$\wedge$	$\vee$	F_0	-0.235	0.328	-0.466
$\wedge$	$\vee$	E_{rms}	-0.195	0.341	-0.440
$\wedge$	$\vee$	E_l	-0.215	0.352	-0.468
$\wedge$	$\wedge$	F_0	-0.092	0.007	-0.087
$\wedge$	$\wedge$	E_{rms}	-0.223	0.092	-0.272
$\wedge$	$\wedge$	E_l	-0.315	0.138	-0.404
$\wedge$	μ	F_0	-0.237	0.271	-0.423
$\wedge$	μ	E_{rms}	-0.216	0.275	-0.408
$\wedge$	μ	E_l	-0.226	0.277	-0.421
$\wedge$	σ	F_0	-0.219	0.314	-0.443
$\wedge$	σ	E_{rms}	-0.191	0.348	-0.443
$\wedge$	σ	E_l	-0.225	0.347	-0.475
$\wedge$	D	-	-0.068	0.418	-0.384
μ	$\vee$	F_0	0.114	0.077	0.062
μ	$\vee$	E_{rms}	0.098	0.138	-0.004
μ	$\vee$	E_l	0.088	0.152	-0.025
μ	$\wedge$	F_0	0.031	-0.136	0.134
μ	$\wedge$	E_{rms}	0.030	-0.159	0.154
μ	$\wedge$	E_l	0.022	-0.169	0.146
μ	μ	F_0	0.051	0.030	0.035
μ	μ	E_{rms}	0.064	0.036	0.039
μ	μ	E_l	0.061	0.022	0.043
μ	σ	F_0	0.049	0.127	-0.043
μ	σ	E_{rms}	0.084	0.160	-0.035
μ	σ	E_l	0.072	0.145	-0.035
μ	D	-	0.146	0.292	-0.082

all within document occurrences of a term. These functions can be seen as two distinct sampling processes: one that prefers selecting high over low values from a sample of occurrence-level features of a term-document pair; and another one which prefers selecting low over high values from the same sample of occurrence-level features. As the chances of encountering an extreme value from a sample increase with the size of the sample, features that are max or min aggregated this way will tend to be positively or, conversely, negatively correlated with TFO scores. This may explain why most of the features aggregated via the max and min functions present the strongest correlation with TFO in the results from Table 5.17.

Contrary to the aggregation functions $ps(i) \in \{\Sigma, \vee, \wedge\}$ which are biased towards within-document frequency counts, calculating the average $ps(i) = \mu$ over occurrence-level features controls for the size of each data sample and is therefore not affected by the within-document frequency counts of the target terms. For $ps(i) = \mu$, the Spearman's ρ_s values shown in Table 5.17 indicate that the acoustic features under study are generally poorly correlated with IR scores when averaged across occurrences. This is consistent with the results obtained in the experiments with the heuristic retrieval functions from Section 5.3, in which the scores resulting from using mean aggregates ($ps(i) = \mu$) were frequently less effective for document and passage ranking than those derived from maximums and summations.

Similar to the correlation analysis just described, a similar analysis can be made of the correlation between occurrence-level features and IR scores. Because occurrence-level features are not aggregated across within document occurrences, performing this type of analysis avoids calculating correlation coefficients over features that are biased towards term frequency estimates. Table 5.18 depicts the results of such an experiment. The values in the table indicate that features extracted from the SDPWS2 collection are weakly correlated with IR scores, whereas those extracted from the BBC1 collection show practically no correlation with IR features. An exception to the latter finding is the duration feature (D), which presents mild associations with BM25 and BIM scores in both collections. In particular, the coefficients calculated for SDPWS2 show that the max ($\vee$), mean (μ) and standard deviation (σ) of the contour features F_0, E_{rms}, and E_l are positively correlated with BIM scores, while the min ($\wedge$) of such features is negatively correlated. Therefore, occurrences with high values for these acoustic features tend to be associated with terms with low document frequency, i.e. with terms that occur rarely in the collection. This observation is consistent with the hypothesis about predictability and prominence of words which states that unpredictable words are more likely to be accented, that is, to be spoken with more extreme acoustic values than words more commonly mentioned in the discourse.

Table 5.18: Spearman's ρ_s rank-order correlation coefficients of occurrence features against BM25, BIM, and TFO scores. The coefficient values are coloured based on the reference scale



(a) BBC1

$ps(k,i)$	$C_Z(n)$	BM25	BIM	TFO
$\vee$	F_0	0.024	0.025	-0.009
$\vee$	E_{rms}	0.039	0.032	0.004
$\vee$	E_l	0.046	0.037	0.007
$\wedge$	F_0	-0.048	-0.039	0.000
$\wedge$	E_{rms}	-0.075	-0.052	-0.021
$\wedge$	E_l	-0.078	-0.057	-0.015
μ	F_0	-0.048	-0.034	-0.002
μ	E_{rms}	-0.031	-0.021	-0.004
μ	E_l	-0.029	-0.021	-0.002
σ	F_0	0.049	0.043	-0.003
σ	E_{rms}	0.041	0.028	0.017
σ	E_l	0.026	0.019	0.010
D	-	0.211	0.166	0.018

(b) SDPWS2

$ps(k,i)$	$C_Z(n)$	BM25	BIM	TFO
$\vee$	F_0	0.101	0.253	-0.140
$\vee$	E_{rms}	0.127	0.323	-0.183
$\vee$	E_l	0.117	0.307	-0.179
$\wedge$	F_0	-0.054	-0.102	0.033
$\wedge$	E_{rms}	-0.078	-0.163	0.062
$\wedge$	E_l	-0.079	-0.197	0.095
μ	F_0	0.055	0.163	-0.107
μ	E_{rms}	0.065	0.199	-0.133
μ	E_l	0.045	0.139	-0.100
σ	F_0	0.099	0.239	-0.120
σ	E_{rms}	0.142	0.350	-0.188
σ	E_l	0.122	0.313	-0.174
D	-	0.176	0.477	-0.269

Regression analysis

Linear regression analyses were conducted to study the extent to which variations in IR scores can be explained by linear combinations of acoustic features. In the first analysis, a linear regression model was fitted with the term-level features to predict each of the BM25, BIM, and TFO scores. A similar analysis was then carried out considering only the occurrence-level features as the variables to be used by the regression model.

In regression analysis, the multi-correlation between the combination of dependent variables and the independent variable is commonly estimated via the coefficient of determination (R^2), calculated as the square of the Pearson's product moment correlation between the model's predictions and the true values of the independent variable. In order to facilitate the comparison between the correlation coefficients reported in Tables 5.17 and 5.18, the Table 5.19 reports, instead of R^2, Spearman's ρ_s coefficients between the IR scores predicted by the linear regression models and their true values.

As expected, the figures from Table 5.19 are consistent with the observations made from the correlation coefficients for individual features, namely: (i) that features are more highly correlated with IR scores when aggregated across occurrences, possibly due to the bias that is introduced in the term-level aggregation process; and (ii) that features extracted from the BBC1 collection tend to be less correlated with IR scores than those extracted from the SDPWS2 collection. With regard to (i), the correlation coefficients from models trained with unbiased features ($ps(i) = \mu$ in Table 5.17, Tables 5.18, 5.19c and 5.19d) still suggest that there may be a meaningful association between the IR score of a term and a linear combination of its acoustic features. With respect to (ii), there may be multiple reasons why the coefficients show weaker correlations in the BBC1 collection. Since only

Table 5.19: Spearman's ρ_s between IR scores predicted by linear regression models trained with acoustic features against the true BM25, BIM, or TFO scores.

(a) BBC1, term-level features

$ps(i)$	BM25	BIM	TFO
Σ	0.273	0.379	0.573
$\vee$	0.269	0.349	0.440
$\wedge$	0.279	0.400	0.495
μ	0.214	0.190	0.059
All	0.317	0.444	0.572

(b) SDPWS2, term-level features

$ps(i)$	BM25	BIM	TFO
Σ	0.589	0.454	0.850
$\vee$	0.505	0.386	0.693
$\wedge$	0.444	0.479	0.711
μ	0.140	0.299	0.112
All	0.609	0.513	0.871

(c) BBC1, occurrence-level features

$C_z(n)$	BM25	BIM	TFO
F_0	0.090	0.073	0.007
E_l	0.149	0.113	0.021
E_{rms}	0.164	0.129	0.024
D	0.211	0.166	0.018
All	0.221	0.175	0.036

(d) SDPWS2, occurrence-level features

$ps(i,k)$	BM25	BIM	TFO
F_0	0.115	0.295	0.156
E_l	0.152	0.396	0.219
E_{rms}	0.160	0.396	0.212
D	0.176	0.477	0.269
All	0.188	0.485	0.262

automatic transcripts were available for the BBC1 documents, the errors introduced by the ASR may have disrupted the grouping of occurrences by term IDs and thus added noise to the correlation estimates. Besides ASR errors, the high diversity of the spoken material from the BBC1 collection, much of which includes multi-party conversations, background music, and speech recorded outdoors, may have introduced extra noise in the estimation of the acoustic features.

Correlation histograms

Silipo and Crestani (2000) presented a series of histograms as evidence of the observed correlation between BM25 and human-annotated stress scores in the OGI corpus. Since these manual annotations were grouped into three discrete categories for stressed words (low, medium, and high), the histograms in (Silipo and Crestani, 2000) showed how BM25 scores were distributed in these three classes. This section attempts to reproduce these histograms based on the acoustic data from the BBC1 and SDPWS2 collections.

Because no manual annotations of syllable stress are available for the BBC1 and SPDPWS2 data, the prediction scores given by a linear regression model trained with automatically extracted acoustic features were used as a substitute for the human-generated stress scores used in Silipo and Crestani's study. For this purpose, the documents from a spoken collection were first split into a training set and a test set, with 60% of the documents assigned for training and 40% for testing. A linear regression model was then fitted with the occurrence-level acoustic features from the training set to predict their associated BM25 scores. This model was then applied to the test set to generate a predicted BM25 score for each term occurrence in the test documents.

Figures 5.10 and 5.11 depict histograms that reflect the distribution of true BM25

Figure 5.10: Distribution of true BM25 scores in the BBC1 collection for term occurrences with:

(a) predicted $BM25 < 3.28$ (b) predicted $BM25 > 3.28$

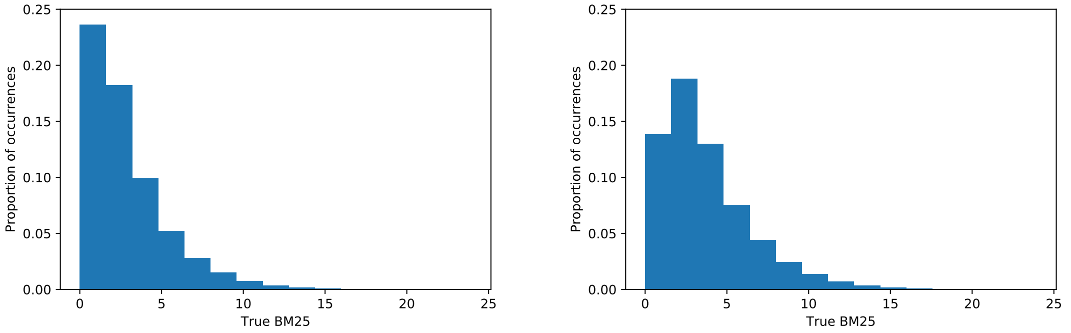
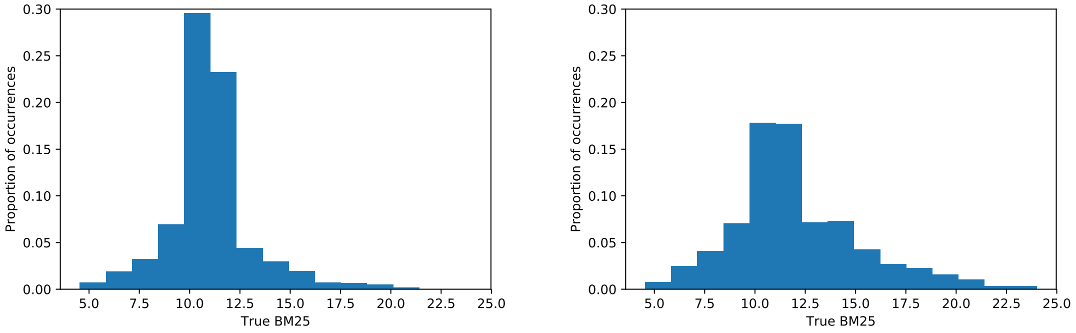


Figure 5.11: Distribution of true BM25 scores in the SDPWS2 collection for term occurrences with:

(a) predicted $BM25 < 11.62$ (b) predicted $BM25 > 11.62$



scores in the test documents for the BBC1 and SDPWS2 collections respectively. Two histograms are depicted in each figure. The proportions shown in the left histograms (a) were calculated by only considering term occurrences whose predicted BM25 scores were less than the mean of all predicted BM25 scores, in an attempt to approximate the class of "low" stressed words. Similarly, the right histograms (b) were calculated over occurrences with predicted BM25 scores greater than the mean. If there was a strong correlation between predicted and true BM25 scores, as was observed in Silipo and Crestani's study, the left and right histograms would show a steep decreasing, respectively increasing, sequence of bars. But, since the grade of correlation between true and predicted BM25 scores is low in the BBC1 and SDPWS2 collections, the proportions of term occurrences with high true BM25 scores does not change significantly between the left and right histograms.

5.4.2 Acoustic-based classification of significant terms

This section describes experiments that seek to determine whether a meaningful difference exists in the acoustic realisation of words when spoken in “relevant” versus “non-relevant” contexts. For this investigation, a statistical classifier was trained to distinguish between query terms appearing in relevant and non-relevant documents. The hypothesis is that this classifier will be able to learn the relationship, if any, between prominence and importance of significant terms, or otherwise evince that such relationship does not hold in reality.

Generation of datasets for classification experiments

A dataset consisting of examples of terms occurring in relevant and non-relevant documents was generated for a given set of queries, documents, and relevance assessments. For this purpose, every occurrence in the collection of each term appearing in a query was labelled as belonging either to the relevant, non-relevant or unknown class.

The labelling process was performed as follows. Consider $C = \{d_1, \ldots, d_N\}$ as the collection of documents, t_i the i-th term of C and t_{kij} its k-th occurrence in d_j. Suppose also that Q is the set of queries $q_1, \ldots, q_L$ and $R_1, \ldots, R_L$ their respective sets of relevance assessments such that R_l contains every document that is known to be relevant to q_l. An occurrence t_{kij} of a term t_i is labelled as relevant if and only if there is a query $q_l \in Q$ such that $t_i \in q_l$ and $d_j \in R_l$. Every other query term occurrence t_{kij} not labelled as relevant is then deemed: non-relevant, if a set $\overline{R}_l$ of documents known to be non-relevant to q_l is available and there is $q_l \in \overline{R}_l$ so that $t_i \in q_l$ and $d_j \in \overline{R}_l$; or unknown, if the relevance status of the document that contains the query term occurrence is not available in the relevance assessments.

Note that under this labelling scheme, all occurrences of query terms appearing in the same relevant document are labelled as relevant. Additionally, note that a term may be present in multiple queries and that the relevance assessments associated with such queries may be inconsistent for a particular document. For instance, an occurrence t_{kij} may be deemed relevant w.r.t. q_l and non-relevant or unknown w.r.t. some other query q_{l+1}. When inconsistencies were encountered, the relevant class was always given preference over the unknown or non-relevant class and the affected training instances labelled as relevant.

Based on the labelling procedure described above, a dataset of training pairs $\{(\vec{x}_{kij}, y_{kij})\}$ was generated for a test collection, where $\vec{x}_{kij}$ is a feature vector for the query term occurrence t_{kij} and y_{kij} is its associated relevance label. The vectors $\vec{x}_{kij}$ were populated with the occurrence-level acoustic features. These included duration D, and the contour features F_0, E_{rms}, and E_l aggregated via $\vee$, $\wedge$, μ, and σ, for a total of 13 features per vector. Table 5.20 presents statistics about the datasets generated based on the SDR data from the BBC1, BBC2, and SDPWS2 collections.

In addition, it is informative to study the predictive power of term-level features, resulting from aggregations of occurrence features. For this purpose, a dataset of term-

Table 5.20: Statistics of datasets generated for occurrence classification experiments with SDR data from the BBC1, BBC2, and SDPWS2 collections. “Rel.” stands for “Relevant”

Collection	**Query set**	**Queries**	**Training instances**				
			Total	**Rel.**	**Non-rel.**	**Unk.**	**Inconsist.**
BBC1	SH13	50	156,848	1,581	0	154,985	282
BBC2	SH14	28	106,950	12,255	11,953	81,014	1,728
	SAVA	30	153,660	10,458	6,810	134,183	2,209
SDPWS2	SD2	110	19,290	6,056	0	10,967	2,267
	SQD1	35	19,344	3,636	0	13,740	1,968
	SQD2	80	42,104	9,756	0	25,086	7,262

Table 5.21: Statistics of datasets generated for term classification experiments with SDR data from the BBC1, BBC2, and SDPWS2 collections.

Collection	**Query set**	**Queries**	**Training instances**				
			Total	**Rel.**	**Non-rel.**	**Unk.**	**Inconsist.**
BBC1	SH13	50	58,655	169	0	58,463	23
BBC2	SH14	28	42,445	1,398	2,657	38,170	220
	SAVA	30	59,789	1,109	1,821	56,739	120
SDPWS2	SD2	110	3,500	680	0	2,590	230
	SQD1	35	4,588	463	0	3,897	228
	SQD2	80	9,976	1,199	0	7,925	852

document training pairs $\{(\vec{x}_{ij}, y_{ij})\}$ was constructed for each test collection where $\vec{x}_{ij}$ is a feature vector populated with term-level features for term i in document d_j, and y_{ij} is a relevance label associated to this term-document pair. Table 5.21 shows statistics about the datasets generated for term-level features.

Experiments and results

The predictive power of acoustic features were investigated in two classification tasks. An occurrence classification task, which consisted of classifying an occurrence-document pair t_{kij} of a query term given input $\vec{x}_{kij}$ into its relevance class for that document y_{kij}. And a term classification task, which consisted of classifying a term-document pair t_{ij} given input $\vec{x}_{ij}$ into its relevant class y_{ij}. The datasets listed in Tables 5.20 and 5.21 were used for the occurrence and term classification tasks respectively.

For each dataset, a 10 folds cross-validation experiment was carried out. Each cross-validation experiment required the training instances to be randomly shuffled in a dataset, grouped by query ID, and split into 10 equal-sized folds, ensuring that all instances associated with the same query ID were kept in the same fold. Subsequently, a classifier was trained with 9 folds and its performance measured on the remaining fold. This process was repeated 10 times for every possible combination of training and testing folds, and the resulting performance scores averaged.

Logistic regression classifiers were used in all classification experiments presented in this section. In particular, the scikit-learn toolkit (Pedregosa et al., 2011) was used to train and evaluate the logistic models. To cope with the high class imbalance that exists in the datasets, different penalisation weights were set for the minority (relevant) and majority (non-relevant or unknown) classes. Specifically, the cost weight of each class was set to $\frac{M}{2C}$, with M being the total number of instances and C the number of instances of that class in the training set. For similar reasons, the generalisation power of the classifiers was measured with balanced accuracy (BAC) (Brodersen et al., 2010) instead of conventional accuracy. Note that trivial classifiers, such as those that output the most common label from the dataset, obtain a BAC score of 50%.

The statistical significance of the classification results was determined via permutation tests. Each test consisted of training a model to predict a random permutation of the original assignment of class labels. Statistical significance was then determined by calculating the proportion of accuracy scores that were larger than the score obtained by training the model on the original labels. The results presented in this section were validated with 1000 random permutations of label assignments.

To better understand the predictive power of the acoustic features in each classification task, it is also useful to consider the accuracy of logistic models trained with scores produced by BIM, TFO, and BM25 as input features. Because the scores produced by the BIM, TFO, and BM25 functions are known to be useful for ranking relevant documents more highly than non-relevant ones, they should in principle be useful features for the classification of term-document pairs into a relevant and non-relevant class, and can thus provide a point of reference for measuring the usefulness of the acoustic features. While the BIM, TFO, and BM25 functions calculate a score for a term-document pair instead of a score for each term occurrence in a document, the scores produced by these functions can be directly extrapolated to individual term occurrences and hence used to populate the input vectors $\vec{x}_{kij}$.

Tables 5.22 and 5.23 present the results of the cross-validation experiments for the occurrence and term classification tasks respectively. Tables 5.22a and 5.23a show results for the task of classifying between relevant and unknown instances only, where non-relevant instances were considered as “unknown” when available, while Tables 5.22b and 5.23b show results for the task of classifying relevant from non-relevant instances. The first three columns in each table show accuracy scores obtained by training models with BIM, TFO, and BM25 features in isolation, while the IR column shows the accuracy obtained when training models with all BM25-based features. Columns 6 to 9 show results obtained for features derived from F_0, E_{rms}, E_l, and D respectively, while the “PROS” column shows the results for models trained with all acoustic features. Finally, the last column in the tables depict the accuracy obtained with models trained with the IR and PROS features in combination.

As expected, the results from Tables 5.22 demonstrate that IR features are substantially more effective than acoustic features at identifying occurrences of query terms in relevant documents. More importantly, the figures also indicate that models trained with the complete set of acoustic features (PROS) are significantly better than chance (50%),

Table 5.22: Cross-validation BAC (%) of logistic regression models trained with occurrence-level acoustic features, BIM, TFO, or BM25 scores, for classifying among:

(a) relevant, and unknown or non-relevant classes;

Dataset	BIM	TFO	BM25	IR	F_0	E_{rms}	E_l	D	PROS	IR+PROS
SH13	**62.69***	**73.32***	**69.07***	**77.10***	50.64	52.28	**54.10***	47.44	**52.73**	**76.71***
SH14	**64.58***	**75.32***	**71.84***	**77.75***	46.67	**51.92***	**49.94***	**52.66***	**50.23***	**77.79***
SAVA	**62.74***	**71.66***	**71.71***	**77.72***	**49.16**	**51.33***	**52.69***	49.04	**51.88***	**77.49***
SD2	**61.86***	**62.92***	**62.63***	**65.79***	**52.46***	**51.46***	49.33	53.14	**53.70***	**64.75***
SQD1	**61.18***	**63.74***	**63.69***	**64.88***	**55.40***	**53.72***	**53.50***	**55.75***	**54.42***	**64.82***
SQD2	**52.98***	**68.27***	**62.24***	**69.25***	**52.47***	**55.07***	**55.16***	**54.21***	**55.21***	**68.51***
Average	61.00	69.21	66.86	72.08	51.13	52.63	52.45	52.04	53.03	71.68

(b) relevant and non-relevant classes.

Dataset	BIM	TFO	BM25	IR	F_0	E_{rms}	E_l	D	PROS	IR+PROS
SH14	**59.37***	**61.52***	**61.69***	**59.85***	47.64	48.46	47.49	46.79	47.52	**59.56***
SAVA	**49.91***	**64.91***	**60.44***	**67.42***	46.58	50.20	49.74	**50.01**	48.24	**67.75***
Average	54.64	63.22	61.06	63.63	47.11	49.33	48.62	48.40	47.88	63.65

albeit not providing any additional benefits on top of IR features (IR+PROS). Among the group of IR features (columns 2-5), results show similar trends to those observed in earlier experiments, specifically, that weights produced by BIM are less effective than those produced by TFO, and that a combination of these two (IR) can produce models that are more accurate at detecting relevant occurrences. Within the group of acoustic features (columns 6-10), the large majority resulted in models that performed significantly better than chance (50%). Models trained with all acoustic features in combination achieved the highest accuracy values on average.

Tables 5.22b and 5.23b report accuracy scores for models trained with non-relevant targets instead of unknown instances. These experiments could only be performed for the SH14 and SAVA datasets, since these are the only collections where examples of non-relevant documents are available. As can be observed from the results, the accuracy of the logistic models decreases significantly when trained to differentiate between relevant and non-relevant instances. The reason for this can be traced back to the way documents were pooled by the Search&Hyperlinking task organisers during the generation of the relevant assessments for the SH14 and SAVA queries. Since only the top-ranked documents ranked by a group of IR systems were included in the pools to be assessed, only the highest scoring documents for a query were assessed as “relevant” or “non-relevant” by a human judge. Because all these documents contain a similarly high number of terms matching the query, the task of distinguishing between relevant and non-relevant documents in this case becomes more difficult. Despite this increase in task difficulty, the results from Table 5.22b indicate that BM25 derived features can still perform better than chance. Yet, models trained with acoustic features could not perform better than a trivial classifier, meaning that occurrences of terms spoken in these two categories of high scoring documents are indistinguishable based on the acoustic features considered.

Table 5.23: Cross-validation BAC (%) of logistic regression models trained with term-level acoustic features, BIM, TFO, or BM25 scores, for classifying among:

(a) relevant, and unknown or non-relevant classes;

Dataset	BIM	TFO	BM25	IR	F_0	E_{rms}	E_l	D	PROS	IR+PROS
SH13	**65.58***	**72.45***	**73.76***	**80.90***	**73.03***	**72.56***	**71.23***	**68.88***	**70.44***	**78.41***
SH14	**60.98***	**71.57***	**68.24***	**74.43***	**69.39***	**70.71***	**71.00***	**68.71***	**71.64***	**75.62***
SAVA	**64.84***	**67.98***	**70.73***	**74.60***	**64.94***	**66.17***	**65.95***	**64.71***	**66.31***	**74.75***
SD2	**61.64***	**61.67***	**66.33***	**67.19***	**62.35***	**60.46***	**60.39***	**60.38***	**61.61***	**66.37***
SQD1	53.06	**63.40***	**61.35***	**64.55***	**63.23***	**61.70***	**62.83***	**63.97***	**61.21***	**61.17***
SQD2	52.12	**62.02***	**60.35***	**63.04***	**61.70***	**62.38***	**62.13***	**60.61***	**61.68***	**62.26***
Average	59.70	66.51	66.79	70.79	65.77	65.66	65.59	64.55	65.48	69.76

(b) relevant and non-relevant classes.

Dataset	BIM	TFO	BM25	IR	F_0	E_{rms}	E_l	D	PROS	IR+PROS
SH14	**55.61***	**61.96***	**60.34***	**63.33***	**59.68***	**59.29***	**59.58***	**58.24***	**59.38***	**62.39***
SAVA	**53.12**	**62.87***	**60.50***	**63.72***	**59.72***	**60.77***	**60.42***	**59.16***	**58.56***	**62.19***
Average	54.37	62.42	60.42	63.53	59.70	60.03	60.00	58.70	58.97	62.29

The accuracy values for the term classification tasks shown in Table 5.23 indicate once more that the predictive power of the acoustic features increases when they are aggregated across term occurrences. This is consistent with the observations made based on the analysis from Section 5.4.1, where term-level features presented strong correlations with respect to term frequency estimates. A similar effect is observed in the results of the classification experiments from Table 5.23. In the latter case, models trained with acoustic features achieved similar BAC scores than models trained with TFO scores. Despite these high correlations, the term-level acoustic features do not give additional improvements in BAC over models trained with BM25 derived features.

On the utilisation of occurrence-level predictions in BM25

In an additional study, we attempted to adapt a BM25 based ranking function to produce term weights that are sensitive to the predictions made by a classifier trained with acoustic features[4]. Similarly to the classification experiments described in the previous section, we trained a binary classifier with inputs $\vec{x}_{kij}$ for each query term occurrence t_{kji} to predict relevant and non-relevant target classes.

For this study, we used an expanded set of 294 acoustic features per occurrence as the input vector $\vec{x}_{kij}$, instead of the 13 features used in the experiments described in the previous section. Our expanded feature-set included the features proposed by Rosenberg (2012) for the task of pitch accent detection, as well as those proposed by Mishra et al. (2012) for the task of word prominence detection. Specifically, this feature set was composed of: aggregations (max, mean, standard deviation and Z-score of maximum) of raw and speaker-normalised F_0, $\log F_0$, and intensity contours; aggregations of their delta,

[4] This study was originally reported in (Racca and Jones, 2015b)

spectral tilt and spectral band contours; voicing ratio, centre of gravity, and area under F_0 and intensity contours as well as the location and amplitude of peak and valleys. All these features were extracted for a target occurrence as well as for a window of 8 context words around the target, by using the AuToBI toolkit v1.5.1 (Rosenberg, 2010).

Instead of logistic regression models, the work from (Racca and Jones, 2015b) used radial basis support vector machines (SVMs) (Cortes and Vapnik, 1995), with a combination of grid search and cross validation to find suitable values for the C and γ parameters. The trained models were then used to predict labels for all term occurrences in the collection, including occurrences of terms that did not appear in any of the queries for training. These predictions were next used in a modified BM25 function to boost the weight of terms whose occurrences in a document were predicted as relevant by the model. The approach adopted to incorporate the model's predictions into BM25 is similar to the CWL approach, previously described in Section 5.2.3, in which the raw term frequency of the term to be scored is replaced by a summation of occurrence-level scores. The alternative summation we used is shown in Equation 5.19,

$$F_{\Sigma}(i,j) = \sum_{k} \alpha^{\hat{y}_{kij}} \tag{5.19}$$

where $\hat{y}_{kij} = f(\vec{x}_{kij}) \in \{-1, 1\}$ is the classifier's prediction for occurrence t_{kij} and α is some positive constant. Essentially, the function accumulates an amount equal to α for every occurrence predicted as relevant and, conversely, an amount equal to α^{-1} for every occurrence predicted as non-relevant.

The effectiveness of a BM25 function that uses $F_{\Sigma}(i,j)$ instead of the raw counts of term frequency was evaluated in the SPR task. A cross-training experiment was carried out for this purpose, with the SDPWS test collection and the SD2 and SQD1 query sets used for generating either training or testing data. In addition, we conducted experiments with manual as well as with ASR transcripts. The final results of these experiments, included in Appendix E, did not draw any clear conclusions as to whether the proposed approach was effective in improving over a standard BM25 baseline in the SPR task. In fact, the retrieval effectiveness of the modified BM25 function was only found statistically significantly better than Okapi BM25, based on a paired t-test, when SD2 queries were used to train the SVM model and Okapi BM25 was set to sub optimal values for the b and k_1 parameters. When better values for b and k_1 were used in BM25, no significant differences were observed between the baseline and the acoustically enhanced version of BM25.

It is important to note that the majority of features included in our extended feature set are based on additional transformations (functionals) applied to the basic F_0 and intensity contours. Even though this extended feature set may provide increased performance in pitch accent detection, classification, and prominence detection tasks, as demonstrated by Mishra et al. (2012); Rosenberg (2012), they do not perform significantly better than the subset of features considered in the other classification experiments described in this thesis.

A possible important limitation of the approach proposed in this study is that the classification task posed to the statistical classifiers, that is, classifying each individual term occurrences into relevant or non-relevant classes, is ill-defined. One reason for this is that many terms in a query are not likely to be equally important at signalling relevant from non-relevant content, yet the labelling procedure adopted for mapping document assessments of relevance onto occurrence-level classes makes no distinction among terms in the queries. A second reason is that the occurrence-level relevance classes may not be entirely valid at the level of occurrences, nor reflective of the underlying ranking task to be solved which requires a classification to be made at the level of documents instead. Since the relevance status of a document is not likely to be determined by the presence or absence of a single term from the query, training a classifier to predict the relevance status of a document based entirely on this single occurrence is not likely to be valuable for the task of ranking documents (or passages) in order of relevance to the query.

In the next section, experiments are described with a learning-to-rank approach which seeks to solve the document ranking problem directly. Instead of training the model to optimise an ill-defined learning objective at the level of term occurrences, the statistical model is trained to optimise the quality of a ranking of documents based on acoustic features extracted from the query terms matching such documents.

5.4.3 Learning-to-rank with acoustic features

In the experiments with feature combinations described in Section 5.3.5, the prominence score for a term-document pair was formed by averaging the available features associated with occurrences of this term in the document. These combined scores were then incorporated into variants of the Okapi BM25 function as described in Section 5.2, and then used for scoring spoken documents or passages for given a query.

Learning-to-rank approaches present a potentially more effective alternative for combining and incorporating new information into existing ranking functions. Because learning-to-rank approaches rely on supervised learning techniques to "learn" a ranking function from examples of query-document pairs and relevance assessments, they can facilitate the integration and combination of non-standard features in the construction of new retrieval models. A commonly cited example in this context is that of PageRank (Page et al., 1999), which can easily be incorporated via learning-to-rank approaches, yet has been shown difficult to be integrated into theoretical IR frameworks effectively (Craswell et al., 2005). Besides facilitating the inclusion of new features, the underlying learning algorithms used in learning-to-rank approaches are capable of exploiting inter-feature relationships or even discovering new feature transformations that are useful for the underlying ranking task.

The experiments described in this section seek to determine if a state-of-the-art learning-to-rank approach is capable of exploiting the document-level acoustic information of terms

Table 5.24: Feature template used in learning-to-rank experiments. The total number of features extracted per query-document pair is 208: 64 for each F_0, E_{rms}, and E_l, plus 16 based on D.

Document	**Term**	**Occurrence**	**Contour**
d	$ps(i)$	$ps(k, i)$	$C_z(n)$
$\Sigma, \vee, \wedge, \mu$	$\Sigma, \vee, \wedge, \mu$	$\vee, \wedge, \mu, \sigma, D$	F_0, E_{rms},E_l

to improve the quality of a given ranking of spoken documents. Rather than learning a ranking function from scratch, the learning-to-rank model is trained to improve upon an initial ranking produced by a well-tuned Okapi BM25 function. If the speech prosody of words is truly useful for SDR, then a learning-to-rank model may be able to exploit this information to produce a re-ranking of documents that surpasses the quality of the initial BM25 ranking.

LambdaMART

Among the learning-to-rank approaches proposed in the literature, LambdaMART (Burges, 2010; Burges et al., 2011) was chosen for this set of experiments. Because LambdaMART models can be trained to iteratively improve upon a given baseline ranking function, they provide a useful tool to assess whether a ranking produced by a standard lexical-based BM25 function can be improved by considering additional prosodic/acoustic features. In addition, this learning-to-rank approach was the winner of the Yahoo! Learning to Rank Challenge (Chapelle and Chang, 2011) and has been demonstrated to perform in-par with other state-of-the-art approaches in different tasks and collections (Tax et al., 2015). Appendix C provides a detailed description of LambdaMART models.

Experimental set-up

For training a LambdaMART model with data from a specific test collection, a set of training pairs $\{(\vec{x}_j, y_j)\}_{j=1}^{M}$ needs to be produced for every query. For each query q_l in a test collection, let D_l be the set of documents that have one or more terms from that query. A training instance $(\vec{x}_j, y_j)$ was created for each $d_j \in D_l$, where $\vec{x}_j$ is a feature-vector with values derived from acoustic features of the query-terms present in d_j, and y_j is the relevance score of d_j, equal to 1 if d_j is relevant to q_l, as stated in the relevance assessments, and 0 otherwise.

The features that comprise the document vector $\vec{x}_j$ were generated by aggregating term-level acoustic features at the document-level, as illustrated in Figure 5.3. Table 5.24 summarises the different features extracted for every query-document pair in a test collection. The total number of acoustic features generated per query-document pair is 208. As was done in the classification experiments described in Section 5.4.2, results are reported for LambdaMART models trained with each of the individual feature groups F_0, E_{rms}, E_l, or D, as well as with the complete set of acoustic features (PROS).

Table 5.25: Statistics of datasets generated for learning-to-rank experiments with SDR data from the BBC1, BBC2, and SDPWS2 collections.

Collection	**Query set**	**Queries**	**Training instances**	
			Total	**Relevant**
BBC1	SH13	50	44,073	50
BBC2	SH14	28	37,323	778
	SAVA	30	44,303	492
SDPWS2	SD2	110	4,686	440
	SQD1	35	2,784	114
	SQD2	80	6,953	300

Table 5.25 presents statistics of the learning-to-rank datasets generated based on the LIMSI and MAN transcripts of the BBC1, BBC2, and SDPWS2 collections. The generated datasets can be arranged into two groups of three: a BBC group (English) with the query set splits SH13, SH14, and SAVA; and a SDPWS2 group (Japanese) with the query set splits SD2, SQD, and SQD2. Learning-to-rank experiments were carried out by cross-validating LambdaMART models across the splits in a group, where each split served as either training, validation, or test data. Validation data was mainly used to avoid overfitting by stopping the training algorithm if no improvements were seen in the validation queries after 50 consecutive iterations of LambdaMART. All models were optimised on and evaluated using mean average precision (MAP). An open source Python implementation of LambdaMART, *RankPy*[5], was used to train all models. In addition to early-stopping, various hyper parameters in LambdaMART can be adjusted to help overcome overfitting. These were set the the values recommended in the RankPy package, as follows:

- Shrinkage: 0.1
- Maximum number of leaf nodes: 5
- Minimum number of instances per leaf: 50
- Minimum number of instances required to split a node: 2

In addition to the input vectors $\vec{x}_j$, a LambdaMART model can be provided with a base ranker, $F_0(\vec{x}_j)$, used to generate initial rankings of documents for every query. When a base ranker is provided, LambdaMART tries to improve upon this baseline by augmenting the ensemble with new trees trained on the residuals of the trees from the ensemble. This set up matches well with the overall objective of the present study, which seeks to ascertain whether acoustic features can be used to improve a well-tuned BM25 ranking function. For this reason, LambdaMART models were initialised with the rankings produced by the variations of Okapi BM25 considered in the previous experiments, specifically, the BIM, TFO, and the full version of BM25. Note that, as shown in Table 5.3a, the best-performing values of the BM25 parameters b and k_1 differ slightly across the query sets in

[5] `https://bitbucket.org/tunystom/rankpy`

the BBC and SDPWS collections. In this respect, a realistic scenario is adopted in which only training queries are assumed to be available at training time, and consequently the baseline BM25 scores were produced with the b and k_1 values that performed best in each respective training set.

Experimental results

Tables 5.26, 5.27, and 5.28 summarise the results obtained by LambdaMART models for every data split. In particular, the three tables depict respectively results obtained for models using the BIM, TFO, and full BM25 functions as base rankers, trained with different feature subsets ($+F_0$, $+E_{rms}$, $+E_l$, and $+D$) and with the complete set of acoustic features (+PROS). The columns BIM, TFO, and BM25 show the effectiveness achieved by such retrieval functions on the test queries. Bold values and * symbols mark statistically significant ($p < 0.05$) and highly significant ($p < 0.01$) differences respectively between LambdaMART and the BIM, TFO, and BM25 baselines. Because the b and k_1 parameters used in each case were tuned on the training queries, the MAP scores of the baselines TFO and BM25 are lower than those reported in Section 5.3 for these ranking models.

The results from Table 5.26 show that LambdaMART can effectively improve upon BIM rankings when trained with the acoustic features considered. This is consistent with the results reported earlier for the experiments with the GH function and re-vindicates the hypothesis that the acoustic features under study can effectively capture within-document term frequency information when aggregated across occurrences. The latter observation is not surprising considering that most of the term-level acoustic features are biased towards term frequency estimates, as demonstrated in the experiments described in Sections 5.4.1 and 5.4.2.

For models that used TFO as base-learner (Table 5.27), the acoustic features did not provide consistent benefits in retrieval effectiveness. Similar observations can be made about models that used BM25 to produce the initial rankings of documents (Table 5.28), with small but nevertheless significant improvements over the baseline in a few of the test conditions. Overall, the average effectiveness of LambdaMART models trained with acoustic features tends to be slightly greater than the BM25 baseline, especially, in experiments with the SDPWS2 collection. Despite this, significant improvements were the exception rather than the rule and were sometimes inconsistent across data splits and feature groups.

Some of the acoustic features under study, in particular those aggregated via a max ($\vee$), min ($\wedge$), or summation (Σ) across term occurrences, are biased towards within-document term frequency estimates. Because these types of features will always be strongly correlated with term frequency estimates, irrespective of the feature's original values, there remains the question of whether the observed improvements over the BIM, TFO, and BM25 baselines are truly due to the acoustic information of terms or to some random effect caused by using an extended set of aggregated features that are well correlated with term frequency estimates. If is true that speakers tend to "highlight" significant words by using extreme acoustic values when they speak, then they will tend to do so consistently. In other words, their assignment of acoustic values to words will not be arbitrary.

In order to determine whether the observed improvements in MAP were a consequence of the particular way that acoustic values were assigned by speakers to term occurrences in the BBC and SDPWS2 collections, or if these were merely due to the bias introduced by the feature aggregation process, the learning-to-rank experiments were repeated with features derived from random permutations of the original acoustic-occurrence assignments. For this purpose, the feature vectors associated with each term occurrence in a collection were permuted randomly so that, at the end of this process, each term occurrence was randomly assigned to the feature vector of another term occurrence in the corpus. The resulting randomised features were then aggregated across occurrences and subsequently across terms and documents to form the datasets used to train, validate, and test the LambdaMART models. This experiment was repeated with 100 distinct random permutations of the vectors in a training set. For each of these, a LambdaMART model was trained and tested on a test set. The resulting 100 MAP scores for each test set were finally used to estimate a p-value.

Results with gray coloured cells in Tables 5.26, 5.27, and 5.28, highlight cases in which the MAP score obtained with the original acoustic features can be deemed significant ($p < 0.05$), based on the fact that such scores are higher than 95% of all MAP scores obtained with random permutations. Among all MAP scores of LambdaMART models that were deemed significantly higher than BIM (Table 5.26) based on a t-test, only a small fraction of these were also deemed meaningful according to a permutation test. This suggests that much of the improvement obtained over the BIM was probably due to the fact that aggregated features are correlated with term frequencies, rather than to the actual information that is encoded in the acoustic features. Thus, the way features were transformed provided a noisy approximation of term frequency information, which is known to be useful in combination with IDF information as that modelled by the BIM. Despite this, there were still cases in which models trained with acoustic features provided significant improvements over both models trained with random permutations and the BIM, TFO, and BM25 baselines. This suggests that acoustic features may, under special circumstances, provide complementary information to term distribution statistics that is useful for ranking relevant spoken documents. However, these meaningful improvements were only observed for few of the test conditions considered, and therefore they do not generalise across test collections.

5.4.4 Summary of experiments with statistical methods

Section 5.4 described additional experiments that sought to gain a better understanding of the relationship between the prosodic realisation of terms, their lexical-based weights as measured by variations of the BM25 function, and the relevance status of the content in

Table 5.26: Retrieval effectiveness of LambdaMART models on test queries when BIM is used as base ranker.

(a) BBC

Train	**Dev**	**Test**	BIM	+F_0	+E_{rms}	+E_l	+D	+PROS
SH14	SAVA	SH13	0.236	**0.362***	**0.398***	**0.377***	**0.394***	**0.428***
SAVA	SH14	SH13	0.239	**0.389***	**0.408***	**0.403***	**0.344**	**0.398***
SH13	SAVA	SH14	0.197	**0.338***	**0.362***	**0.356***	**0.308***	**0.363***
SAVA	SH13	SH14	0.209	**0.381***	**0.369***	**0.369***	**0.332***	**0.370***
SH13	SH14	SAVA	0.194	**0.301***	**0.300***	**0.334***	**0.254***	**0.316***
SH14	SH13	SAVA	0.184	**0.329***	**0.308***	**0.337***	**0.322***	**0.323***
Average			0.210	0.350	0.357	0.363	0.326	0.366

(b) SDPWS2

Train	**Dev**	**Test**	BIM	+F_0	+E_{rms}	+E_l	+D	+PROS
SQD1	SQD2	SD2	0.615	**0.697***	**0.731***	**0.738***	**0.694***	**0.705***
SQD2	SQD1	SD2	0.637	**0.720***	**0.735***	**0.735***	**0.702***	**0.727***
SD2	SQD2	SQD1	0.511	0.573	0.607	0.563	0.577	0.595
SQD2	SD2	SQD1	0.511	0.589	0.590	**0.609**	**0.591**	0.594
SD2	SQD1	SQD2	0.474	**0.559***	**0.565***	**0.551***	**0.560***	**0.549***
SQD1	SD2	SQD2	0.473	**0.592***	**0.625***	**0.637***	**0.597***	**0.584***
Average			0.537	0.622	0.642	0.639	0.620	0.626

Table 5.27: Retrieval effectiveness of LambdaMART models on test queries when TFO is used as base ranker.

(a) BBC

Train	**Dev**	**Test**	TFO	+F_0	+E_{rms}	+E_l	+D	+PROS
SH14	SAVA	SH13	0.437	0.457	0.455	0.451	0.458	0.461
SAVA	SH14	SH13	0.459	0.482	0.459	0.459	0.475	0.469
SH13	SAVA	SH14	0.428	0.429	**0.438***	0.433	**0.437***	0.434
SAVA	SH13	SH14	0.409	**0.392**	**0.382**	0.405	0.401	0.391
SH13	SH14	SAVA	0.358	0.362	**0.363**	**0.363**	0.359	**0.363**
SH14	SH13	SAVA	0.360	0.367	0.361	0.367	0.363	0.367
Average			0.409	0.415	0.410	0.413	0.415	0.414

(b) SDPWS2

Train	**Dev**	**Test**	TFO	+F_0	+E_{rms}	+E_l	+D	+PROS
SQD1	SQD2	SD2	0.654	0.667	**0.675***	**0.683**	**0.675**	**0.693***
SQD2	SQD1	SD2	0.653	**0.671**	0.665	0.667	**0.671**	**0.671**
SD2	SQD2	SQD1	0.662	0.673	0.658	0.665	0.670	0.687
SQD2	SD2	SQD1	0.667	0.674	0.650	0.650	0.669	0.668
SD2	SQD1	SQD2	0.653	0.654	0.655	0.667	0.664	0.650
SQD1	SD2	SQD2	0.670	0.678	0.675	0.680	0.677	0.670
Average			0.660	0.670	0.663	0.669	0.671	0.673

Table 5.28: Retrieval effectiveness of LambdaMART models on test queries when Okapi BM25 is used as base ranker.

(a) BBC

Train	**Dev**	**Test**	BM25	$+F_0$	$+E_{rms}$	$+E_l$	$+D$	+PROS
SH14	SAVA	SH13	0.476	0.486	0.441	0.483	0.476	0.455
SAVA	SH14	SH13	0.479	0.482	0.474	0.489	0.454	0.469
SH13	SAVA	SH14	0.407	0.413	0.410	**0.416***	0.413	0.404
SAVA	SH13	SH14	0.416	0.394	0.403	0.421	0.413	0.418
SH13	SH14	SAVA	0.371	0.366	**0.375**	0.375	0.373	**0.375**
SH14	SH13	SAVA	0.378	0.385	0.385	0.384	0.384	0.383
Average			0.421	0.421	0.415	0.428	0.419	0.417

(b) SDPWS2

Train	**Dev**	**Test**	BM25	$+F_0$	$+E_{rms}$	$+E_l$	$+D$	+PROS
SQD1	SQD2	SD2	0.713	0.725	**0.732**	0.730	**0.735**	0.730
SQD2	SQD1	SD2	0.697	**0.719**	0.712	0.718	**0.718**	**0.715**
SD2	SQD2	SQD1	0.706	0.685	0.694	0.708	0.691	0.702
SQD2	SD2	SQD1	0.698	0.691	0.685	0.704	0.669	0.691
SD2	SQD1	SQD2	0.643	0.645	0.646	0.645	0.655	0.645
SQD1	SD2	SQD2	0.664	**0.672**	0.661	0.674	**0.686**	0.679
Average			0.687	0.689	0.688	0.696	0.692	0.694

which such terms are spoken. Such relationships were indirectly studied through a series of data analysis and machine learning experiments.

First, the correlation between acoustic features of terms and their BM25 scores was investigated. This work extends that of Silipo and Crestani (2000) with the speech data from the BBC and SDPWS collections. The correlation analysis performed over term-level features indicated that when these are aggregated via max ($\vee$), min ($\wedge$), and summation (Σ), the resulting scores tend to be strongly correlated with within-document term-frequencies, irrespective of the feature values that term occurrences acquire. By contrast, when term-level features are averaged across occurrences, they are weakly correlated with BM25, BIM and TFO weights. The latter trend was also observed when estimating the correlation of occurrence-level features against BM25 weights, which was significantly weaker than that observed for term-level features. Considering multiple acoustic features and combining these linearly generally produces scores that are more strongly correlated with BM25, BIM and TFO weights, as demonstrated by the linear regression experiments.

Second, logistic regression classifiers were used to determine whether speakers assign special acoustics to words spoken in contexts for which such word is topically relevant. In this study, a word was considered topically relevant if it appears in a document which is deemed relevant to a query containing the word. The results of this experiment demonstrated that a linear classifier could, to a minor extent, identify differences between words pronounced in relevant and non-relevant contexts when trained with acoustic features in isolation. However, term weights produced by variations of the BM25 function were significantly more effective at distinguishing relevant from non-relevant occurrences than acoustic features. Using BM25 weights in combination with acoustically-derived weights to train the logistic models did not result in increased classification performance over that achieved with the BM25 weights alone. Furthermore, the fact that models trained with TFO weights were similarly accurate to models trained with term-level acoustic features, but less accurate than models trained with occurrence-level acoustic features, supports the observation that term-level features are indeed affected by the aggregation procedure and biased towards term frequency estimates.

Third, LambdaMART models were trained with features at the level of documents to ascertain if such models are capable of exploiting the acoustic information of words in a SDR task more effectively. The models were trained to improve upon a given initial ranking of documents, produced by the BIM, TFO or BM25 retrieval functions. LambdaMART models were shown to provide increased retrieval effectiveness over a ranking of documents produced by BIM, but they did not show clear consistent improvements over rankings produced by TFO and BM25 across the majority of the test conditions.

5.5 Summary

This chapter investigated the value of prosodic information as a complement to lexical information to produce enhanced term weights for SCR. For this purpose, a set of speaker-normalised acoustic correlates of pitch, loudness and duration were extracted for each index term from the BBC and SDPWS collections.

A diverse set of experiments were then carried out with heuristic retrieval functions that sought to incorporate these acoustic features into the calculation of term weights to determine the importance that individual terms matching a query should be given in the retrieval process. Results from these experiments demonstrated that the acoustic information of individual words could provide benefits in retrieval effectiveness on top of a lexical-based retrieval function only when the term weights produced by the latter are of low quality, uniform, or otherwise poorly estimated.

The relationship between lexical and acoustic term weights was next studied through the analysis of the correlation of acoustic features and lexical weights. This analysis indicated that scores derived from the set of acoustic features are weakly correlated with lexical weights produced by the BM25 retrieval function. Further experiments were then conducted to investigate the relationship between terms considered topically relevant and their prosodic information, by proxy of relevance assessments of documents and queries available in the BBC and SDPWS collections. These experiments showed that while there was some value in using acoustic information for identifying topically relevant terms, the acoustic features did not provide any benefits over, nor complemented effectively with lexical features. Additional experiments with a learning-to-rank approach also indicated that acoustic features may be of value only for improving over low-quality rankings produced by sub-optimal lexical-based functions.

Overall, the experimental work described in this chapter suggests that to a minor extent the speech prosody of words can capture useful information about the significance of words pronounced in a spoken document. However, compared to evidence that can be derived from word distribution statistics, the evidence that prosodic information can provide about the importance of words is not sufficiently strong to be useful for SCR. Besides providing a rather weak and noisy signal about word importance, the prosodic correlates considered in this investigation do not seem to provide complementary information to word distribution statistics that could be used to improve the ranking of spoken documents or passages.

Chapter 6

Robust SCR through Passage Contextualisation

When documents are long and multi-topical, and in applications that seek to minimise audio playback time, spoken passage retrieval is normally preferred over full spoken document retrieval. Despite the recent progress that has been made in the quality of speech recognition systems, ASR errors still pose a big challenge for SCR applications, especially, in cases where the information units to be retrieved are short in length. Since short retrieval passages contain fewer term repetitions, recognition errors may have a much larger impact on the ability of a system to retrieve these elements effectively.

In passage and XML retrieval, contextualisation techniques (Kekäläinen et al., 2009; Arvola et al., 2011; Carmel et al., 2013) seek to improve the rank of a relevant element by considering information from its surrounding elements and its container document. Recent research has demonstrated that some of these techniques are also particularly effective in SCR applications (Nanjo et al., 2014; Shiang et al., 2014). However, no previous research has explicitly studied their potential to provide robustness to speech recognition errors.

This chapter evaluates existing contextualisation techniques, including a recently proposed technique based on positional language models (PLM) (Lv and Zhai, 2009) on the task of retrieving relevant spoken passages in response to a spoken query. The benefits of these techniques are studied when queries and documents are transcribed with increasingly higher error rates, in order to simulate increasingly difficult retrieval conditions.

This chapter is structured as follows. Section 6.1 provides an extended introduction and motivates the use of contextualisation techniques for SCR. Section 6.2 presents the various contextualisation techniques considered in this investigation. Experiments with these techniques are next presented in Section 6.3. Lastly, Section 6.4 summarises research findings.

6.1 Motivation

Although the quality of ASR systems has improved significantly over the past few years, ASR errors still pose a challenge to traditional text retrieval techniques. This occurs in domains where speech is informal, conversational, or spontaneous (Larson and Jones, 2012b), or when the elements to be retrieved by the SCR system are short in length or lack sufficient contextual information and verbosity to be retrieved effectively (Allan, 2001). Context and verbosity are desirable properties of a retrievable element because they can increase its chances of matching one or more query terms, even when many of its terms are misrecognised by the ASR system. In general, the more repetitions of important terms used to convey the topic and the more exhaustively this topic is covered by the terms in the element to be retrieved, the more robust will be its matching process against ASR errors.

The domain and level of spontaneity of the speech content may well affect the diversity of the vocabulary used to convey information as well as the amount of word repetition (verbosity). These characteristics of speech may in turn make the task of finding relevant information more or less difficult for an SCR system. For instance, in broadcast news, presenters frequently read written reports whose content and word-usage has been carefully selected to facilitate the understanding of the material while maximising communication effectiveness. By contrast, in less formal speech, topics tend to be conveyed somewhat more vaguely, by using a more limited vocabulary, making use of fewer content-bearing words and/or synonyms.

Related to the increase in difficulty in retrieving elements with poor or non-descriptive vocabulary, there is also the problem of structuring long multi-topic documents into suitable retrieval elements. Existing content structuring approaches based on structural cues or text segmentation techniques produce a static fixed set of segments which may not always align well with topic boundaries or with the elements that will best satisfy individual user information needs. In addition, the length of the segments produced by such segmentation approaches has a potential effect on the robustness of an SCR system to ASR errors. While an SCR system may still be able to retrieve a long element containing a few number of misrecognised words from the query at top-ranks, such system will arguably have more difficulties in retrieving a shorter version of this element at similar ranks, as in the latter case the impact of ASR errors will be greater relative to the length of the element.

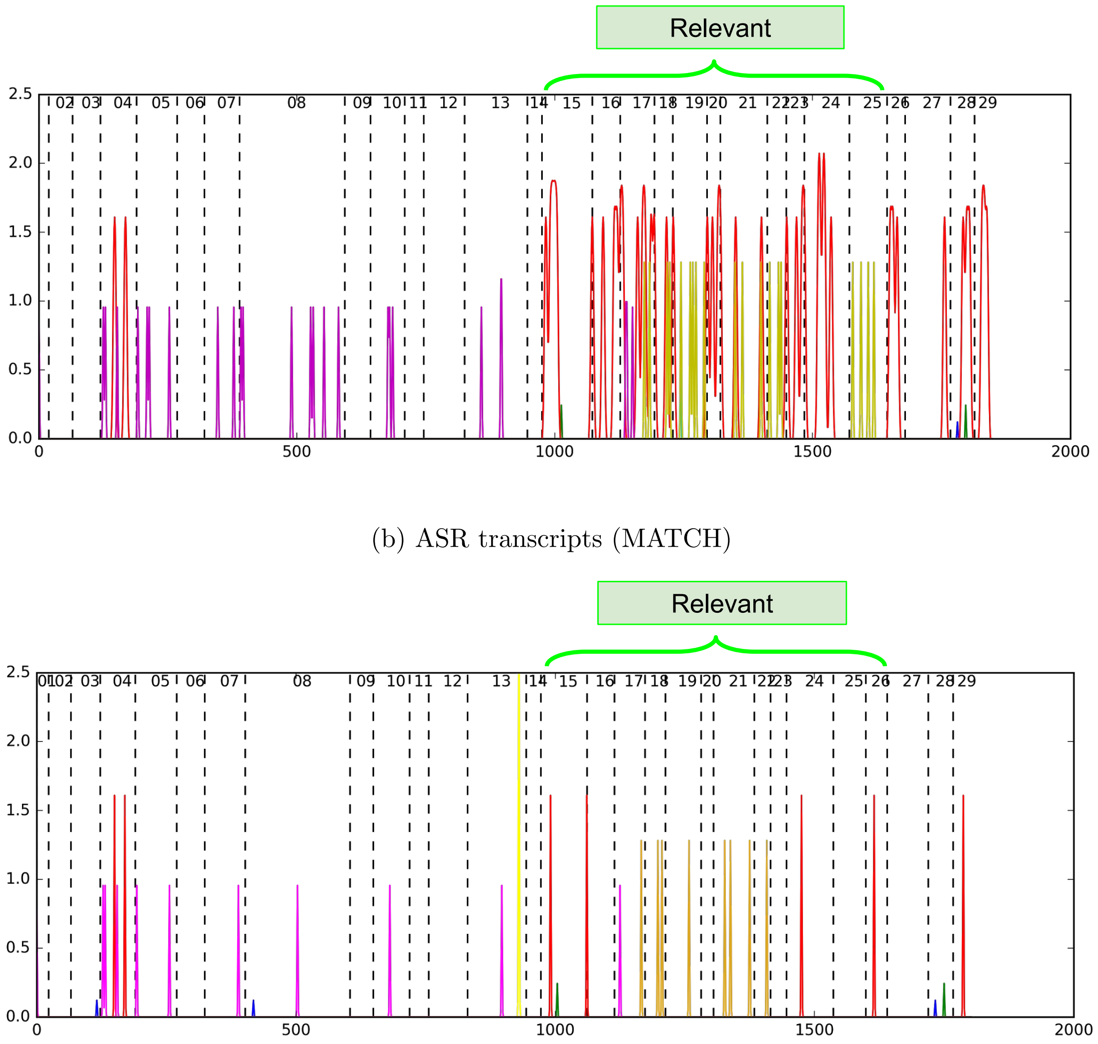
The extent to which an SCR system is robust to ASR errors is thus likely to depend on the length, verbosity, vocabulary diversity, and boundary quality of the elements that are considered as retrievable units by the SCR system. To see why all these factors are important, consider the spoken document example from Figure 6.1. The plots from the figure depict the locations of terms from a query appearing in the manual and ASR transcripts of a spoken lecture from the SDPWS2 collection. Every coloured vertical line in the plots indicates positions where a particular query term occurs within the document, with each colour representing the occurrences of a different query term and height proportional to their inverse document frequencies. Vertical dashed lines indicate the boundaries of retrieval elements, defined at positions where slide transitions were made in the lecture (see Section 4.2.4), numbered from 01 to 29 in this document, with elements 15-26 being relevant to the query.

The plots from Figure 6.1 depict clearly the effects that deletion and substitution errors can produce on a speech transcript and the potential impact these may have on retrieval effectiveness. The example also shows a case of sub-optimal segmentation, where the relevant section has been fragmented into several smaller retrieval elements. As the example shows, ASR errors can substantially reduce the number of query terms appearing in the transcript, particularly within regions that are relevant to the query. Thus, regions that would otherwise contain a high number of query term occurrences and be assigned high relevance scores by the SCR system can be "diluted" or "weakened" by the effects of recognition errors and thus be assigned less prominent relevance scores instead. Intuitively, these negative effects are expected to worsen if the amount of ASR errors increase in the transcripts.

As the example exposes, an SCR system that uses a text retrieval method to rank spoken elements based on the amount of term overlap between each element and the query, is more likely to suffer from the impact of ASR errors if considering each retrieval element independently in the scoring and ranking processes. However, because the content corresponding to the entire relevant section in this case was conveyed by using a high number and diverse range of terms related to the query, an SCR system may still be able to return all relevant elements from this document at high-ranks if considering their surrounding context (neighbouring elements) when computing each element's relevance score. Thus, while ASR errors can cause query terms to disappear from the transcript and "dilute" regions with a high density of query terms, considering term occurrences from neighbouring elements in the relevance scoring process may help in recovering the original density information of the relevant elements.

Besides being potentially useful against ASR errors and inaccurate segmentation of the spoken material, retrieval techniques that consider the context of its retrieval elements may also be able to capture explicit dependencies among related elements. Traditional IR models assume that the relevance of a document is independent of the relevance of other documents from the collection. Although this assumption may seem reasonable in document retrieval applications, it certainly seems less justifiable in the case of passage retrieval where many of the elements to be ranked may in fact belong to a single document. Elements that belong to the same document are more likely to be about similar topics and, therefore, more likely to condition the probability of relevance of other elements that also occur in that document. In lectures or academic presentations, for example, it is normal for a presenter to provide an introduction at the beginning of the talk which, even though

Figure 6.1: Locations where query terms appear in a manual and ASR transcript for a spoken query and document from the SDPWS2 collection. Vertical coloured lines mark the occurrences of a specific query term with height given by each term's inverse document frequency. Occurrences from the same term share the same colour. Also, dashed vertical lines mark the boundaries of individual retrieval elements (slide-group segments) in the document transcripts.



(a) Manual transcripts (MAN)

(b) ASR transcripts (MATCH)

it may occur some minutes before the full presentation of a particular topic, may still be of importance for this topic and possibly contain some useful terms which may not be mentioned later in the presentation. In such circumstances, it seems logical to consider a longer informational unit around the target element, in the hope to facilitate its retrieval at top ranks.

The context of an element, that is, the information that is present in its container document, has been shown to be valuable for improving element-retrieval effectiveness (Kekäläinen et al., 2009; Arvola et al., 2011). The process of taking context into account when computing the relevance score of an element is known as contextualisation (Kekäläinen et al., 2009). Various contextualisation techniques have been proven effective not only in text retrieval tasks such as XML retrieval (Arvola et al., 2011) and passage retrieval (Carmel et al., 2013; Keikha et al., 2014), but also in SCR (Nanjo et al., 2014; Shiang et al., 2014) tasks. Despite this, no previous studies have investigated the extent to which contextualisation techniques can provide increased robustness to ASR errors in the context of passage retrieval, when retrieval elements are pre-defined short excerpts of potentially errorful transcripts.

The remainder of this chapter studies the impact of incorporating context into the task of retrieving short spoken passages given a spoken query from a collection of long spoken documents. The research question under study is RQ-2, stated in Section 1.2 as whether contextualisation techniques can provide increased robustness to ASR errors when relevance scores are calculated via text retrieval methods.

6.2 Contextualisation techniques

Contextualisation techniques for element-retrieval seek to rank an element based on its content and the contents of its neighbouring elements in a document. In these techniques, elements are scored depending not only on the query terms occurring within the element itself but also on those occurring in other positions within the document. Two simple and widely adopted contextualisation approaches consist of interpolating the scores of an element with those of its document to consider global context (Nanjo et al., 2014), or with those from a fixed number of surrounding elements to consider local context (Shiang et al., 2014). In contrast, techniques based on positional models (PM) (Carmel et al., 2013; Keikha et al., 2014) allow consideration of longer-spans of context ignoring element boundaries. This section describes these contextualisation techniques in more detail.

6.2.1 Document score interpolation (DSI)

A simple approach to contextualising a retrieval element with information from its document is to combine the element's relevance score, calculated by a ranking function, with the score of its source document (Bartell et al., 1994; Fox and Shaw, 1993; Callan, 1994; Belkin et al., 1995; Huang et al., 2004; Abdul-Jaleel et al., 2004; Nanjo et al., 2014).

Firstly, elements and documents are scored independently by a ranking function to form two separate ranked lists of results. Secondly, the elements retrieved initially are re-ranked according to the relevance scores of their documents. By making element score computation sensitive to document scores, low-scoring elements may acquire increased relevance scores if contained within a high-scoring document. In the remainder of this thesis, this method is referred to as document score interpolation (DSI).

As mentioned in Section 2.4, methods for the effective combination of relevance scores produced by different ranking functions or "experts" have been investigated in previous research (Bartell et al., 1994; Fox and Shaw, 1993; Belkin et al., 1995). Among the methods that have been proposed for score combination, in the experiments of this thesis a simple weighted linear combination of scores or CombSUM (Fox and Shaw, 1993; Belkin et al., 1995) is adopted for combining the retrieval scores of documents and elements or passages. This combination approach has been shown to perform well in text-retrieval tasks as well as in image retrieval tasks (Chatzichristofis and Arampatzis, 2010).

Based on the CombSUM method, the relevance score of a passage p within document d for a query q is given by Equation 6.1.

$$S_{DSI}(q,p) = \lambda \, S_{BM25}(q,d) \; + \; (1-\lambda) \, S_{BM25}(q,p) \tag{6.1}$$

where the interpolation parameter λ adjusts the influence of the document score over the combined score. Intuitively, the λ parameter controls the amount of contribution considered from the passage's context in the final relevance score of the passage. Note that in this contextualisation technique, all passages in document d will receive the same equal contribution from d. If d obtains a high score for the query, then the score of its passages will be dominated by the score of d. In Equation 6.1, the document scores $S_{BM25}(q,d)$ are calculated based on frequency statistics estimated from the collection of documents, whereas the passage scores $S_{BM25}(q,p)$ are based on statistics estimated from the collection of passages only. As it is normally recommended in the application of this technique, in the experiments presented in this thesis with the DSI method, document and passage scores are range-normalised between 0 and 1 before being combined via Equation 6.1.

6.2.2 Positional models (PMs)

Positional models (PMs) seek to improve IR effectiveness by exploiting information about the positions where query terms occur in a document. A representative example of these type of models are positional language models (PLMs) (Lv and Zhai, 2009) which were introduced in IR as a mechanism to integrate evidence from term proximity features and passages into the language modelling framework for IR (Ponte and Croft, 1998). A PLM estimates the probability $P(i|c,d)$ that term i is generated at position c in document d. Thus, for every document d in the collection, and for every position c within every document, a PLM estimates a probability distribution over all terms centred at position

c in d.

In a PLM, the estimation of $P(i|c, d)$ is based on the so-called pseudo-frequency of term i, calculated by considering the distance between c and all occurrences of i in d. When calculating these pseudo-frequencies, the intuition is that the more distant an occurrence of the term is to position c, the less influence this term is expected to have around this position in the document, and so the less representative the term is expected to be of the topic being discussed around position c. Conversely, if a term i occurs at some position l, then the influence of this term is said to "propagate" to distant positions within the document. The extent to which the occurrence at position l propagates to other positions gradually decays with their distance from l.

In practice, the pseudo-frequency of term i is calculated by means of a kernel decay function that determines the extent to which an occurrence propagates to distant positions in the document. Depending on the decay rate and shape of such a kernel function, any occurrence of i in d can possibly affect the pseudo-frequency of term i for every other position in d. Conversely, the pseudo-frequencies of term i at position c can possibly be influenced by all occurrences of i in d as long as they are close enough to c. The kernel function in PMs is commonly parametrised by a propagation parameter σ which adjusts the influence that a term occurrence has over distant positions.

Several kernel density functions have been proposed in the past for PMs. Figure 6.2 shows plots for a representative set of these kernels. Among these, the Gaussian kernel shown in Equation 6.2

$$K(l, c) = \exp\left[\frac{-(l-c)^2}{2\sigma^2}\right] \tag{6.2}$$

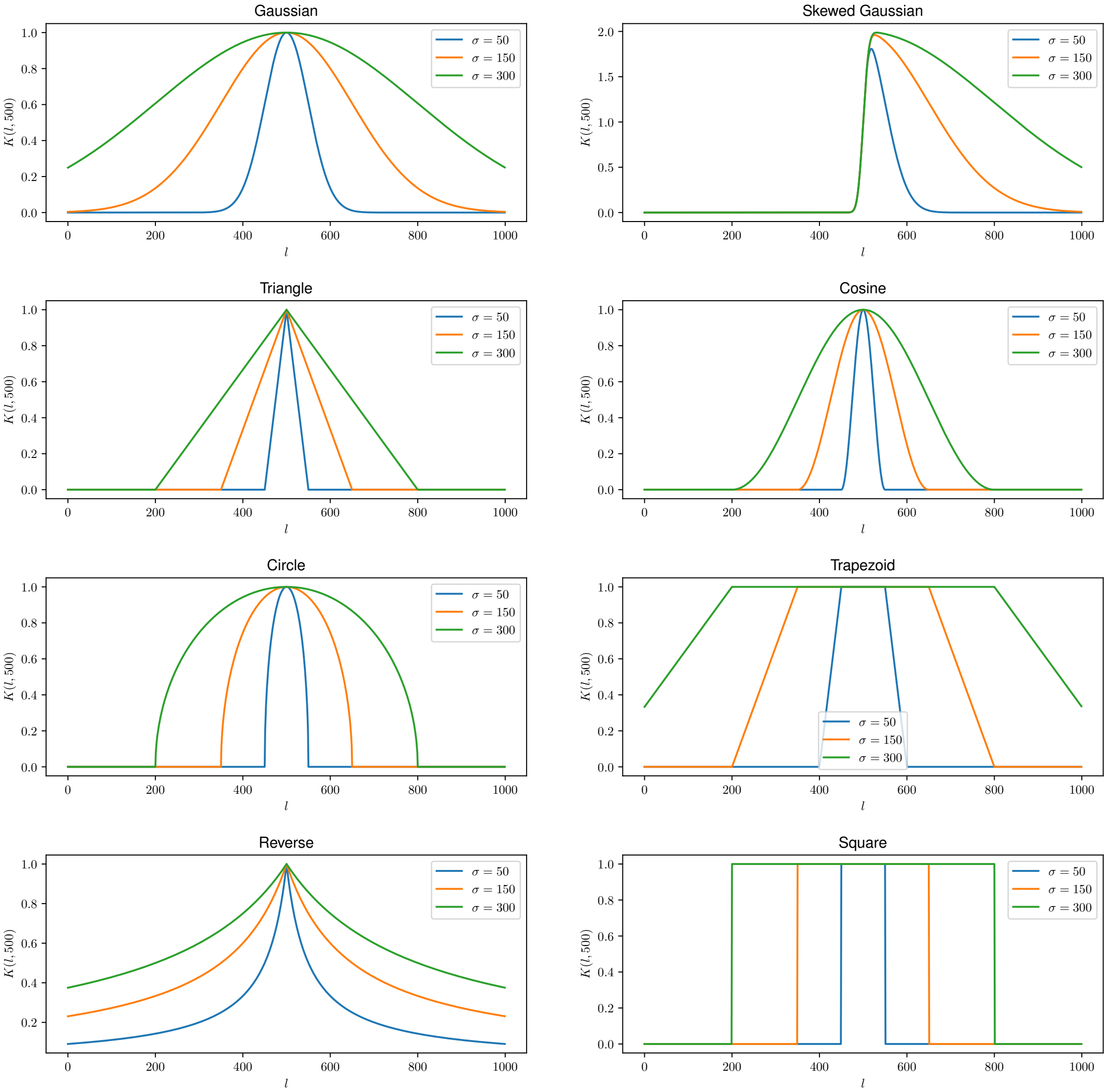
has been frequently shown effective in document and passage retrieval tasks with positional models (Lv and Zhai, 2009; Carmel et al., 2013). An exception to this is the work of Keikha et al. (2014), that showed that the skewed Gaussian kernel shown in Equation 6.3

$$K(l, c) = \exp\left[\frac{-(l-c)^2}{2\sigma^2}\right]\left[1 + erf\left(\frac{\alpha\,(l-c)}{\sqrt{2}}\right)\right] \tag{6.3}$$

with a positive skewness parameter $\alpha > 0$, can often outperform a Gaussian kernel in the task of finding answers to non-facto id questions within text articles. Keikha et al. (2014) attributed the superior performance of the positive skewed Gaussian kernel to its asymmetric shape, which has the effect of giving higher propagation values to positions located after the occurrence of a term than those located before. In the task of document retrieval, Song et al. (2011) showed that the Reverse kernel, shown in Figure 6.2, can provide superior retrieval performance compared to the Gaussian kernel when used in a PM to capture query term proximity heuristics.

In recent work, PMs were proposed as a contextualisation technique for passage retrieval (Carmel et al., 2013). In this work, a standard TF-IDF approach was used to compute the relevance score of a passage p within document d, where the frequency of

Figure 6.2: Kernel density functions proposed in previous research to calculate pseudo-frequency counts in PMs.



term i in p is given by its pseudo-frequency estimate, calculated as shown in Equation 6.4,

$$ptf_i = \sum_{c \in pos(i,d)} \sum_{l=p_1}^{p_n} K(l, c) \tag{6.4}$$

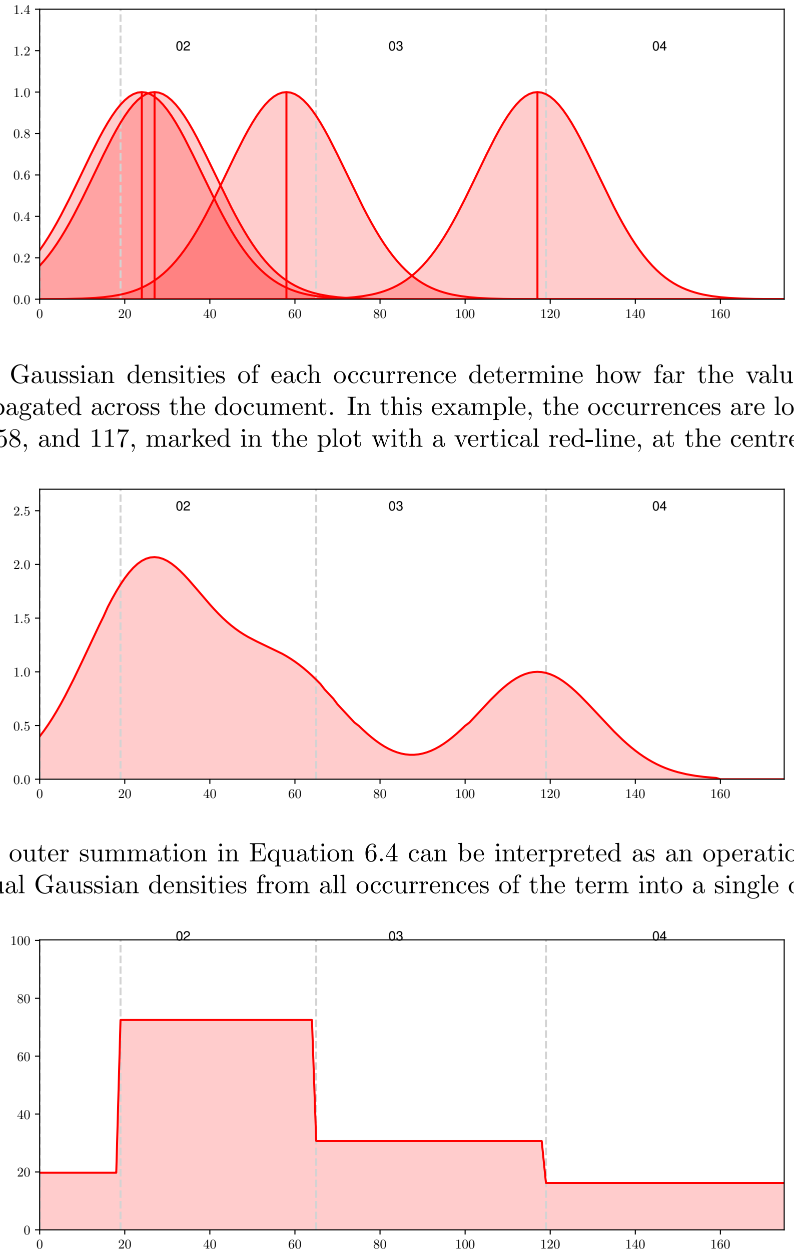
where $pos(i, d)$ denotes the set of positions where term i occurs in d, $p_1, \ldots, p_n$ are the spanning positions of p in d, that is, all positions within the passage boundaries, and K is a kernel density function. Based on Equation 6.4, the pseudo-frequency of term i for a passage p is calculated as a sum of discrete integrals over the kernel function for the range of positions across which p expands in d. Figure 6.3 gives a graphic description of how ptf_i is calculated according to Equation 6.4 when the Gaussian kernel is used. The final effect is that the value of an individual occurrence gets propagated onto nearby passages, even when these passages may not strictly "contain" an occurrence of the term.

If Equation 6.4 is used directly to compute ptf_i, then longer passages may unfairly obtain greater pseudo-frequency values. To avoid longer passages from receiving unmerited pseudo-frequency counts, Carmel et al. (2013) propose to apply the inner summation in Equation 6.4 across a fixed number of positions, independent of the length of the passages to be scored. To minimise the number of kernel calculations, in the experiments reported in this chapter with PMs, the inner summation is only applied at the position l within the passage that maximises $K(l, c)$ for every $c \in pos(i, d)$. In other words, the kernel function is only evaluated at the positions within a passage where the kernel gives its maximum value for every term occurrence. For a term i occurring at position c, this corresponds to evaluating $K(l, c)$ at $l = p_1$ or $l = p_n$ if $c < p_1$ or $c > p_n$ respectively, or at $l = c$ otherwise. Given this modification of Equation 6.4, ptf_i will increment by 1 each time the term i appears within the boundaries of p, and by $K(p_1, c) < 1$ or $K(p_n, c) < 1$ every time such term appears before or after the passage respectively.

Figure 6.4 shows the pseudo-frequency values that result from this approximation of Equation 6.4, for the same example from Figure 6.3. The final effect obtained is similar to that obtained from using Equation 6.4: term counts in a passage get propagated to neighbouring passages. Note that if the spread parameter σ of the kernel is 0, then passages only receive counts from the occurrences they contain. Whereas if $\sigma = \infty$, then all passages obtain the same value of ptf_i, equal to the frequency of the term in the document.

Although PMs were originally proposed within the language modelling framework for IR (Lv and Zhai, 2009), the idea of using a position dependent term count that gets propagated to distant positions in a document is general enough to be applied within other IR frameworks. In this respect, Carmel et al. (2013) used pseudo-frequency counts in a standard TF-IDF framework, while Song et al. (2011) did so within the probabilistic relevance framework (PRF). Similarly to the work from Song et al. (2011), the PM used in the experiments reported in this thesis is based on an adaptation of the probabilistic approach to pseudo-frequency counts. This adaptation is based on a similar idea than that used for the integration of prominence scores described in Section 5.2.4 for the CWL method.

Figure 6.3: Example of how pseudo-frequency counts are calculated when the Gaussian kernel is used. In this example the pseudo-frequency of a term is calculated for the passages 01, 02, 03, and 04 in a document that has 4 occurrences of the term.

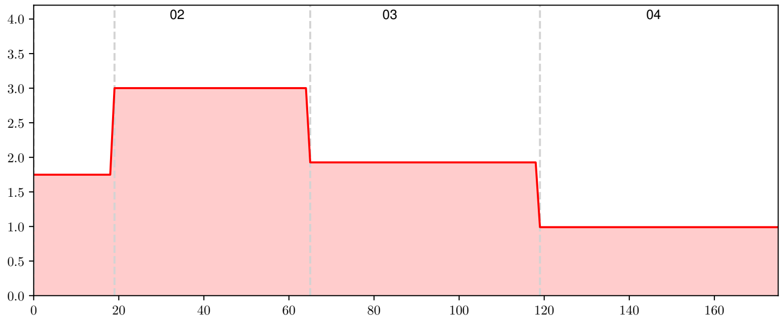


(a) The Gaussian densities of each occurrence determine how far the value of an occurrence is propagated across the document. In this example, the occurrences are located at positions 24, 27, 58, and 117, marked in the plot with a vertical red-line, at the centre of the Gaussians.

(b) The outer summation in Equation 6.4 can be interpreted as an operation that adds the individual Gaussian densities from all occurrences of the term into a single density contour.

(c) The inner summation in Equation 6.4 calculates the area under the density contour across all positions in each passage to obtain its final pseudo-frequency for the term.

Figure 6.4: Example of pseudo-frequency counts computed based on the maximum kernel values that can be obtained for a passage, for every term occurrence. The pseudo-frequency for passage 02 is equal to 3 because the first three occurrences of the term are fully-contained within this passage, while the occurrence at position 117 is not close enough to provide any additional contribution for this passage. Passage 04 obtains a ptf_i close to 1 as occurrence at position 117 appears just before the start of this passage.



By changing the document representation to consider pseudo-frequency counts instead of term frequencies for each term, while maintaining Poisson distributional assumptions for these pseudo-counts, the resulting model can be approximated by a BM25-like function, where the term frequencies tf_i are replaced by the pseudo-frequency counts ptf_i. The resulting retrieval function is then given by Equation 6.5.

$$S_{PM}(q,p) \;=\; \sum_{i \in q,p} \; \frac{(k_1+1)\,ptf_i}{ptf_i + k_1\,(1-b+b\,\frac{docl}{avel})} \; \frac{(k_3+1)\,qf_i}{k_3+qf_i} \quad cfw(i) \tag{6.5}$$

Note that, since $ptf_i = tf_i$ when $\sigma = 0$, the original BM25 formulation for scoring passages can be recovered from $S_{PM}(q,p)$ by setting $\sigma = 0$. Also, when $\sigma = \infty$, the pseudo-frequency ptf_i of the term equals the number of occurrences of i in the document d, and Equation 6.5 produces the same score as $S_{BM25}(q,d)$.

Recall that according to the original formulation of BM25, *docl* in Equation 6.5 is the length of the passage to be scored, and *avel* the average length across all passages in the collection. Ideally, these values should be updated in Equation 6.5 to reflect how the length of the original passage changes based on the pseudo-frequency estimates (Lv and Zhai, 2009). In the implementation of S_{PM} used in the experiments of this chapter, *docl* and *avel* were based on the original term frequency counts instead. Although this modification facilitated the implementation of the model within the Terrier framework, it presents some limitations. First, note that passage lengths given by pseudo-frequencies will be generally proportional to the original length of the passage as long as the spread parameter σ is set to small values. Contrary to this, for high values of σ, short passages will receive contributions from almost every term occurrence in the document, and thus their pseudo-frequency length will be substantially greater than their original length. In addition, since passages located at both ends of a document can only receive propagated counts from occurrences after or before such end points, these type of passages will tend to acquire lower pseudo-frequency counts than a passage located at the middle of the document. Despite these known limitations, using estimates of *docl* and *avel* based on pseudo-frequency counts would require us to re-estimate *docl* and *avel* every time a distinct value of σ is used, which would make any attempts to find optimal values of σ impractical.

While the PM technique seeks to contextualise the contents of a passage by propagating terms occurring close to it, the DSI technique described in Section 6.2.1 does so by extending the passage with contributions from all terms in the document. In this respect, the PM and DSI techniques can be interpreted as contextualising a passage with local and global context respectively. While PM puts more emphasis on local context, DSI makes no distinction between distant and local context. Because it may be beneficial to contextualise a passage with different levels of context granularity (Ogilvie and Callan, 2005; Arvola et al., 2011), the experiments presented in the next section additionally studied the performance of a technique that combines passage scores obtained with S_{PM} and document scores obtained with S_{BM25}. This scoring function is shown in Equation 6.6, where

$$S_{DSI\text{-}PM}(q,p) = \lambda\, S_{BM25}(q,d) + (1-\lambda)\, S_{PM}(q,p), \tag{6.6}$$

p is a passage contained in document d and q is the query. Equivalently, this technique can be given in terms of a interpolation of scores produced by a PM with $\sigma = \infty$ for capturing global context, and a second PM with a small value for σ for capturing locally focused context.

6.3 Experiments with contextualisation techniques

This section describes a series of retrieval experiments that seek to determine whether the contextualisation techniques presented in Section 6.2 can provide increased retrieval robustness against ASR errors.

6.3.1 Task and test collections

The potential benefits of using contextualisation techniques were investigated in a spoken passage retrieval task (SPR), where the elements to be retrieved contain significantly less occurrences of query terms and retrieval methods are thus more sensitive to recognition errors in the transcripts. Also, to see whether such techniques are helpful for alleviating the impact of ASR errors, their retrieval effectiveness was studied on transcripts with varying levels of recognition errors.

Recall from Chapter 4 that the BBC as well as the SDPWS speech collections were transcribed by different ASR systems. Compared to the various transcripts available for the SDPWS collection, those available for the BBC collection present similar levels of recognition accuracy. This is evidenced by the measures reported in Tables 4.3 and

6.3 for these document collections. Since the transcripts from the SDPWS2 collection were purposely generated by using acoustic and language models of decreasing quality, they provide a more diverse range of possible error levels in the transcripts. Table 6.1 summarises the list of transcripts available for the SDPWS2 collection along with short IDs used in the rest of this section to refer to each transcript type. Additionally, Table 6.2 reports passage length statistics for every transcript type, while Table 6.3 reports speech recognition accuracy.

Among the topic sets available for the SDPWS2 collection, the SQD1 and SQD2 sets contain spoken queries, whose transcripts are also available at different levels of transcription quality. This permits us to simulate additional retrieval conditions of increasing difficulty, by evaluating retrieval performance over increasingly noisier combinations of query and document transcripts. Table 6.4 lists the different combinations of query and passage transcripts used for evaluation sorted by their combined ranked-index accuracy (RIA). Although other query-passage combinations could have also been considered, in particular those involving low-quality query transcripts and high-quality passage transcripts (e.g. A0 for queries and M for passages), we limit our investigation to the combinations from Table 4.21 as these already capture a wide range of transcription quality levels. Also, recall that Kaldi models were not released by the NTCIR task organisers, but only the A0 transcripts of the SQD2 queries. For this reason, combinations involving A0 transcripts for the SQD1 queries were not considered in the experiments of this chapter.

As opposed to the WERs for the SQD1 and SQD2 query sets reported in Table 4.21, the recognition accuracy measures reported in Table 6.4 were calculated against passage transcripts by restricting terms to only those occurring in the queries from a query set. More specifically, if $acc(p_r, p_h)$ denotes a measure of recognition quality that compares the set p_r of term counts in the reference passage against the set p_h of terms counts in the hypothesised passage, then the value of this measure restricted to terms from a reference query q_r and an hypothesised version of this query q_h is $acc(p_r \cap q_r, p_h \cap q_h)$. For a query q, this metric can be calculated for every passage in the collection and their results averaged. Similarly, the same can be done for each query in a set of queries and these results can then be averaged across all queries from the set. The figures from Table 6.4 show these query-set averages for every accuracy measure. These figures give a rough idea of how many differences (or similarities in the case of BIA and RIA) exist between the set of matching terms obtained by using perfect transcripts for both query and passages, and that obtained by using noisy transcripts.

The SPR task considered with the SDPWS2 transcripts consists of ranking slide-group segments (SGS), and corresponds to the same task described in Section 5.3.1. Retrieval effectiveness in this task is evaluated with MAP. Part of the retrieval experiments reported in this chapter were conducted as part of our participation at the NTCIR-12 SpokenQuery&Doc-2 (SQD2) task whose official results are available in (Akiba et al., 2016). Since the spoken queries from the SQD2 task present similar characteristics than

Table 6.1: Transcripts from the SDPWS2 collection used in the contextualisation experiments.

SpokenQuery&Doc ID	Short ID
MAN	M
K-MATCH	A0
MATCH	A1
UNMATCH-LM	A2
UNMATCH-AMLM	A3

Table 6.2: Length statistics of segmented transcripts from the SDPWS2 collection.

Transcript	Passages	Avg. len.	S.D. len.	Max. len.
M	2,328	74.8	67.6	757
A0	2,330	74.2	67.4	760
A1	2,334	73.6	67.6	736
A2	2,335	80.7	74.8	806
A3	2,330	67.1	62.6	680

those used at the SQD1 task, part of the experimental work presented in this chapter focuses on maximising passage retrieval effectiveness in the SDPWS2 collection for this type of queries.

6.3.2 Maximising retrieval effectiveness via QF and exponential IDF

Besides the application of contextualisation techniques, additional methods were explored with the goal of maximising passage retrieval effectiveness for the SDPWS2 collection with the SQD1 and SQD2 queries.

Retrieval from small document collections with long queries

As described in Section 4.2.3, the spoken queries from the SQD1 and SQD2 sets were created following a set of guidelines that encouraged speakers to produce long queries, containing a large number of spoken terms. This is clear from the query length statistics presented in Table 4.20, which show that written queries from the SD2 set contain 6.77 terms on average, while those from the SQD1 and SQD2 sets contain, respectively, 24.13 and 30.77 terms on average in their manual transcripts. In addition to long queries, with 98 presentation transcripts and 2329 slide-group passages, the size of the SDPWS2 collection is several orders of magnitude smaller than most standard test document collections used

Table 6.3: Recognition accuracy of passages as measured by WER and index similarity metrics for the SDPWS2 collection.

Transcript	#Terms	WER	UTER	TER	BIA	RIA
M	6,230	0%	0	0	1.00	1.00
A0	6,350	22.0%	0.19	0.39	0.65	0.72
A1	6,131	43.7%	0.34	0.70	0.43	0.53
A2	11,219	67.5%	0.49	1.22	0.20	0.30
A3	14,190	70.5%	0.57	1.20	0.17	0.28

Table 6.4: Recognition accuracy of query terms in passages from the SDPWS2 collection for different combinations of query and document transcripts.

(a) SQD1 queries.

Transcripts		**Measures**			
Query	**Passage**	**UTER**	**TER**	**BIA**	**RIA**
M	M	0.00	0.00	1.00	1.00
M	A0	0.09	0.20	0.83	0.86
M	A1	0.18	0.45	0.65	0.70
M	A2	0.32	0.59	0.53	0.59
A1	A1	0.27	0.94	0.43	0.50
M	A3	0.46	0.78	0.39	0.48
A1	A2	0.41	0.95	0.35	0.42
A2	A2	0.47	0.97	0.30	0.37
A1	A3	0.52	1.10	0.26	0.35
A2	A3	0.58	1.05	0.24	0.31
A3	A3	0.60	1.10	0.20	0.28

(b) SQD2 queries.

Transcripts		**Measures**			
Query	**Passage**	**UTER**	**TER**	**BIA**	**RIA**
M	M	0.00	0.00	1.00	1.00
M	A0	0.10	0.20	0.84	0.87
M	A1	0.18	0.43	0.68	0.74
A0	A0	0.19	0.51	0.63	0.72
A0	A1	0.26	0.62	0.55	0.64
M	A2	0.32	0.59	0.53	0.59
A1	A1	0.26	0.83	0.48	0.57
A0	A2	0.43	0.71	0.45	0.53
M	A3	0.47	0.79	0.41	0.51
A1	A2	0.41	0.87	0.39	0.48
A0	A3	0.56	0.88	0.32	0.42
A2	A2	0.45	1.00	0.33	0.40
A1	A3	0.54	1.06	0.28	0.39
A2	A3	0.58	1.09	0.25	0.34
A3	A3	0.59	1.16	0.23	0.33

in IR research. A third distinctive characteristic of the academic talks from the SDPWS2 collection is that they are highly homogeneous in terms of the range of topics and domains discussed. As described in Section 4.2.2, most of the talks from the SDPWS2 are highly technical and contain a significant amount of domain-specific vocabulary.

Retrieving content from documents with the characteristics of the SDPWS2 collection may pose additional challenges to conventional retrieval methods. Because of the size of the collection, a large proportion of terms from all possible terms in the Japanese language will be missing or underrepresented. Thus, terms that are frequently used and that would normally occur in a large proportion of documents in a larger collection, may occur in a significantly smaller proportion of documents in the SDPWS2 collection. Under these circumstances, IDF scores may not provide a reliable estimate of the relative importance that terms should be given when calculating their contribution to document relevance scores. In particular, IDF estimates for underrepresented terms would be unusually high, and therefore closer in magnitude than the IDF scores given to the less frequent terms in the collection. For instance, the term “家” (“home”, “family”, “household”), which is one of the top 200 most frequent words in Japanese[1] appears in 4 passages in the SPDWS2 collection, while the terms “ディリクレ” (“Dirichlet”) and “コンピュータ” (“computer”) do so in 2 and 21 passages respectively. Considering that the total number of passages in the SDPWS2 collection is $N = 2329$, then the IDF scores for “home”, “Dirichlet”, and “computer” are 9.01, 9.86, and 6.74 respectively[2]. Even though the terms for “Dirichlet” and “computer” would arguably be more useful for retrieval if used in a query, the term for “home”, which a-priori seems to be less useful for content retrieval, acquires an overrated IDF score which puts it in a similar level of importance than other terms with likely higher power to discriminate between relevant and irrelevant content in the SDPWS2 collection.

The potential difficulties that unreliable IDF scores may pose to retrieval in the SDPWS2 collection are accentuated if considering the characteristics of the SQD1 and SQD2 queries. Because these queries are extremely verbose, they tend to contain a high number of low content-bearing terms, such as the term for “home” from the previous example. These low-quality terms may not only increase the number of spurious document matchings in the retrieval process, but also have a major impact on the overall relevance score assigned to documents. Specifically, if a query contains a large number of low content-bearing terms with underestimated document frequency, these terms may dominate the summation of term scores for a document and thus diminish the contributions to relevance scores from more topically informative terms.

Improving term weight estimations in small collections

One technique that has been used in the past to improve the estimation of IDF scores in small spoken collections is to use an external document collection to either re-calculate the

[1] `http://corpus.leeds.ac.uk/frqc/internet-jp.num`

[2] based on the collection frequency weight (cfw) from Equation 2.7 and $\log_2$

document frequencies or expand the contents of the documents from the original collection with topically related terms (Johnson et al., 2000, 1999a; Singhal et al., 1999). Although expansion techniques were successfully applied to collections of broadcast news in the TREC SDR tasks, they are only effective if the external collection used is representative of the collection that is the target of retrieval. As opposed to collections of broadcast news, for which parallel corpora exist and are easily available, the SDPWS2 content is highly specific to a particular technical domain for which it is difficult to find appropriate external data to use for the implementation of expansion techniques or for the re-estimation of IDF weights.

In the absence of an appropriate external collection, three alternative approaches were adopted in this thesis to ameliorate the effects of using poorly estimated document frequency statistics and verbose queries in the experiments conducted with the SDPWS2 collection. The first technique was described in Section 4.2.2 and consists of removing low content-bearing terms from the queries and documents. By removing stop words and only keeping terms identified as verbs and nouns in the transcripts, the number of term matchings corresponding to unimportant terms can be reduced dramatically. The second and third techniques consist of exploiting term frequencies in the query (QF) and raising the value of IDF to the power of some positive number $d \geq 1$. The following sections motivate and describe these two techniques.

Within-query term frequency (QF)

When queries are long, it is often beneficial to exploit term frequencies in the query, which may provide useful information about which terms should be given increased weights during document score computations. In the Okapi BM25 model (Equation 2.9), within-query term frequencies are accounted for by the factor $\frac{(k3+1)qf_i}{k_3+qf_i}$, where qf_i denotes the count of term i in the query. This query-frequency factor (QF) is parametrised by the k_3 constant, which controls the rate at which the factor increases with every unit increment of qf_i, as well as the point at which it reaches its asymptotic maximum. The general assumption underlying the use of QF is that terms that occur more frequently in the query are more representative of the user's underlying information need. For retrieval from the SDPWS2 collection, the hope is that QF estimates may help signal important terms in the query and then increase their overall contribution to relevance scores over less important terms. Following the example before, if the term "Diritchlet" appears more frequently in a long query than low-quality terms like "home", then enabling the QF factor in BM25 would increase the difference between the weights assigned to these terms.

Exponential inverse document frequency (EIDF)

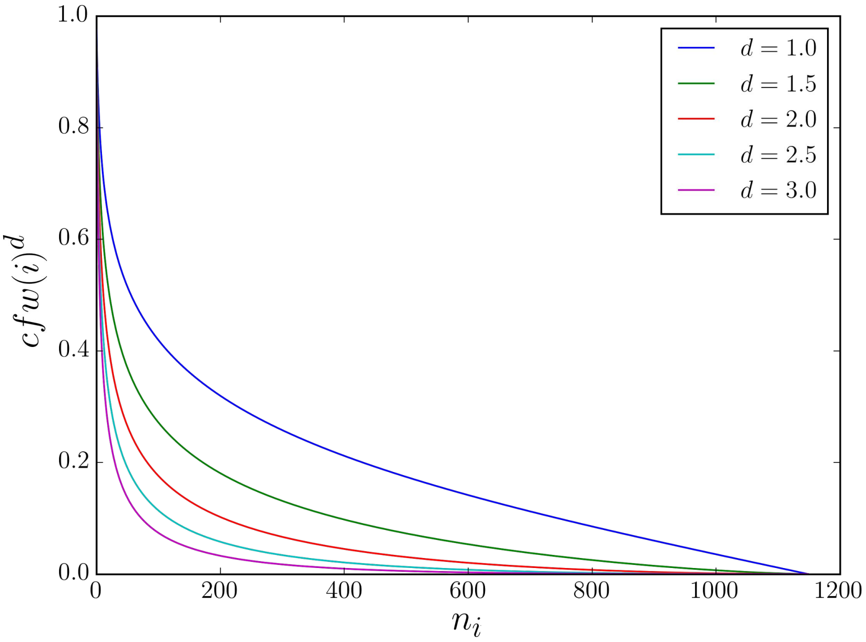
A small collection of documents can be interpreted as a sample of documents taken from a larger population of documents. If the sample is too small, then the document frequencies of terms calculated for the sample will not be representative of the document frequencies calculated from the population. Terms with high document frequencies in the population may be underrepresented in the sample and obtain low document frequencies and consequently high IDF scores. If the sample were to be augmented with documents from the population one document at a time, the document frequencies of terms in the extended sample would slowly converge to those seen in the population. The effect would likely be that document frequencies of highly frequent terms in the population would increase at a faster rate than those of rare terms, gradually reflecting the larger differences in document frequencies that exist between these two term groups. In other words, the difference of IDF scores between terms with high and low population frequencies would gradually increase, establishing a bigger separation between these two groups.

If population frequencies are unknown, a similar separation effect can be achieved between the IDF scores of low and high frequency terms by raising the standard IDF values to the power of some constant $d \geq 1$. To achieve this effect in the experiments reported in this thesis, a fourth parameter $d \geq 1$ was then included in the Okapi BM25 function (Equation 2.9) as the exponent of the collection frequency weight $cfw(i)^d$. The d parameter can then be adjusted to control the shape and slope of the $cfw(i)$ function and thus increase the relative difference between weights assigned to frequent and rare terms. Figure 6.5 shows this effect for different values of the exponent d. By using this alternative function with $d = 3$, the IDF scores assigned to terms "home", "Dirichlet", and "computer" in the SDPWS2 passage collection would be 732.3, 959.4, and 307.1 respectively.

Although the proposed modification for IDF scores from Equation 6.5 may seem unconventional, a similar modification has previously been proposed by Zhai (2001) and implemented as the TF-IDF model in the Lemur Toolkit[3]. Besides this work, in work parallel to ours, Murata et al. (2014, 2016) proposed a different alternative function for computing IDF scores that also seeks to degrade the scores of underrepresented terms in the collection in favour of terms that are more discriminative for retrieval. That work was conducted in the context of the instance search task at TRECVID (Over et al., 2014), which poses the task of finding instances of persons, objects, or places, within videos given an example image. Murata et al. (2014) observed that using the standard formula of $cfw(i)$ in the Okapi BM25 function performed poorly in this task when "key-points", these are, pixel-derived features extracted from the query and documents, are treated as the "terms" upon which the matching process is performed. The authors attributed the poor performance of the original $cfw(i)$ weights to common background pixels occurring in the query and video images which tended to dominate the final BM25 scores of the videos over "foreground" key-points that provided stronger evidence of relevance. The alternative $cfw(i)$ function proposed by Murata et al. (2016) is given in Equation 6.7,

[3] http://www.lemurproject.org

Figure 6.5: Exponential collection frequency weight ($cfw(i)^d$) for different values of the exponent d and $N = 2329$ (size of the SDPWS2 collection). For ease of comparison, all values are normalised between 0 and 1 although the true value scales may vary significantly for different values of d.



where γ acts as a tuning constant.

$$BEIDF(i) = \log \frac{e^{-n_i/\gamma}(N - n_i + e^{n_i/\gamma} - e^{-n_i/\gamma} + 1)}{(e^{n_i/\gamma} - e^{-n_i/\gamma} + 1)(n_i + e^{-n_i/\gamma})} \tag{6.7}$$

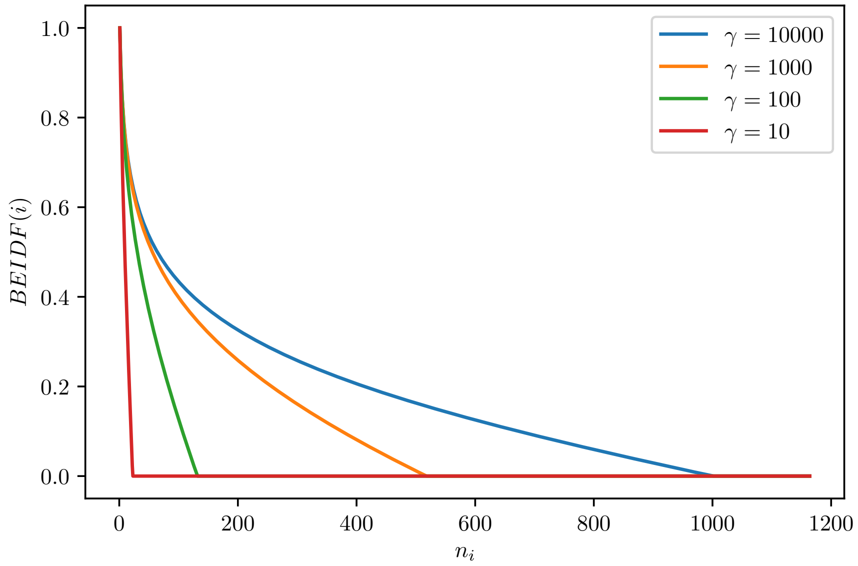
In practice, to avoid considering negative values in the document scores, Murata et al. modify Equation 6.7 to output 0 whenever the argument of the log is less than 1.

Figure 6.6 plots $BEIDF(i)$ for various values of γ. For large values of γ, the function produces similar weights to those produced by the standard $cfw(i)$ function, whereas for small values of γ the function marks a sharp boundary between terms with low and high document frequency.

Experiments with QF and exponential IDF

Passage retrieval experiments were carried out to explore if within-query frequencies and exponential IDF can help to alleviate the problems associated with verbose queries and small collections. For this purpose, the effectiveness of the BM25 function with and without QF and exponential IDF (EIDF) was measured for various combinations of test collections and transcripts. Table 6.5 presents MAP scores obtained with the standard BM25 function with: (i) QF and EIDF disabled; (ii) only QF enabled; (iii) only EIDF enabled ($cfw(i)^d$); and (iv) both QF and EIDF enabled. In each evaluation condition, the BM25 parameters b, k_1, k_3, and d were optimised for each topic set and passage collection

Figure 6.6: Bayesian Exponential IDF ($BEIDF(i)$) for different values of the parameter γ and $N = 2329$ (size of the SDPWS2 collection). For ease of comparison, all values produced by the function are normalised between 0 and 1.



by using the coordinate ascent algorithm described in Appendix D.

The results in Table 6.5 show that increased retrieval effectiveness can be obtained if exploiting within-query term frequencies (QF) and using exponential collection frequency weights (EIDF) in the standard BM25 function. Nonetheless, these techniques seem to provide substantial improvements only for the last 4 experimental conditions in Table 6.5, where retrieval is done with long spoken queries over a small collection of passages (SDPWS2). For experiments conducted with the BBC collection, which is relatively larger than the SDPWS2 collection, and with the SD2 queries, which are shorter than the queries from the SQD1 and SQD2 sets, enabling QF and EIDF in the BM25 function

Table 6.5: Passage retrieval effectiveness of Okapi BM25 with QF and EIDF disabled ($k_3 = 0$, $d = 1$), with QF enabled ($k_3 > 0$), with EIDF enabled ($d > 1$), and with both QF and EIDF enabled. For these results, BM25 parameters are optimised on test queries.

Collection	**Topics**	**Transcript**		**Models**			
		Queries	**Documents**	**BM25**	**+QF**	**+EIDF**	**+ALL**
BBC	SH13	MAN	LIMSI	.315	.315	.315	.315
	SH14	MAN	LIMSI	.337	.337	.337	.337
	SH14	MAN	NST	.330	.330	.330	.330
	SAVA	MAN	LIMSI	.304	.304	.304	.304
	SAVA	MAN	NST	.242	.242	.246	.246
SDPWS2	SD2	MAN	MAN	.450	.450	.456	.457
	SQD1	MAN	MAN	.241	.257	.254	.291
	SQD1	MATCH	MATCH	.168	.209	.176	.210
	SQD2	MAN	MAN	.258	**.290**	**.268**	**.303***
	SQD2	K-MATCH	K-MATCH	.236	.255	.245	**.270***

Table 6.6: Passage retrieval effectiveness of Okapi BM25 when $cfw(i)^d$ (EIDF) or Equation 6.7 (BEIDF) are used as collection frequency weight. Results are for best performing values of the parameters d and γ in each condition.

Topics	**Transcript**		**Models**	
	Queries	**Documents**	**EIDF**	**BEIDF**
SQD1	MAN	MAN	.254	.240
SQD1	MATCH	MATCH	.176	.170
SQD2	MAN	MAN	.268	.266
SQD2	K-MATCH	K-MATCH	.245	.247

does not provide any improvements in retrieval effectiveness. In conditions where these techniques are effective, applying both techniques in combination results better than using either in isolation. Overall, the results indicate that it is beneficial to use QF and EIDF in BM25 when performing retrieval with long queries from small collections.

In the previous section, two variations of EIDF were presented: $cfw(i)^d$ and the BEIDF from Equation 6.7. Table 6.6 compares the retrieval effectiveness achieved by these two variations of EIDF on the SQD1 and SQD2 topics and the SDPWS2 collection. As can be seen from the results, the two variations of EIDF perform similarly and obtain comparable MAP scores in these test conditions. The fact that both formulations of EIDF are equally effective in this set-up suggest that the simpler $cfw(i)^d$ formula is able to replicate the effects produced by the more theoretically sound BEIDF function. Thus, despite its ad-hoc nature, the $cfw(i)^d$ formula seems to provide a good practical approximation of BEIDF.

6.3.3 Contextualisation experiments

To study the potential for context to improve passage ranking for retrieval in noisy conditions, the effectiveness of the contextualisation techniques presented in Section 6.2 was evaluated on various combinations of query and document transcripts of the SDPWS2 collection. More specifically, experiments were conducted to measure the effectiveness of the document score interpolation (DSI) technique presented in Section 6.2.1, the positional variation of BM25 (PM) described in Section 6.2.2, and a combination of these two (DSI-PM). The DSI technique contextualises a passage with the contents of its container document, while the PM technique does so by putting heavier emphasis on local rather than global context. The technique that combines DSI with PM makes use of both global and local context when calculating the relevance score of a passage. In all experiments conducted with these contextualisation models, BM25 weights were calculated by enabling the QF factor ($k_3 > 0$) as well as using exponential cfw weights ($d > 1$), since these modifications result in improved retrieval quality as demonstrated in Section 6.3.2.

To understand the behaviour of these contextualisation techniques in increasingly noisier conditions, experiments were conducted with different combinations of query and document transcripts. Each combination of transcripts imposes a different evaluation condition with a varying level of noise, each of which may require adjusting model parameters differently in order to achieve optimal performance. Furthermore, in order to study the relative importance assigned to contextual evidence in increasingly noisier conditions, it is informative to find values for the contextualisation parameters σ and λ that provide the best performance in each noise condition. Consequently, parameters were optimised for each model by seeking to maximise retrieval effectiveness in each noise condition. To obtain an unbiased estimate of the relative performance of these techniques, parameters were first optimised on the SQD1 queries (training data), and retrieval models using these optimal parameters were then evaluated on the SQD2 queries (test data).

In the PM technique, 5 parameters were optimised: b that adjusts the degree of length normalisation; k_1 and k_3 which control the rate of increase of the TF factor as the (pseudo) frequency of a term increase in the passage and query respectively; the newly incorporated parameter d which controls the rate of decrease of the IDF factor as the collection frequency of a term increase; and σ that widens the scope of occurrences of query terms so that they can influence the score of more distant passages. The DSI technique produces two independent rankings, one for documents and another one for passages. Since the optimal BM25 parameters may differ for document and passage rankings, 9 parameters were optimised for the DSI technique: 4 corresponding to b, k_1, k_3, d for each of the BM25 functions that produce document and passage rankings, and λ that controls the influence of document evidence in the passage scores. Lastly, for the DSI-PM technique, 10 parameters were optimised: the 9 parameters corresponding to DSI plus σ used in the PM function to create the initial scores of passages.

Table 6.7 reports MAP scores obtained with a baseline BM25, which does not contextualise passages, and with the PM, DSI, and DSI-PM contextualisation models for queries that were used as training data (SQD). In this table, MAP scores in bold are statistically significantly greater than those obtained with the BM25 baseline based on a paired t-test ($p < 0.05$). These results demonstrate that it is possible to obtain substantial and consistent improvements in passage retrieval effectiveness by using contextualisation techniques for this set of spoken queries. Furthermore, the relative improvements over the BM25 baseline tend to be greater for noisier combinations of query and document transcripts. In some cases, the same level of retrieval performance obtained with high-quality transcripts can be obtained with low-quality transcripts by using contextualisation models. For instance, BM25 obtains a MAP of .241 for M queries and A0 transcripts, while DSI-PM can reach .258 MAP for the substantially noisier A1-A3 combination.

Table 6.8 reports MAP scores obtained by the BM25, DSI, PM, and DSI-PM models on the test queries (SQD2). In this case, the parameters used in each model and evaluation condition were those found optimal for the SQD1 queries, thus the figures from Table 6.8 should be considered as better indicators of the generalisation power of the contextualisation models compared to those presented in Table 6.7. For transcript combinations of the form A0-X, that is, all those that involve using A0 transcripts for the queries and some other transcript type X for the documents, the parameters used were those obtained for

Table 6.7: Retrieval effectiveness (MAP) of contextualisation models for training queries (SQD1). These results are for the best performing parameter settings found for the same set of queries (SQD1) in each evaluation condition. Percentages next to MAP scores show relative improvements with respect to the BM25 baseline.

RIA	**Transcripts**		**Models**						
	Query	**Doc.**	**BM25**	**PM**		**DSI**		**DSI-PM**	
100%	M	M	.291	.312	+7%	**.353**	+21%	**.366**	+26%
86%	M	A0	.241	**.315**	+29%	**.311**	+29%	**.320**	+33%
70%	M	A1	.218	.279	+28%	.253	+16%	.285	+31%
59%	M	A2	.093	**.179**	+92%	**.176**	+89%	**.194**	+109%
50%	A1	A1	.219	.298	+36%	**.303**	+38%	.290	+32%
48%	M	A3	.154	**.261**	+69%	**.206**	+34%	**.275**	+79%
42%	A1	A2	.097	**.160**	+65%	.147	+52%	**.170**	+75%
37%	A2	A2	.112	.141	+26%	.184	+64%	.201	+79%
35%	A1	A3	.125	**.248**	+98%	**.162**	+30%	**.258**	+106%
31%	A2	A3	.098	**.195**	+99%	.143	+46%	**.192**	+96%
28%	A3	A3	.101	**.186**	+84%	**.165**	+63%	**.202**	+100%

SQD1 for the combinations M-X. MAP values in bold and those marked with *, †, and ◇ indicate statistically significant differences with respect to BM25, DIS, PM, and DIS-PM respectively based on a *MaxT* permutation test that corrects for multiple hypothesis testing (Boytsov et al., 2013). In this case, *MaxT* tests were performed to compare every pair of runs from a single evaluation condition (row in Table 6.8). For the *MaxT* tests, the number of permutations used was $B = 100,000$ and the level of significance set to $\alpha = 0.05$.

Overall, the results from Table 6.8 indicate that using global (DSI) and local (PM) context either in isolation or in combination (DSI-PM) provide significant gains in retrieval effectiveness across most evaluation conditions. Moreover, the DSI-PM method which makes use of both local and global context to expand the passage representation, tends to obtain higher MAP scores on average than if using the DSI or PM methods alone. Similarly to the observations made from the results of the training queries (SQD) in Table 6.7, the relative gains of using context in highly noisy conditions ($RIA < 60\%$) are greater on average than in less noisy conditions.

Effects of varying the contextualisation parameters

The results from Tables 6.7 and 6.8 indicate that the retrieval effectiveness of the contextualisation methods degrades at a lower rate than that of a standard passage retrieval approach, when the queries and document transcripts contain higher amounts of ASR errors. These results demonstrate that these techniques can make retrieval methods more robust to ASR errors when the units to be retrieved are short in length and its retrieval more likely to be negatively affected by transcription errors.

Recall from the descriptions of the DSI and PM models, that the λ and σ parameters control the emphasis that is given respectively to the global and local context that sur-

Table 6.8: Retrieval effectiveness (MAP) of contextualisation models for test queries (SQD2). These results are for the best performing parameter settings found for the training queries (SQD1) in each evaluation condition. Percentages next to MAP scores show relative improvements with respect to the BM25 baseline.

RIA	**Transcripts**		**Models**						
	Query	**Doc.**	**BM25**	**PM**		**DSI**		**DSI-PM**	
100%	M	M	.272	.291	+6%	**.305**	+12%	.314	+15%
87%	M	A0	.261	.294	+12%	.274	+5%	.299	+14%
74%	M	A1	.228	.267	+16%	**.272**	+19%	.278	+21%
72%	A0	A0	.261	.292	+11%	.254	-3%	.293	+12%
64%	A0	A1	.236	.267	+13%	**.271**	+14%	.274	+15%
59%	M	A2	.085	**.178**	+108%	**.160**	+87%	**.174**	+103%
57%	A1	A1	.186	**.239**	+28%	**.249**	+34%	**.277**	+49%
53%	A0	A2	.091	**.177**	+94%	**.166**	+82%	**.172**	+88%
51%	M	A3	.119	**.209**	+75%	**.214**	+80%	**.192**	+60%
48%	A1	A2	.097	.124	+27%	**.126**	+29%	.139	+43%
42%	A0	A3	.111	**.173**	+55%	**.191**	+72%	**.167**	+50%
40%	A2	A2	.117	.096	-21%	.112	-4%	.155†	+33%
39%	A1	A3	.048	**.144***	+200%	**.142◇**	+196%	**.134**	+178%
34%	A2	A3	.066	.100	+52%	**.122**	+86%	.126	+91%
33%	A3	A3	.095	.145	+52%	**.146**	+53%	**.168**	+76%

rounds a passage in the calculation of its relevance score. A question that remains to be answered is thus whether these models can benefit from using larger amounts of context in increasingly noisier conditions. In other words, is it effective to increase the emphasis given to context in the score of a passage when the queries and documents contain a higher number of transcription errors? This section seeks to answer this question by studying the effects of varying the contextualisation parameters λ and σ in the DIS and PM scoring functions.

Figure 6.7 shows how MAP scores vary for increasing values of σ in six representative evaluation conditions. Each of the six lines in the plot was generated based on the optimal parameter settings for the SQD2 queries, by evaluating the PM function for $\sigma = 0, \ldots, 800$ while leaving fixed b, k_1, k_3, and d. Recall that larger values of σ increase the width of the Gaussian kernels and thus magnify the influence that individual term occurrences have over distant passages. For perfect or quasi-perfect transcripts (M-M and A0-A0), the model achieves maximum performance for $\sigma = 76$ and $\sigma = 111$ respectively, while for moderately noisy transcripts (A1-A1 and M-A3) it does so for $\sigma = 296$ and $\sigma = 341$ respectively. Finally, in extremely noisy conditions (A2-A2 and A2-A3), the maximum points are located at $\sigma = 530$ and $\sigma = 682$ respectively. These observations provide supporting evidence for the claim that longer spans of context become increasingly useful for retrieval as ASR errors increase in the transcripts.

Figure 6.8 shows the effects of changing the interpolation parameter λ from 0 to 1 in the DSI model for the same set of transcript combinations plotted in Figure 6.7. Recall that for smaller values of λ, the DSI model places more emphasis on passage scores than document scores, whereas for $\lambda \approx 1$ more emphasis is put on document scores than on passage

Figure 6.7: MAP scores on SQD2 obtained with the PM method for six representative evaluation conditions and $\sigma \in [0, 800]$. Plots for the other transcript combinations evaluated follow similar trends.

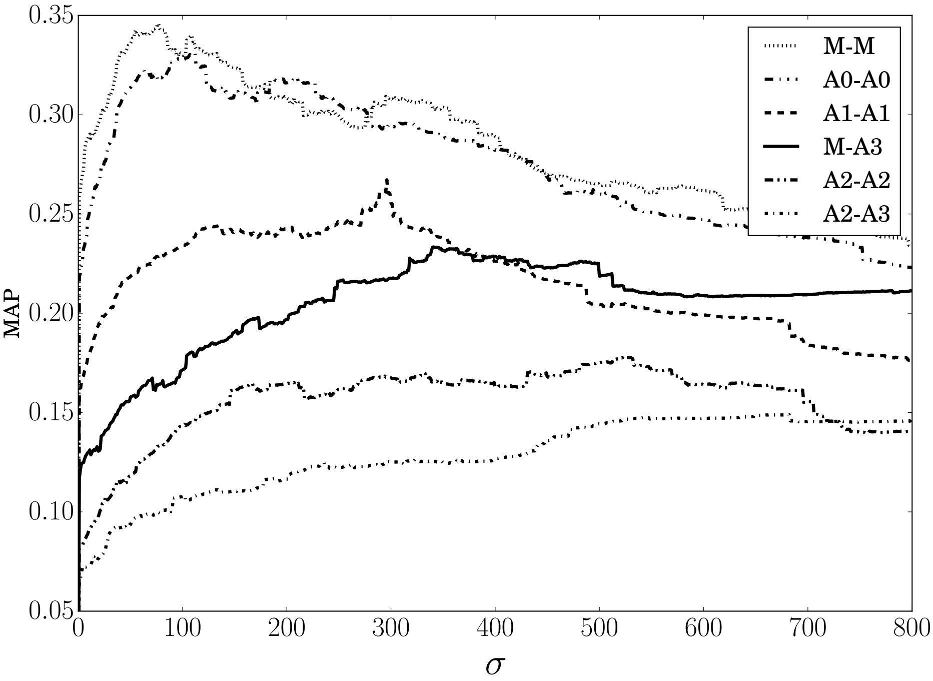
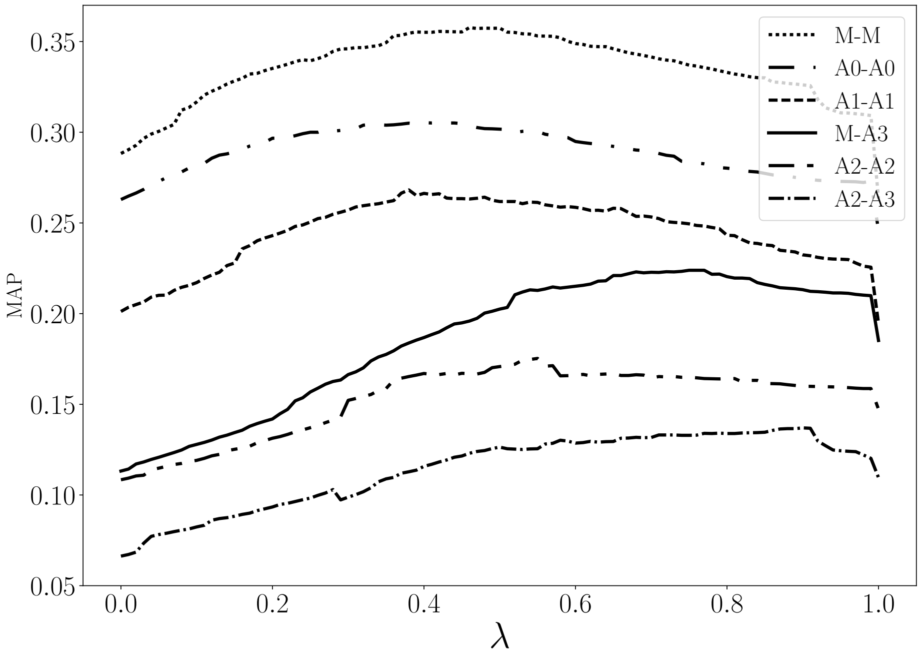


Figure 6.8: MAP scores on SQD2 queries obtained with the DSI method for six representative conditions and $\lambda \in [0, 1]$.



scores. Based on these plots, it can be seen that higher passage retrieval effectiveness can be obtained when document scores are used in combination with passage scores. The curves in the plot appear to demonstrate two different trends. On one hand, the MAP curves associated with high-quality transcripts tend to have peaks at lower values of the $[0, 1]$ range, specifically at $\lambda = 0.46$, $\lambda = 0.40$, and $\lambda = 0.38$ for M-M, A0-A0, and A1-A1 respectively. On the other hand, curves associated with low-quality transcripts tend to have maximums at higher values of the $[0, 1]$ interval, corresponding to $\lambda = 0.77$, $\lambda = 0.55$, and $\lambda = 0.90$ for M-A3, A2-A2, and A2-A3 respectively. Thus, document scores (global context) become increasingly beneficial for passage retrieval as the amount of mismatches due to ASR errors increase in the transcripts.

6.3.4 Confidence adaptive contextualisation

The experiments presented in Section 6.3.3 show that a relevant passage can be ranked more effectively if evidence from additional occurrences of query terms around the passage is considered in the calculation of the passage's relevance score. These results also show that as the amount of speech recognition errors increase in the transcripts, greater importance can be given to context contributions in order to increase the robustness of these retrieval functions to ASR errors. In highly noisy conditions, term statistics become less reliable if calculated within short passages in the transcripts, so considering scores or term counts based on a expanded version of the passages or the full-document normally result in enhanced passage scores and rankings.

Along with recognition hypothesis of words spoken in an utterance, the majority of ASR systems can also produce confidence scores for words or word sequences. Recall from Section 2.2.4 that a confidence score is a numeric estimate, typically in the $[0, 1]$ range, of the level of uncertainty the ASR system has about a particular recognised word or sequence. Given the observation that context becomes increasingly important in transcripts with increasingly higher error rates, it seems reasonable to consider the possibility of adjusting the contextualisation parameters of the DIS and PM functions according to the levels of speech recognition confidence found in the transcripts. The passage scoring process could then be modified to rely more strongly on context for scoring passages that have been recognised with low confidence, while reducing the contribution of context for passages that have been recognised with high confidence. The key intuition is that context becomes less useful for passages with reliable transcriptions, since these would normally provide accurate term frequency statistics from which their relevance score can be determined with an effective level of accuracy. In contrast, context has higher potential to improve the ranking of relevant passages with low-confidence speech recognition, since these are likely to provide less reliable count statistics which would translate into unreliable passage scores.

Adapting contextualisation parameters to confidence estimates

Given confidence scores for each term in a document transcript, a rough confidence estimate for a passage, $c(p) \in [0, 1]$, can be obtained by averaging the confidence scores of the terms contained in the passage. The complement of the passage's confidence score, $u(p) = 1 - c(p)$ can therefore serve as an indicator of the uncertainty with which the contents of the passage were transcribed. In order to increase the incidence of context in the calculation of the relevance score of low-confidence passages, the contextualisation parameters in the DSI and PM methods can be increased proportionally to $u(p)$.

For the DSI function shown in Equation 6.1, the interpolation parameter $\lambda \in [0, 1]$ controls the incidence that the document context has over the final score of the passage. Instead of using the same value of λ for all passages, a passage-dependent value $\lambda_p = \frac{u(p)+\lambda}{2}$ equal to the average of the original λ and the uncertainty score of the passage can be used as the interpolation parameter. Under this alternative set-up, the DSI function will tend to put more emphasis on document scores when scoring passages with high transcription uncertainly.

In the case of the PM retrieval function (Equation 6.5), the $\sigma > 0$ parameter controls the width of the Gaussian kernel, and therefore determines the extend to which a term can propagate to distant positions and influence the score of neighbouring passages. Given a maximum kernel width $\sigma^{\vee}$, a possible definition for a passage-dependent σ_p based on the passage's transcription uncertainty is $\sigma_p = u(p)\, \sigma^{\vee}$. The final effect of using these passage-specific values is then to increase the extent to which a term occurring in the document can influence the score of a passage that has high transcription uncertainty (low confidence). In this case, $\sigma_\vee$ will determine the maximum value that σ_p can acquire, assigned only to passages with extreme uncertainty levels ($u(p) = 1$).

Experiments with adaptive contextualisation techniques

To investigate the potential benefits of adapting the contextualisation incidence parameters based on confidence scores, SPR experiments were carried out with the SQD1 and SQD2 queries over a selection of document transcript combinations from the SDPWS2 collection.

Recall from the description of the SDPWS2 transcripts in Section 4.2.2 that each character sequence from a 1-best hypothesis was re-tokenised by using the morphological analyser MeCab. Since the LMs used for recognition were generated by using the analyser ChaSen, the re-tokenisation process with MeCab frequently produced a different word sequence than the one present in the ASR's 1-best hypothesis. This new word sequence normally included tokens that were not present in the vocabulary of the LM used to decode the ASR hypotheses. Although time stamps could be obtained for MeCab's tokens by running force-alignment, obtaining posterior-based confidence scores for these alternative tokens is not trivial. For this reason, the experiments reported in this section were carried out with the original tokens from the ASR transcripts (ChaSen's) for which there exist reliable confidence scores based on word posterior probabilities. Because of these tokenisation differences, the MAP scores of the experiments reported in this section may differ from those reported in Section 6.3.3.

Since the focus of these experiments is on the utilisation of confidence scores, it is important to ensure that the confidence estimates produced by the Kaldi and Julius ASR systems provide meaningful information. Figure 6.9 shows the distribution of confidence scores extracted from the terms from the transcripts K-MATCH (A0), MATCH (A1), UNMATCH-LM (A2), and UNMATCH-AMLM (A3). Recall that only the first of these (K-MATCH) was produced by Kaldi, while the remaining ones were generated by Julius. The plots show that the vast majority of the confidence scores produced by Kaldi are equal to or near 1.0. In contrast, those produced by Julius are more evenly distributed in the $[0, 1]$ interval and thus seem to be more informative overall. Furthermore, there is a clear distinction in the distribution of confidence scores between high (MATCH) and low (UNMATCH) quality transcripts. As expected, confidence scores tend to be greater in high-quality transcripts. Considering these characteristics of the available transcripts, the experiments from this section were limited to transcripts produced by the Julius system.

Similarly to what was done in the experiments described in Section 6.3.3, in the experiments with adaptive context, the SQD1 queries were used as training data, for optimising the parameters of the DSI and PM functions. This included the parameters of both versions of each ranking function, those which used fixed context incidence parameters λ and σ, and those which adapted the initial values of these parameters by using the passages' uncertainty scores. The best parameter configurations were finally used to evaluate the models on the SQD2 (test) queries.

Figure 6.9: Distribution of confidence scores associated to terms from the SDPWS2 transcripts.

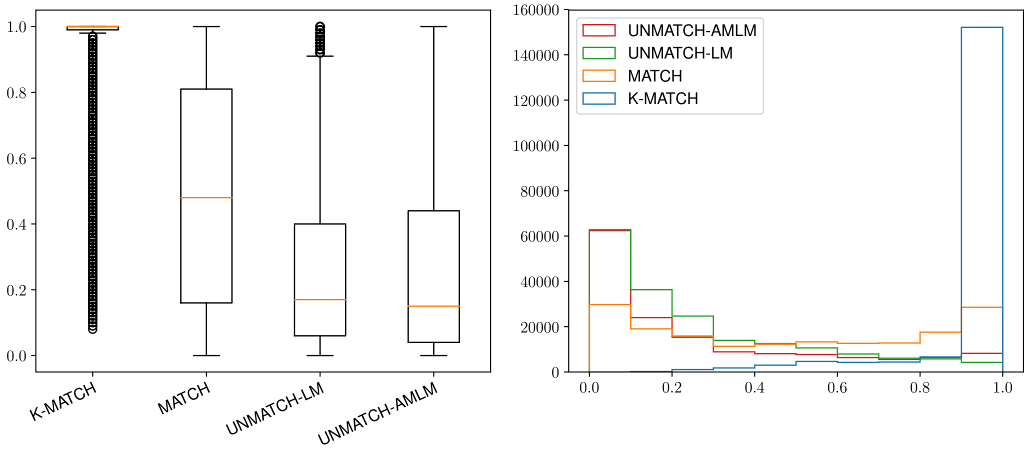


Table 6.9 shows the MAP scores obtained by the adaptive (PM+U and DSI+U) and non-adaptive (PM and DSI) contextualisation models on the SQD2 (test) queries. Values in bold mark statistically significant differences ($p < 0.05$) between PM and PM+U, or DSI and DSI+U, based on paired t-tests. The MAP differences between adaptive and non-adaptive models were in general not statistically significant, meaning that there is not a clear indication that the proposed adaptation approaches can improve the effectiveness of the PM and DIS methods. A possible explanation for this is that small variations in the context parameters may not dramatically affect the final rankings of passages. This can be partially seen in the plots from Figures 6.7 and 6.8, where large differences in MAP can only be achieved by large variations of the contextualisation parameters. However, the adaptation approaches produced highly variable incidence parameters in practice across different passages, so the above explanation may not provide a complete answer. Additional experimentation with alternative incidence parameters designed to be more sensitive to small variations in uncertainty scores, including term and occurrence specific λs in PM, plus uncertainty scores based on query terms only, produced detrimental results in retrieval effectiveness.

Despite significant differences not being found between adaptive and non-adaptive methods, the MAP scores of DSI+U tended to be consistently higher than those of DSI for most transcript combinations, while the MAP values of PM+U tended to be generally lower than those of PM. A possible reason for this effect is that the adaptive version of PM may suffer from improper frequency normalisation when using different propagation values for different passages. In this case, a passage with high transcription uncertainty would obtain a higher propagation value of λ_p and, with this, greater pseudo frequency estimates from query terms. Thus, increasing the value of λ_p for a passage produces the effect of enlarging the passage. In the design of retrieval functions, it is usually beneficial to apply length normalisation to avoid overestimating the relevance scores of long documents which are more likely to contain higher term occurrences independently of their relevance

Table 6.9: Retrieval effectiveness measured in MAP of contextualisation models that adapt context parameters according to uncertainty scores (PM+U and DIS+U) compared to that obtained with non-adaptive models (PM and DSI). These results are for the SQD2 queries with the best performing parameter settings found for the SQD1 queries, and for transcripts containing the original tokens produced by the ChaSen analyser.

RIA	**Transcripts**		**Models**				
	Query	**Doc.**	**BM25**	**PM**	**PM+U**	**DSI**	**DSI+U**
74%	M	A1	.190	.239	.235	.229	.231
59%	M	A2	.080	.166	.162	.132	.124
57%	A1	A1	.124	.177	.172	.147	.152
51%	M	A3	.107	**.192**	.154	.179	.184
48%	A1	A2	.076	.073	.071	.070	.072
40%	A2	A2	.095	.122	.103	.120	.126
39%	A1	A3	.077	.132	.147	.091	.092
34%	A2	A3	.059	.089	.082	.097	.099
33%	A3	A3	.095	.149	.144	.134	.143

status with respect to the query. For similar reasons, the proposed adaptation of PM may benefit from a more advanced normalisation technique that could consider the λ_p assigned to the passage besides its length to dampen the weight of term frequencies assigned to this passage's relevance score.

6.4 Summary

This chapter presented an initial investigation of some of the benefits that contextualisation techniques can provide to an SCR system in the task of ranking a pre-defined collection of spoken passages in order of relevance to a query. Exploiting contextual information for ranking the passages by considering the information from their container documents can be beneficial in SPR for a number of reasons.

First, by considering all possible occurrences of query terms in the document, contextualisation techniques can potentially alleviate the impact that ASR errors can have on retrieval effectiveness when the elements to be ranked are small and contain a high number of mismatches with respect to the query. Second, by disregarding the presence of strict boundaries between the passages and considering a "softened" version of the boundaries, models that employ contextualisation can be more robust to inaccurate segmentation of the spoken documents. Finally, a context-aware passage retrieval model can to some extent break with the independence assumptions that retrieval models make about the relevance status of elements contained in the same document by conditioning the score of a passage based on the score of its context.

Among all these possibilities, the experiments presented in this chapter focused on assessing whether contextualisation techniques can improve retrieval robustness to ASR errors. This was done by studying the variations of retrieval quality achieved by various contextualisation techniques in a SPR task, for different levels of noise in the query and document transcripts. Three contextualisation techniques were evaluated and compared against a well-tuned non-contextualised retrieval model: a document score interpolation (DSI), which considers global context, a positional model (PM), which emphasises local context, and their combination (DSI-PM). Results of retrieval experiments with transcripts of varying quality validate previous findings that highlight the importance of using context in element-retrieval and SCR tasks, and indicate that a combination of local and global context performs best for SCR.

Further analysis revealed that considering greater extents of local and global context can improve SCR effectiveness as ASR errors increase in the transcripts. This last observation motivated further experiments with techniques that can adapt the contextualisation incidence parameters based on the level of transcription uncertainty given by the ASR system. The results from these experiments showed that adaptive techniques did not provide any significant improvements in retrieval effectiveness over non-adaptive contextualisation techniques. Although not significant, minor differences between using adaptive and non-adaptive techniques were still observed, which motivates further investigation in this direction. More complex adaptation techniques than those explored in this work could be developed to appropriately account for length normalisation issues when passages with different uncertainty levels are contextualised with disproportionate amounts of contexts. Additionally, higher-quality uncertainty scores could be estimated based on more advanced methods for confidence score calibration, such as those described by Yu et al. (2011).

In addition to investigating the value of contextualisation techniques for robust SCR, this chapter studied some of the challenges that verbose queries and small collections pose to existing retrieval methods. In such circumstances, inverse document frequencies are poorly estimated and cannot provide an accurate account of the true discrimination power that a term has for selecting documents. The overall effect is that low content-bearing terms get assigned similarly high IDF scores than high-quality terms. These low-quality terms tend to dominate the relevance score of the documents because the number of distinct terms with these characteristics tend to be greater in verbose queries.

Two methods were explored to mitigate the issues associated with verbose queries and poorly estimated document frequencies: (i) exploiting within-query term frequencies (QF); and (ii) using exponential inverse document frequencies (IDF). Experiments showed that using these two techniques in a BM25 function provide increased retrieval effectiveness when the collection of documents is small and the queries are extremely verbose (SDPWS2 collection with SQD queries). Contrary to this, for larger spoken collections (BBC) or shorter queries (SD2), using QF and exponential IDF in BM25 does not produce improved effectiveness.

In this chapter, contextualisation techniques were shown to provide enhanced robustness to ASR errors in a passage retrieval task. While contextualising passages by considering longer excerpts of content resulted in increased retrieval robustness against ASR errors, the ability of these techniques to tackle segmentation errors in the pre-defined passages was not appropriately evaluated. The next chapter presents a large-scale evaluation of several content structuring methods, including methods based on contextualisation, for when pre-defined passages are not immediately available and need to be determined by the SCR system.

Chapter 7

Content Structuring and Evaluation in SCR

SCR from collections of long multitopical documents requires effective content structuring strategies to be applied if the amount of irrelevant material presented to the user is be minimised and their efficiency maximised. As discussed throughout Chapters 2 and 3, several content segmentation strategies have been proposed in the past with the objective of achieving this goal. The basic strategy adopted has consisted of dividing spoken documents into smaller retrieval units by using one of the text segmentation algorithms described in Section 2.3, and then to present these segments to the user as a ranking of audio snippets which commence the playback at the beginning of the retrieved speech segment.

While a significant amount of research has focused on developing new content structuring strategies, a comparable amount of effort has been devoted to the development of new evaluation methodologies and measures to appropriately quantify the "quality" of a ranked list of search results, when retrieval units are not known in advance and are instead expected to be defined by the search system. However, the difficulties associated with the evaluation of retrieval methods in these conditions, plus additional considerations that are relevant for the estimation of user satisfaction in SCR, have resulted in the creation of evaluation measures with undesirable properties. These measures tend to favour some segmentation and retrieval strategies more than others, and to overlook aspects that are important for user satisfaction.

This chapter examines the limitations of these current SCR evaluation measures when used to compare different content structuring strategies, and then describes a novel evaluation framework that seeks to quantify user satisfaction more accurately. The proposed measure is then used to evaluate a large number of content structuring methods in a SCR task in order to determine which of these is most effective and to better understand their advantages and disadvantages. The remainder of this chapter is structured as follows. Section 7.1 overviews the evaluation problem and discusses aspects that should be accounted for when designing an evaluation measure for SCR. Section 7.2 reviews existing evaluation measures proposed for SCR and related retrieval tasks, while Section 7.3 describes our proposed evaluation framework. Section 7.4 describes experiments that compare structuring methods, while Section 7.5 summarises our findings.

7.1 Evaluation of unstructured content retrieval

This section discusses different aspects of unstructured retrieval tasks, user behaviour, and user satisfaction, which are important to consider when designing evaluation measures for SCR.

7.1.1 Overview and the pool bias problem

Traditionally, IR systems have been evaluated in terms of their ability to distinguish between relevant and non-relevant documents or, as is the case of ranked retrieval, in terms of the proportion of relevant documents that are ranked on top of non-relevant ones. A ranking in which all of the relevant documents are placed on top of the list is considered the most effective and satisfactory way to present the search results to the user. Any variation of this ranking that interleaves relevant with non-relevant documents is considered sub-optimal and, consequently, less useful to the user.

The traditional approach to evaluating IR systems relies on the assumption that most users will be satisfied if presented with a set of predefined "documents" containing the information of interest. This implicitly implies that such a set of documents exists or that they can be constructed from the content available in advance of the indexing process. In other words, the assumption is that the collection is or can be structured somehow into a set of ideal documents for retrieval. A document is therefore considered an indivisible or atomic retrieval target from the system's perspective and systems are hence evaluated in terms of their ability to retrieve these basic units of information in order of relevance. In these circumstances, the standard "pooling" methodology in which a sample of potentially relevant documents is generated from multiple ranking algorithms and then submitted for manual assessment is appropriate and can be applied without major difficulties. Once a sample of relevant documents are available for a query, any of the evaluation measures that were presented in Section 2.1.3, such as MAP, nDCG, or ERR, can be used to quantify the ranking effectiveness of a system.

Compared to traditional document retrieval tasks, measuring effectiveness in unstructured content retrieval presents additional difficulties. If the "natural" documents in the collection are not suitable as retrieval units, either because they are long, multi-topical, or cumbersome to navigate through, and if in addition it is not clear how to best divide them into smaller suitable sub-units for retrieval, then the assumption of the existence of a document representing an ideal retrieval unit becomes less reasonable and so are the evaluation methods and effectiveness measures which depend upon this concept. This motivated the development of alternative pooling strategies and effectiveness measures which are more appropriate for the evaluation of unstructured content retrieval tasks in which there is not a predefined retrieval unit.

Because there is not a predefined set of document units to be retrieved from a unstructured collection, retrieval systems are left with no other option than to produce location pointers, indicating the starting and optionally ending offsets within the collection where the relevant information may be found by the user. Most pooling strategies for gathering relevance data assume that the output of different retrieval systems are samples taken from the same predefined set of documents. The union of these samples is then calculated, to remove any possible duplicates, and then each document is manually judged for relevance by a human assessor, independently from other documents. In the case of unstructured collections, this procedure cannot be immediately applied without modification. Besides the non-trivial problem of identifying near-duplicate results among ranked lists of pointers, there is also a significant increase in the difficulty of the task of assessing the relevance of content that is taken out of context, as this may not contain sufficient evidence for the assessor to provide a reliable judgement of relevance.

Researchers have adopted a wide array of alternative methods for collecting relevance assessments in these circumstances. For instance, in the cross-language speech retrieval task (CL-SR), assessors were asked to manually find regions containing information relevant to a search topic by issuing a related query to a retrieval system, and then using this to guide their decision making process when judging pointers from a pool of search results (Oard et al., 2006). Relevance assessments for the Search and Hyperlinking 2013 (SH13) task were collected in a similar fashion. As described in Section 4.1.4, annotators were asked to determine the boundaries of a relevant section to a query with the help of a SCR system. The ground truth data collected at the NTCIR SpokenDoc SD2 and SQD1 tasks, described in Section 4.2.4, was collected through a pooling procedure, but assessors were asked to refine the pointers from the pool of results in order to determine the true extent of the relevant content within a document.

While in all these examples annotators were explicitly requested to revise the pointers produced by the retrieval systems, and were given access to the full contents of the documents pointed by these results, the relevance assessment studies carried out at the SH14 and SAVA tasks employed an annotation tool which forbid assessors from further refining the pointers from the pool and restricted their access to the full contents of the documents. Restrictions like these can introduce different kinds of biases in the resulting relevance judgements, especially if the systems used for generating the results in the pool adopt similar content structuring strategies and ranking algorithms. Consider for instance Figure 4.6 which shows the distribution of lengths of passages included in the pools for the relevance assessments of the SH13, SH14, and SAVA topics. Most passages in the SAVA pool are 120 seconds length, while those from the SH13 and SH14 pools vary more widely. Biases like these in the ground truth can potentially propagate to the figures produced by evaluation measures, which will tend to favour SCR approaches that are similar to those used to produce the ground truth.

7.1.2 Representation and visualisation of search results

Retrieval from collections of long multi-topical unstructured documents requires methods that are able to determine the exact locations where the relevant content is located. The results produced in this case may take one or more forms, depending on the way these will be presented to the user. One possibility is to present results as a ranked list of document-offset pairs, indicating the ID of the document and the offset at which relevant information may be encountered within this document. An extension of this form consists of presenting both starting and ending offsets for each document, to indicate where the relevant information may span to within the document. The first result type, in which only starting points are suggested, can be referred to as "one-sided" result. Conversely, the second form in which both extremes of a suggested region or passage are specified can be referred to as "two-sided" result.

In one-sided content retrieval, search systems are designed to produce a ranked list of best entry points suggesting where a user should begin inspecting a document. Users can then be advised to inspect the results in the proposed order, by starting their search from the offsets returned for each document. One-sided evaluation measures are then those specifically designed to evaluate the quality of a ranked list of best entry points. Within this scheme, an effective retrieval system is considered to be one that assigns top ranks to pointers that are close to the onset of some span known to contain the information of interest.

Besides starting point recommendations, users may also find benefit from ending point suggestions included in two-sided results. This is because ending points can highlight regions within a document that are not worth inspecting, since they may be substantially less likely to contain any relevant information as predicted by the ranking function. By considering this extra information, users can make more informed decisions about when to stop searching when seeking for relevant content in a document. Two-sided evaluation measures are then those that quantify the usefulness of a ranked list of passages, specifying both starting and ending offsets, for finding information relevant to a query. This category of measures seeks to award systems that rank passages containing relevant information more highly than others.

Both one-sided and two-sided results can be presented to the user as a flat list of items ordered by estimates of relevance or, alternatively, as a ranked list of items grouped by document ID. In the first case, items associated with one document may interleave with others associated with a different document in the ranked list. This can cause discomfort with some users who may prefer inspecting all interesting regions of a document first before moving onto the next one. A search system could avoid this by grouping results by document. In this case, groups could be presented in order, ranked by the scores of their associated document or alternatively by some combination of their highest scoring pointers or passages. Within each document, their highest scoring items could be then presented as a ranked list or their relevance scores graphically shown as a density function superposed with the document's timeline.

Under these variants of result representation, visualisation, and presentation layouts, it is not completely clear under which variant retrieval systems would be most effectively evaluated. Ultimately, the way results are to be presented to the user, as well as the expectations about how users will interact with them, should guide the design of evaluation measures and relevance assessment studies. In the absence of information about how results will be presented in a retrieval application, retrieval results can simply be seen as an ordered list of suggested locations, which if inspected by the user in the specified order, will satisfy the information need of the user while minimising the user's effort. Related to this dilemma is the question of whether retrieval systems should be seen as tools that facilitate the location of the relevant information or that can additionally facilitate its consumption. In other words, should systems only provide pointers to where the relevant content is located or should they also include hints or transform the content somehow so that users can make better use of the relevant information for their final goal?

7.1.3 Browsing dimensions and user satisfaction

The main goal of a retrieval system is to maximise user satisfaction. In standard IR tasks, this is assumed to occur when the user can effectively find content that satisfies his/her information need by revising the search results without requiring the inspection of any piece of irrelevant information. This objective may vary slightly depending on whether the user is interested in finding all the relevant material (recall-oriented) or whether their need is satisfied with any of the relevant documents available (precision-oriented). Furthermore, when graded relevance assessments are considered, increased levels of user satisfaction are assumed to be achieved when highly relevant documents are ranked above less relevant ones.

Under these considerations, there are a number of dimensions along which a retrieval system could improve upon to increase the satisfaction of users. First, a system could reduce the amount of non-relevant material the user needs to audition by ranking relevant content at top ranks. Second, the system could increase the amount and quality of relevant material which can be effectively accessed by the user from inspecting the search results by ranking all pieces of highly relevant content at top ranks. While the first aspect relates to user effort, the second relates to gain, this is, the amount of benefit a user can obtain from navigating through the list of search results.

In retrieval tasks like passage retrieval, XML retrieval or SCR, where one of the main goals is to take the user to the exact locations where the relevant material is positioned within long documents, the "effort" dimension plays a critical role in the estimation of user satisfaction. In this case, users will be maximally satisfied if provided with an ordered list of document offsets which would permit them to detect every piece of relevant information available, without having to inspect any non-relevant material.

There are three ways in which a retrieval system could reduce the amount of non-relevant material the user will be exposed to in these tasks. The first consists of reducing the number of results pointing to documents that do not lead to any relevant material or that point to redundant material, this is, content already seen or processed by the user. The second consists of producing document offsets at locations that could facilitate the detection of any relevant content they may contain; these are locations that are close enough to the onsets of any relevant region. The third consists of ranking these high-quality entry points above lower-quality ones.

All of the above aspects focus on minimising effort along two dimensions of content browsing. The first browsing dimension, "vertical", represents navigation across the different pointer surrogates in the ranking of results, corresponding to the entries returned by the system in the search results page. The action of moving along the items in this list has some non-negligible effort associated with it. The second browsing dimension, "horizontal" relates to the process of navigating within a specific document, starting from one of the entry points suggested by the system, in the search of relevant content. This dimension presupposes an additional effort or cost on the part of the user, on top of that associated with vertical browsing.

In the design of an evaluation measure, the effort associated with vertical and horizontal browsing may depend on the peculiarities of the retrieval task and characteristics of the content. For instance, most evaluation measures designed for document retrieval consider that effort derived from horizontal browsing is negligible, and only takes into account vertical browsing effort. In tasks where retrieval results consist of pointers to text documents, horizontal browsing then acquires increased importance, and therefore evaluation measures try to account for this type of user effort. SCR perhaps presents the most extreme case, where vertical effort is arguably less substantial than horizontal browsing effort, since the cost associated with listening to audio material is higher than the cost of scrolling down through a ranked list of text snippets.

7.1.4 Browsing and navigation of multimedia content

Modern audio playback tools provide various controls which can significantly speed up the browsing of speech material compared to that of real-time listening. Standard video cassette recorder (VCR) based controls include normal playback, backward and forward seek operations, which permit the user to jump back and forth in the audio track in steps of 5, 10, and 60 seconds, speed-up controls, which permit them to increase the playback speed by up to 2x, and random access through an interactive seeker-bar, which can be used to jump into any arbitrary time-point in the audio track.

Little research has been done in the past to study how users may interact with VCR-like playback tools when faced with the task of searching for information within speech content.

An exception to this is the study described by Crockford and Agius (2006). This study involved 200 participants who were asked to find optimal entry points within a collection of 12 video clips where relevant information could be found about a particular topic. A video browsing tool was developed for this purpose, which included some of the typical VCR-like controls described above, and permitted all user interactions with the player to be recorded. One of the main findings from this study was that users tend to perform the search task faster over time as they become familiar with the contents of the collection. Another important finding was that participants employed a common set of browsing and search strategies for auditioning a video. Straight viewing or linear playback was used in 20% of cases, while random seek strategies were used less frequently. The most frequently strategy adopted by users (46% of cases) consisted of increasing the playback speed of the content. This strategy was also found to be the most effective at reducing auditioning time, providing an average of 24% time reduction compared to straight viewing. The analysis by the authors also suggests that users prefer reviewing multimedia content in a linear and sequential fashion in the direction that the media would naturally follow as opposed to browsing backwards in time.

A more recent user study described by Cobârzan and Schoeffmann (2014) investigated how users interact with modern web video players when searching for excerpts of video content. The participants in this study were given two types of known-item retrieval tasks to perform manually. Both tasks required them to use the playback, pause, and the seeker-bar controls of a typical web video player to search for a specific scene within a video within three minutes. In the first task, participants were shown the target scene and asked to re-find it after some time. In the second condition, they were shown a text description of the scene along with some relevant keyframes to enable faster visual identification. Although the experimental setup, content, as well as topics used in this study were heavily geared towards visual information, the results shed light on the navigation strategies that users tend to use for finding content and are thus relevant to the SCR case. In 60-70% of occasions, users employed a linear-search stepped strategy at the beginning of their search, consisting of switching between normal playback and forward seeks. Between 20% and 10% of users preferred commencing by seeking 30 and 60 seconds forward respectively, and only 1-3% preferred to jump onto a random location. Users also considered using straight playback in or forward seeking 80% of times, in contrast to the 20% of users who considered to seek backwards in time.

As demonstrated by these studies, in the context of SCR, VCR-like controls permit users to reduce the time they need to invest in scanning a search result, while seeking for relevant information. Most users make frequent use of the controls provided for implementing their browsing strategy. There is a clear preference for forward seeking strategies over backward seeking. These are often optimised by switching between normal playback, fixed seeks, and increase of playback speed. Because of the impact these controls may have on reducing horizontal browsing effort, they should in principle be considered in the design of evaluation measures for SCR.

Besides playback controls, other visual aids can be integrated in SCR systems in order to reduce horizontal and vertical browsing effort. A popular technique is to use thumbnails to annotate the seeker-bar with different types of metadata which are shown when the user selects a particular point in time in the bar. In video retrieval systems, thumbnails typically consists of keyframes, especially selected from the video contents so that they facilitate the identification of relevant material. In SCR applications, the seeker-bar can be annotated with a partial view of the ASR transcripts, or with a set of keywords extracted from them, and selected based on the user's query. In passage retrieval from text documents, highly scoring regions of text can be highlighted so that users can find these more quickly. These visual aids can substantially reduce horizontal browsing time, and thus should also ideally be considered in the design of evaluation measures for SCR.

7.2 Evaluation measures for unstructured content retrieval

Several variants of one-sided and two-sided effectiveness measures have been proposed in the past for evaluating the quality of a flat ranked list of location pointers or passages within unstructured documents. This section reviews a representative set of these measures and identifies some of their limitations.

The gain-discount framework

To facilitate the comparison across different families of measures, these are analysed within the gain-discount framework from Zhang et al. (2010); Carterette (2011), and Smucker and Clarke (2012), described in Section 2.1.3. Recall that this framework decomposes effectiveness measures into gain (g_k) and discount (d_k) factors. For convenience, the general formula is presented again in Equation 7.1.

$$\frac{1}{\mathcal{N}} \sum_{k=1}^{\infty} g_k \, d_k \tag{7.1}$$

The discount factor (d_k) is a monotonically non-increasing function of the ranks, which decreases every time the user inspects a new element at rank k to reflect the decrease of the user's interests in reviewing documents at lower ranks; the gain factor (g_k) represents the added benefit associated with assessing the element ranked at position k; and $\mathcal{N}$ is a normalisation factor. Note that the inverse discount d_k^{-1} can be interpreted as a measure of user effort, and that Equation 7.1 can be alternatively written as the ratio of gain to effort (Jiang and Allan, 2016). Additionally, if $\sum_k d_k = 1$, the discounting function induces a probability distribution over ranks, where each d_k corresponds to the user's continuation probability at rank k. Under these interpretations, the discount and gain factors can be explicitly related to the gain and effort dimensions of user satisfaction described in Section 7.1.3.

7.2.1 One-sided measures based on temporal distance

One-sided evaluation measures focus on the spatial/temporal distance that may exist between the entry points returned by the retrieval system and the location of the relevant information. The simplifying assumption is that this distance is representative of the amount of horizontal browsing effort that users need to invest in assessing a search result. Two important one-sided evaluation measures used in the past for quantifying SCR effectiveness are generalised average precision (gAP) and tolerance to irrelevance ($T2I$).

Generalised average precision (gAP)

Generalised average precision (gAP) was originally proposed as an extension of AP to graded relevance assessments (Kekäläinen and Järvelin, 2002). This measure was later adapted for SCR (Liu and Oard, 2006) in the context of the Cross-lingual Speech Retrieval (CL-SR) task and for text passage retrieval in the INEX Ad-hoc Best in Context task (Kamps et al., 2007). In gAP, the gain derived by the user from visiting an entry point retrieved at rank k depends on the entry point's distance with respect to the beginning of some relevant segment. Only retrieved pointers occurring within a certain minimal distance from the relevant material can result in non-zero gain, while those that are too far away result in a gain of 0. It is also assumed that users cannot derive any gain from finding relevant sections which they have already been found at previous ranks. If an entry point falls inbetween two regions of relevant material, it is assumed that the user will only reach the closest section to the point and will therefore not consume the second section, which may be visited by the user at subsequent ranks.

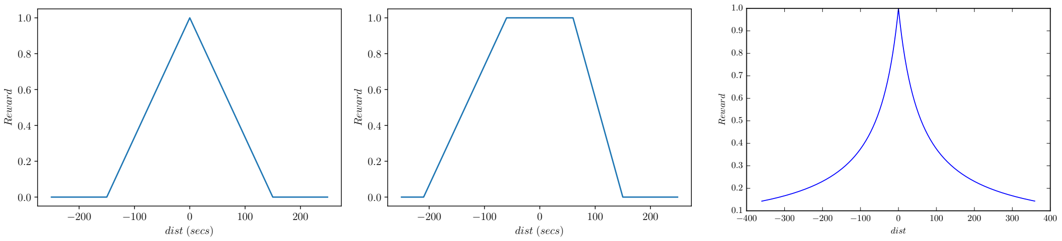
The original definition for gAP given by Liu and Oard (2006) can be instantiated under the gain-discount framework as shown in Equation 7.2,

$$\mathcal{N} = 1, \quad g_k = \frac{1}{k}\sum_{i=1}^{k} r_i, \quad r_k = \max(1 - \frac{dist_k}{10G}, 0), \quad d_k = \frac{u_k}{R}, \quad u_k = \begin{cases} 1 & \text{if } r_k > 0 \\ 0 & \text{otherwise} \end{cases} \tag{7.2}$$

where $dist_k \geq 0$ is the distance between the jump-in point retrieved at rank k and the nearest relevant onset point from the ground truth, and R denotes the total number of relevant items in the collection. If the entry point retrieved appears in a document that has no relevant content, then $dist_k$ is considered to be $10G$ and no extra gain is awarded. Similarly, $disk_k$ is $10G$ if there is another point retrieved at some previous rank $i < k$ whose closest relevant segment in the ground truth is the same as that for k.

Figure 7.1a shows a plot of the distance-based function r_k from Equation 7.2, known as the reward or penalty function. Based on simulations with this metric, Liu and Oard (2006) proposed a penalty function which reduces the credit of an entry point by 0.1 absolute for every $G = 15$ (granularity) seconds of distance shift. The G parameter thus controls the slope and "width" of the reward function. Larger values of G correspond

Figure 7.1: Reward or penalty distance function for gAP used in different evaluation campaigns. For Figures 7.1a and 7.1b, distance ($dist$) is measured in seconds, while for Figure 7.1c is in number of characters.



(a) Liu and Oard (2006). (b) Galuščáková et al. (2012). (c) Kamps et al. (2007).

to wider reward windows and the assumption that users will still derive gain from more distant entry points.

Similarly to AP, gAP can be interpreted in terms of a user model which progresses down the ranked list of pointers from top to bottom. First note that $\sum_{k=1}^{\infty} u_k = R$ if all possible jump-in points are returned in the ranked list, so that d_k defines a probability distribution over the rankings produced by the retrieval system. If these are seen as stopping probabilities, then gAP can be interpreted as the expected value of precision, weighted by some distance function r_k, if users are equally likely to stop their search at ranks with entry points that are sufficiently close to the start of some unseen relevant segment.

Alternative reward functions and limitations of gAP

Galuščáková et al. (2012) conducted a user study to determine the extent to which the triangular function proposed by Liu and Oard (2006) reflects the real tolerance of users in the context of a SCR task. Participants from this study were asked to judge the quality of arbitrary playback points, randomly generated before and after the beginning of relevant passages. Analysis of user interactions indicated that users generally spent 25% less time in identifying a relevant passage when given a point in time before the start of the passage than one after. Users tended to give up their search when given an entry point 3 to 5 minutes further away from the onset of a relevant region, and tended not to invest significantly different amounts of effort when commencing within 1 minute of the relevant material.

Based on these observations, Galuščáková et al. proposed an alternative reward function that prefers points located before rather than after the true start of the relevant content, and that equally rewards points appearing within a reasonable distance (1 minute) of the relevant start point. This improved reward function is shown in Figure 7.1b. Yet another reward function for gAP, shown in Figure 7.1c, was proposed by Kamps et al. (2007) in the context of the INEX Ad-hoc Best in Context task.

A fundamental problem with gAP is that the measure can be practically insensitive to changes in the rankings, while being extremely sensitive to changes in the quality of the entry points. To illustrate this, consider the reward function from Equation 7.2 and two rankings, A and B, generated in response to a query for which there is only one relevant passage, i.e, $R = 1$. Suppose that both rankings only contain one entry point that is within the $10G$ tolerance window from the start of the relevant passage. More specifically, suppose that A's point is ranked in position k_A and that is $9G$ away from the relevant passage, while B's is ranked in position k_B and is aligned perfectly with the beginning of the relevant passage. By replacing these values in Equation 7.2 for rankings A and B and equalling both instances of the equations, it can be shown that the gAP score for ranking A will be equal to that for ranking B if $10k_A = k_B$. Thus, system A must rank its entry point at least 10 times better than system B in order to obtain the same gAP score. If G is set to a small value, say, 10 seconds, the measure would assign equal rewards to an entry point ranked at position 5 that is within 90 seconds from the relevant content (A), and a perfect point ranked at position 50 (B).

Whether users will prefer the ranking A over B in the above example will ultimately depend on the user's ability to identify irrelevant results quickly and the amount of time that they are willing to invest in such process. If the user takes on average 10 seconds to realise that a jump-in point does not lead to any relevant content, then she would be expected to reach the relevant content after 140 seconds if using ranking A and after 495 seconds if using ranking B. Clearly, gAP overvalues entry point accuracy over ranking effectiveness. This is a consequence of the formulation of the measure, which underestimates the amount of horizontal effort that the user must invest in inspecting irrelevant results.

Tolerance to irrelevance ($T2I$)

Tolerance to Irrelevance ($T2I$) is a general model of user behaviour designed to measure the effort that a user must invest in scanning a ranked list of best entry points (De Vries et al., 2004). The model assumes that users will be willing to invest no more than t_{NR} seconds of their time in reviewing irrelevant content per entry point assessed. For this reason, the threshold t_{NR} is referred to as the user's tolerance to irrelevance. Figure 7.2 provides a graphical representation of this basic model as a finite state automaton. When browsing the k-th result, the user stays in the non-relevant state (NR) until: (a) encountering some relevant region, in which case the user moves onto the relevant state (R); or (b) until the user's tolerance expires, in which case, the user abandons horizontal browsing and proceeds to inspect the next result in the ranked list.

Under the assumptions of this abstract model, different effectiveness measures have been derived. One of these was proposed in Aly et al. (2013a) for use in the S&H tasks (Eskevich et al., 2014, 2015). Under the gain-discount framework, this measure can be instantiated by taking g_k, d_k, and u_k from Equation 7.2, and redefining r_k as 1 if $dist_k \leq t_{NR}$ and 0 otherwise. Note that this measure is equivalent to gAP if using a

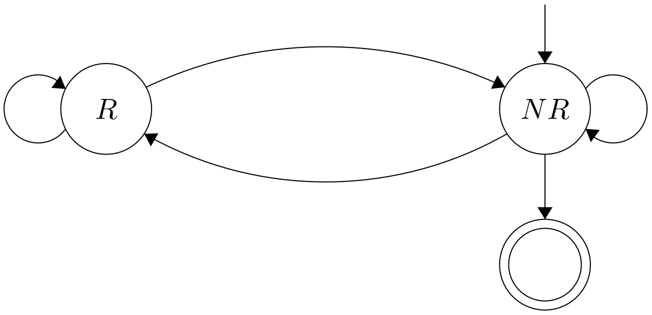


Figure 7.2: FSA for the T2I user model proposed by De Vries et al. (2004).

squared reward function with width given by t_{NR}. Unlike decaying reward functions, the squared function used in Aly et al. (2013a) implements a strict reward policy that is not realistic in terms of user satisfaction as per Galuščáková and Pecina's 2012 studies. In addition, this simple variation of $T2I$ does not remove the imbalance problem of gAP. If poor ranking quality is under-penalised and horizontal effort is not appropriately accounted for, a system may obtain an increased gAP or $T2I$ score simply by returning multiple entry point candidates within an hypothesised highly relevant region, as this would maximise the probability of "hitting" a relevant target by chance, without facing the risk of being heavily penalised by the measure.

7.2.2 Two-sided measures based on text or temporal units

Besides measuring retrieval effectiveness by estimating the quality of starting offsets, several retrieval measures have been developed to also account for the ending offsets suggested by the retrieval systems. This section reviews this type of measures, emphasising those that have been used in SCR research.

Measures based on overlap over pre-defined units

Most two-sided evaluation measures designed for unstructured retrieval are simple extensions of average precision (AP). One such extension, called overlap AP (oAP) (Aly et al., 2013a), was presented in Section 5.3.1. In oAP, retrieved results that overlap with some region deemed relevant in the ground truth are considered relevant. Precision at k is then calculated as the proportion of top k results that overlap any relevant region. These precision values are then averaged across ranks at which results are deemed relevant, and then divided by the number of relevant regions from the ground truth (R) to obtain oAP.

An obvious problem with the oAP measure is that it does not appropriately account for horizontal browsing effort. Instead, users are assumed to be equally satisfied if presented with a perfect result that captures the exact extents of a relevant region, as well as with another one that minimally overlaps with some relevant material but whose starting point is too distant to be considered useful by users. For this reason, retrieval systems can easily maximise *oAP* by returning offsets covering entire documents. Thus, while *gAP* based measures tend to favour retrieval systems that produce a large number of entry points around a putative relevant region, *oAP* will tend to favour rankings containing a lower number of results that cover long spans of content across documents.

Another set of evaluation measures for unstructured retrieval presupposes that the document collection can be divided into a set of minimal indivisible units, such as sentences or utterances, which could be used later to guide the collection of relevance assessments and the estimation of retrieval effectiveness. In particular, the set of measures used at the NTCIR SpokenDoc tasks (Akiba et al., 2011, 2013a, 2014, 2016) assume that relevance assessments and retrieval results are passages made of multiple units taken from a common set of speech utterances or inter-pausal units (IPUs). A measure of quality of a given ranked list of IPU-passages for a query can then be calculated based on the number of IPUs in the retrieved passages that are relevant and the ranks at which these are returned.

In particular, point-wise AP (*pwAP*) calculates AP by treating a retrieved passage as relevant if its middle IPU is relevant. Unlike *oAP*, this measure cannot be gamed by a system that returns long passages. Yet, *pwAP* will tend to favour methods that return a single IPU per passage, as this will increase the chances of maximising the number of relevant elements retrieved, relative to the total number of relevant IPUs available in the ground truth.

Another effectiveness measure based on a common granular segmentation is termed utterance AP (*uAP*). In *uAP*, the ranking of IPU passages is converted into a ranking of IPUs, where the first IPUs are those contained in the first best ranked passage, the second IPUs are those from the second best-ranked passage, and so on until exhausting all passages from the original ranking. *uAP* is then obtained by calculating AP over the transformed ranking by treating individual IPUs as retrieved documents.

The ranking transformation applied in *uAP* can be thought of as an operation that "flattens" the horizontal browsing dimension of the search results into a single vertical (ranking) dimension. An appealing property of this approach is that horizontal browsing effort can then implicitly be accounted for by considering the rankings at which IPUs are found. For *uAP* in particular, this is automatically regarded by the discounting factor in AP, k^{-1}, which decays asymptotically with increasing rank positions. Under this scheme, the user is then assumed to traverse the retrieved passages one by one, by examining all IPUs of one passage before moving onto the next one. In this process, the user derives gain when encountering relevant IPUs and invests a constant amount of effort per IPU.

Despite its advantages, a limitation of *uAP* is that it assumes that users will never browse content beyond the boundaries of a passage, even in the hypothetical case where a region of relevant IPUs occurs immediately after or before the boundaries of the retrieved passage. In addition, since the discounting factor is applied to within-result positions instead of ranking positions, the measure does not properly represent the conditions in which users will be shown the search results. Particularly, the discounting factor applied to a “flattenned” ranking will decrease at a higher rate with respect to ranking positions, making the measure overly sensitive to top results in the rankings.

Lastly, fractional AP (fAP) calculates an AP-like estimate based on the fraction of IPUs within a passage that are relevant (within-passage precision) and the proportion of IPUs from the relevant passage that are retrieved by the passage (within-passage recall). Within-passage precision wP and recall wR for a passage k are thus calculated as shown in Equation 7.3,

$$wP_k = \max_{r \in \mathcal{R}} \frac{|k \cap r|}{|k|} \qquad\qquad wR_k = \max_{r \in \mathcal{R}} \frac{|k \cap r|}{|r|} \tag{7.3}$$

where $\mathcal{R}$ denotes the set of all relevant IPUs and k is a set of all IPUs contained by the passage. Under these definitions, fAP can be instantiated as shown in Equation 7.4,

$$\mathcal{N} = R, \qquad\qquad g_k = wR_k \sum_{i=1}^{k} wP_i, \qquad\qquad d_k = k^{-1} \tag{7.4}$$

where R denotes the number of relevant passages from the ground truth. Similar to the other measures presented in this section, fAP does not consider the fact that users may decide to continue inspecting material beyond the boundaries of a passage and is thus prone to overpenalise near-misses. In comparison to uAP, fAP does not flatten the vertical and horizontal browsing dimensions and thus avoids issues with regard to improper discounting.

General character and time based measures

The fractional AP (fAP) measure described in the previous section can be generalised to the case when IPUs are not available. A way to achieve this is to consider units other than utterances or sentences to use as a common segmentation of the content. In the case of text documents, individual characters within a document may serve as minimal retrieval units, while for speech collections short 1-second fragments of speech could be used. A number of evaluation measures have been developed along these lines, which attempt to compute precision and recall estimates based on these more granular retrieval units for any arbitrary spans of content (Wade and Allan, 2005; Kamps et al., 2007; Eskevich et al., 2012c).

Various evaluation measures based on character precision and recall were proposed for text passage retrieval in the context of the TREC HARD tasks (Allan, 2004). An highly influential measure proposed by Wade and Allan (2005) is termed character average precision (cAP). Variations of this measure have also been used at the INEX Ad-hoc Focused task to evaluate passage retrieval effectiveness (Kamps et al., 2007). One such variation is character precision at a passage-rank cut-off k, which can be instantiated as shown in Equation 7.5,

$$\mathcal{N} = 1 \qquad g_k = \sum_{i=1}^{k} r_i \qquad d_k = \left(\sum_{i=1}^{k} l_i \right)^{-1} \tag{7.5}$$

where r_k is the length of the relevant text contained in the passage retrieved at rank k and l_k is the length of that passage. In all cases, the length of a passage is measured by the number of characters it contains. In this particular form, the discount function d_k can be thought as representing the amount of interest from the user in continuing reading as the amount of read text increases. Alternatively, if the inverse of the discount factor is considered, then character precision measures effort in terms of the amount of text that the user would need to read. If character precision is computed for a fixed k, then the measure would tend to favour methods which return a large number of short passages containing some relevant content, instead of a single passage containing them all. In fact, as shown by Wade and Allan (2005), for a ranked list of passages, the value of the character precision at k could be artificially increased for this ranking by splitting each passage by half, and forming a new ranking of passages based on these passage halves.

Character average precision (cAP) can be obtained by calculating character precision at ranks at which there is a change in gain. This can be instantiated as shown in Equation 7.6, where R denotes the total number of passages known to be relevant to the query, and r_k and l_k are defined as in Equation 7.5.

$$\mathcal{N} = R \qquad g_k = \sum_{i=1}^{k} r_i u_i \qquad d_k = \left(\sum_{i=1}^{k} l_i \right)^{-1} \qquad u_k = \begin{cases} 1 & \text{if } r_k > 0 \\ 0 & \text{otherwise} \end{cases} \tag{7.6}$$

Instead of averaging at each passage in the ranking, an alternative is to average the character precision values at each character retrieved. This can be achieved by applying a similar transformation as that of uAP to the ranking of text passages. This transformation converts the original ranking into a flat ranking of individual characters over which AP can be calculated. Instead of considering R as the total number of known relevant passages for the query, a more appropriate recall base in this case may be the total number of characters known to be relevant, that is, the sum of lengths of relevant passages from the ground truth. Note that under the probabilistic interpretation of AP, if cAP is averaged over passage rankings, it can be interpreted as the expected average precision if the user decides to stop its search at any relevant passage with uniform probability. Instead, if the average is performed over character rankings, a similar interpretation would apply to individual character positions within the returned passages. This would effectively consider the possibility of a user stopping the search in the middle of a relevant passage, which seems less intuitively reasonable.

Eskevich et al. (2012c) adapted cAP to consider temporal minimal units instead of character-level units. This family of effectiveness measures were mainly developed for the evaluation of SCR systems where horizontal browsing effort would be more appropriately quantified as auditioning time instead of read characters. Segment precision (sP) at rank k can be defined as in Equation 7.5, where r_k and l_k denote, respectively, the number of seconds of speech content in the k-th passage that are relevant and the total number of seconds of content included in passage k. Average segment precision (sAP) is then calculated over the ranking of passages, by averaging sP at every position in the ranking where a retrieved passage containing some relevant content is found. Thus, sAP can be instantiated under the gain-discount framework as shown in Equation 7.6, by taking R as the number of passages returned that contain some relevant material.

As with cAP and uAP, an alternative way to calculate sAP is to calculate standard AP over a flattened version of the original ranking of passages, by considering R in this case as the total number of seconds known relevant to the query. In this respect, Eskevich et al. (2012c) argue that averaging over passage rankings is preferable since this is more akin to the vertical browsing process that users will engage with. However, averaging over passage rankings without properly accounting for the amount of relevant material returned at each rank may in general favour retrieval methods that return shorter passages, as these are more likely to maximise sP while only covering a minimal part of a relevant region.

A common limitation of two-sided evaluation measures is that they neglect the position within the relevant region at which a returned passage may occur. As long as the levels of within-passage precision and recall remain the same, two-sided measures will assign the same amount of reward to a passage overlapping with the beginning and the ending of a relevant region. It is reasonable to think that users will prefer the first type of passages over the latter, as passages pointing to the onsets of the relevant material more closely are likely to require less browsing effort from the user. To better account for horizontal browsing effort, Eskevich et al. (2012c) proposed a two-sided measure based on sAP which penalises a system for returning entry points that are too distant from the onset of the relevant material. This measure, termed average segment distance-weighted precision ($dwsAP$), augments sAP with a distance-based reward function, similar to the one used in gAP. Because $dwsAP$ implements the same mechanism as gAP for penalising entry points, it tends to underestimate ranking effectiveness in favour of entry-point accuracy, thus favouring retrieval methods which return a large number of short passages around a potentially relevant region.

7.2.3 Browsing and interaction oriented measures

The majority of the evaluation measures proposed for evaluation of retrieval effectiveness in SCR, passage, and XML retrieval tasks, consist of relatively simple extensions of standard AP or are designed to transform the original ranking of results into an alternative representation on which traditional IR evaluation measures can be applied. A recent trend in IR research is to move away from these system-oriented measures towards user-centric frameworks, based on models of browsing behaviour, which can better account for additional factors that affect user satisfaction. In this paradigm, the design of evaluation measures is grounded in hypotheses and observations about how users may interact and navigate through a list of search results. Much of this effort has been devoted to providing new interpretations of traditional IR measures under the view of different models of user behaviour (Zhang et al., 2010; Moffat and Zobel, 2008; Carterette, 2011; Smucker and Clarke, 2012; Jiang and Allan, 2016). At the same time, novel effectiveness measures have been proposed from this perspective on evaluation of search systems. This section reviews some of these measures.

Trail-texts and the U-measure

Sakai and Dou (2013) developed an evaluation framework that can be used to obtain estimations of user satisfaction across a large range of IR tasks, where the results to be evaluated can be comprised of any arbitrary piece of text, from full documents, to multi-document summaries or aggregated infoboxes. The basic idea is to construct a "trail-text" based on all text that has been read by the user during a search session. A trail-text thus contains the text accessed by the user while interacting with the retrieval system, that is, the concatenation of all text read in the order these were inspected by the user.

Ideally, trail-texts must be obtained by recording the actions of one or more users while interacting with the retrieval system under evaluation, either through an eye tracking device or by analysing click logs. The advantage of considering multiple trail-texts instead of one is that this permits to estimate how the performance of a system may vary across users. In the absence of real user data, Sakai and Dou (2013) suggest carrying out user simulations to obtain a set of simulated trail-texts. These simulations can be either based on deterministic or probabilistic models of user behaviour (Carterette et al., 2011). In the work from (Sakai and Dou, 2013), the authors generated trail-texts deterministically, by concatenating the text of all search results in a result list in their ranking-induced order.

In the process of generating a trail-text from a ranked list of search results, Sakai and Dou applied different rules to the individual results from the list depending on whether or not they could lead the user to some relevant information. In particular, results pointing to relevant content were appended to the trail-text preceded by the text of its associated snippet, while those not pointing to any relevant information were discarded and only their snippets appended to the trail-text. Note that, in practice, this procedure for creating trail-texts is similar to the ranking transformation described in Section 7.2.2 for two-sided measures, whereby the ranking of elements is flattened into a linear vertical ranking of minimal information units. The difference in this case is purely conceptual: trails-texts can be seen as the path that a user decides to follow while interacting with the retrieval system, rather than the stream of information that the search system presents to the user.

Different evaluation measures can then be defined over the contents of a trail-text in order to measure various aspects of user satisfaction. For instance, the U-measure proposed by Sakai and Dou (2013) is defined under the gain-discount framework. In this measure, gain and decay are functions of character positions within a trail-text. The gain factor, g_k, for the k-th position within the trail-text is considered as 0 if such a position belongs to a passage that is not deemed relevant. Otherwise g_k acquires a fixed value of v_l that depends on the length (l) of the relevant passage associated with the k-th position. The discounting factor d_k adopted in the U-measure consists of the linear decay function shown in Equation 7.7,

$$d_k = \max(1 - \frac{k}{L}, 0) \tag{7.7}$$

where L represents the maximum number of characters of text the user is willing to read in a search session. Note that this function decays linearly as the user navigates the individual characters of the trail-text.

Time-traces and improved SCR measures

The counter-part of a trail-text in a temporal medium, such as a video or speech recording, can be termed a "time-trace" (Clarke and Smucker, 2014). In the context of an SCR application, a time-trace can be constructed by appending the fragments of speech content the user has listened to while navigating through a ranked list of SCR results. Similarly to the text retrieval scenario, a time-trace can be generated by either recording real users interacting with a SCR system, or by running user simulations.

The majority of the evaluation measures presented in Sections 7.2.1 and 7.2.2 can, in fact, be interpreted as assuming one particular deterministic type of user behaviour, which results in a specific form of time-trace. For instance, gAP can be thought as a time-trace creation process that appends the non-relevant "gap" existing between a retrieved entry point and the beginning of a relevant region to the time-trace, followed by the relevant content that comes next. The measures oAP, uAP, fAP, and sAP can be thought of as constructing a time-trace by flattening the passages in the ranked list, in a similar procedure to that adopted in the U-measure. Once a user behaviour model is assumed and a time-trace created based on it, these effectiveness measures frequently compute AP at specific positions within the time-trace.

Within this user-oriented paradigm, existing evaluation measures can be adapted based on different expectations about how users may produce a time-trace from a given rank list of SCR results. One possible modification that can be made to the majority of two-sided measures listed above is to model the fact that, once found, users will be willing to continue listening to the relevant material, as well as to any extra adjacent relevant span occurring within a certain range from the end of this relevant material. Thus, instead of considering that users will only inspect content that lies within the boundaries of a retrieved passage, effectiveness measures could also be designed to account for the possible actions that users may take while seeking for relevant material given a starting and ending offset as a suggested region for inspection.

A possible deterministic model of user behaviour that can capture these characteristics may be defined as follows. A user selects a search result and starts listening to the audio material from the entry point suggested by the SCR system. If the user finds any relevant information before reaching the ending point proposed by the system, then the user is assumed to continue listening to the audio until the ending of such relevant material, thus deriving gain from every second of relevant speech found. In cases where the entry point is already contained within a relevant region, the user is then assumed to seek backwards in time until reaching the beginning of such relevant material, and subsequently start deriving gain from that point moving forward. If the region suggested by the system does not lead the user to any relevant content, then the user is assumed to invest an amount of time proportional to the length of the complete audio passage suggested, without acquiring any additional gain. As an evaluation measure inspired by sAP, we proposed average interpolated segment precision ($AiSP$) (Racca and Jones, 2015a), based on this deterministic, albeit more realistic, user model of browsing behaviour. This measure was put into practise at the MediaEval SAVA (Eskevich et al., 2015) and in the TRECVid Hyperlinking (Awad et al., 2016, 2017) benchmarks.

Time-biased gain (TBG) and time well spent (TWS)

Time-biased gain (TBG) and time-well spent (TWS) are a family of effectiveness measures proposed in (Smucker and Clarke, 2012; Clarke and Smucker, 2014), which seek to account for the time a user needs to invest in identifying, assessing, and extracting relevant material from a ranked list of search results produced by an IR system. The key observation made by the authors is that most traditional evaluation measures for IR are unrealistic, in the sense that they assume users will traverse a ranked list from top to bottom, spending an equal amount of time per result, when in reality, the time required to derive gain from every document varies according to several factors. The factors identified that may influence the time a user spends on a retrieved document include the length of the document, the quality of its surrogate, its relevance status, or whether the document is a near-duplicate of some other document already seen by the user. The time-biased gain (TBG) framework attempts to account for these additional aspects in the estimation of retrieval effectiveness. Instead of incorporating these factors as additional set of rules to shape the value of a traditional effectiveness measure, Smucker and Clarke (2012) proposed to measure their impact in terms of the extra time they would add to the information seeking process carried out by the user.

TBG is defined in terms of the gain-discount framework. The cumulative gain derived by the user after having invested an amount of time equal to t in inspecting a number of search results is denoted by $G(t)$. In a similar fashion, the discount factor, $D(t)$, is defined in terms of time and denotes the probability that the user will continue seeking for information after having invested t units of time. In situations where gain cannot be determined as a continuous function of time, for instance in document retrieval tasks with binary judgements, time-biased gain is defined as shown in Equation 7.8,

$$\frac{1}{\mathcal{N}} \sum_{k=1}^{\infty} g_k \, D(T(k)) \tag{7.8}$$

where k ranges across document ranks as opposed to time, g_k is the gain associated with the k-th result in the ranking, and $T(k)$ the expected time at which the user will reach the k-th document.

In order to determine an appropriate value for $T(k)$, Smucker and Clarke (2012) conducted a user study to gather data from real user interactions with a document retrieval system. Three variables were then used to approximate $T(k)$: (i) the time a user spends revising a text surrogate; (ii) that of assessing a document of length l_i; and (iii) the probability that the user will click on a document given its relevance status (c_i). Analysis from the interaction data resulted in the values shown in Equation 7.9 for the calibration of $T(k)$ as a function of the rank k, measured in seconds.

$$T(k) = \sum_{i=1}^{k-1} 4.4 + (0.018 \, l_i + 7.8) \, c_i \tag{7.9}$$

As for the remainder of the components of TBG from Equation 7.9, Smucker and Clarke set the normalisation factor to $\mathcal{N} = 1$, and the discount factor to an exponential decay function of time, proportional to $D(T(k)) = \exp(-T(k))$. The selection of an exponential decay function for modelling user continuation probabilities has been supported by multiple independent studies based on the analysis of query logs from commercial search engines (Moffat and Zobel, 2008; Smucker and Clarke, 2012).

In follow-up work, Clarke and Smucker (2014) developed the time well spent (TWS) measure. Similar to TBG, TWS calculates gain and decay with respect to the time the user has to invest in looking for relevant information in a ranked list of results. For modelling decay, the authors used a log-normal probability distribution, with parameters estimated based on query log data. As opposed to TBG, the gain function $G(t)$ of TWS is parametrised by time instead of rank positions. Gain is then based on the *time well spent* of the user, calculated as the ratio between the time the user spent on relevant content versus the total time invested by the user in the search session. In other words, gain is defined as a benefit-effort ratio, where benefit is measured as time spent on consuming relevant material and effort as the total time spent while interacting with both relevant and irrelevant content.

In practice, computing gain in TWS for a ranked list of results requires the adoption of a certain model of user behaviour that can be used to generate a time-trace of consumed material. The cumulative gain $G(t)$ can then be calculated for a time-trace up to time t, by summing the number of seconds associated with relevant material in the trace up to t divided by t. In the absence of real user interaction data, a time-trace may be generated under the assumption of a deterministic model of user behaviour, such as those described for the U-measure and the $AiSP$ measure, or based on stochastic or probabilistic user simulations (Carterette et al., 2011; Clarke and Smucker, 2014). In their development of TWS, Clarke and Smucker propose using stochastic simulations.

In general terms, a simulated user can be described by parameters θ drawn from a distribution over a parameter space Θ. One such parameter in θ, denoted by t_{max}, may correspond to the maximum amount of time a user is willing to invest in the search. Given a sample value of t_{max}, a user interaction of this length can then be simulated by generating a trace of the user's activities while interacting with the ranked list of results under evaluation. This procedure can be repeated m times for m different user models sampled from Θ to generate m time-traces upon which TWS can be calculated. Clarke and Smucker (2014) adopt this procedure to obtain a distribution of TWS values for a ranked list of results, which they then use to measure "effect sizes" and user variance in TWS. A clear advantage over considering only one user model is that multiple models permit measuring most of the impact that different user behaviours and browsing strategies may produce over the output of an effectiveness measure. Instead, evaluation measures based on a single model of browsing behaviour are exposed to different types of biases which may be introduced by the underlying assumptions made by each model.

7.3 A new user-centric evaluation framework for SCR

This section describes the development of a novel user-centric framework for evaluating SCR results in unstructured collections that tackles many of the limitations of current evaluation measures. Unlike the majority of measures used in the past which can only capture a single dimension of browsing effort, the framework described in this section can account for both vertical and horizontal efforts in an effective and balanced manner, without over estimating the importance that one dimension has over the other. In addition, the proposed framework is based on an explicit probabilistic model of horizontal browsing and navigation of speech, which permits consideration of the effects of playback VCR-like controls and of different presentation layouts more easily within the estimation of retrieval effectiveness. This section begins by describing the horizontal browsing model, and then moves onto augmenting the model to capture vertical browsing behaviour.

7.3.1 Horizontal browsing model

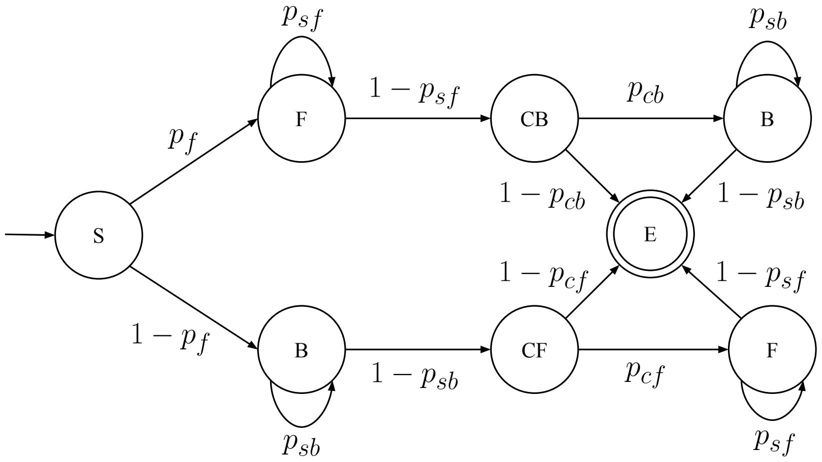
When users selects an item from the ranked list produced by a SCR system, they are provided with a playback interface with which they can start playing the audio track from the point in time suggested by the system. At this stage, the main objective of users is to determine whether any relevant material can be found by navigating the audio from this starting point. It is assumed that users will only derive gain from assessing a search result if the listen-in point suggested by the system leads them to some relevant content, and if they are capable of locating such material within the document within a reasonable amount of time. If the SCR system also produces a suggested ending point, then users may exploit this information if it is found to be useful. Although, in order to keep the development of this browsing model simple, we assume one-sided search results from now on and only assess the quality of the entry points retrieved by a hypothetical SCR system.

In order to accelerate the horizontal searching process, users can avail of the more advanced playback controls offered in most VCR-like interfaces. As discussed in Section 7.1.4, the most common browsing strategy adopted by users consists of a combination of straight playback and forward-seek operations. Alternatively, users may begin their search backwards in time, by moving the seeker bar a few seconds back and commence their playback from that point on. If this initial exploration does not lead users to any relevant material, they may continue their search by exploring locations they had not examined before around the recommended entry point, either in the forward or backward directions. Users are assumed to continue with this information seeking process until: (i) they encounter some relevant material; or (ii) they decide to abandon the search as the material in this document is not worth their time. Note that a user may reach state (ii) even if some relevant material could have been reached if they had decided to invest additional time in the search.

Most of these requisites can be accommodated within a simple probabilistic model that disentangles the information seeking process just described into a sequence of states. Figure 7.3 shows the proposed model as a probabilistic finite state automaton with states $\mathcal{Q} = \{S, F, B, CB, CF, E\}$ and state-transitions given by probabilities $p_f, p_{sf}, p_{sb}, p_{cf}$, and p_{cb}. The state S denotes the starting state and represents the point at which a user must choose between a forward (F) or a backward (B) seeking strategy to commence the search. The user then moves on to the forward state F with probability p_f, or on to the backward state B with probability $1 - p_f$. While in state F, the user is assumed to have started the playback of the audio at normal speed in order to begin listening to the material. The user is then assumed to listen to the next i time units of audio with probability p_{sf}^i. This is represented in Figure 7.3 by the self transition in the F state with the p_{sf} label. After listening to i time units, the user may decide to stop seeking forward with probability $1 - p_{sf}$, and move onto state CB. This state stands for "backward continuation" and represents the possibility that a user may want to continue searching for relevant content by exploring the content located before the entry point, in the backwards direction, after failing to find any relevant material in the forward direction. From the CB state, the user can then move onto the backward state B with probability p_{cb} or abandon the search and move on to state E with complementary probability. States B and CF are analogous to states F and CB for when the user commences seeking for relevant information in the backward direction and optionally continues searching in the forward direction.

Under this model of browsing behaviour, an individual user can be characterised by a vector of probabilities $u = (p_f, p_{sf}, p_{sb}, p_{cf}, p_{cb})$. A different selection of any of these

Figure 7.3: Proposed model of horizontal browsing. The states S and E are starting and ending states respectively, while F, B, CB, and CF correspond to the forward browsing, backward browsing, backward continuation, and forward continuation respectively.



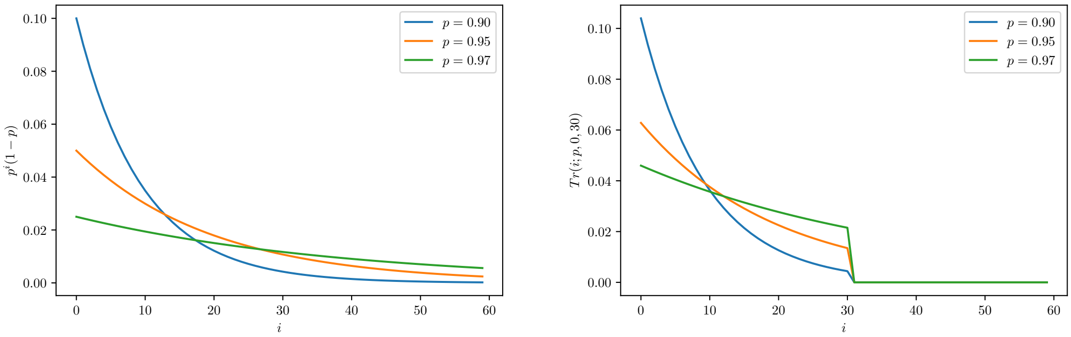
parameters would correspond to a different user model. For instance, $p_f = 1.0$, $p_{sf} = 0.99$, and $p_{cb} = 0.0$ results in a user who would never search backwards. Ideally, these probabilities could be learned from analysing real user interactions or by mining search logs. In the absence of this information, one can opt for user simulations by synthesising a number of independent users as done by Carterette et al. (2011), and Clarke and Smucker (2014).

Forward and backward browsing

The forward and backward seeking states F and B represent the "core" components of this user model. Note that each of these states induces an independent searching process, where the number of time units the user consumes before stopping, I, is a random variable assumed to follow a geometric distribution with continuation or persistence probability given by p_{sf} and p_{sb} for states F and B respectively. The election of a geometric distribution to model horizontal browsing is inspired by the model underlying rank-biased precision (RBP) (Moffat and Zobel, 2008), described in Section 2.1.3. Recall that RBP adopts a discount factor equal to the probability density function (PDF) of a geometric distribution $d_k = p^{k-1}(1 - p)$, where p acts as the persistence or continuation probability of the user at a particular rank position $k > 0$. The user model underlying states F and B is similar to that of RBP, presented in Figure 2.1, although adapted to model horizontal browsing in our proposed framework.

An important difference between RBP and the states F and B in our user model is that RBP assumes that there are an infinite number of documents in the ranking which the user could explore. Thus, in RBP, the probability mass of the geometric distribution spreads over an infinite number of ranking positions, starting from rank $k = 1$ up to infinity. Although rankings are always finite in practice, this assumption is still reasonable for

Figure 7.4: Probability density functions of geometric and truncated ([0, 30]) geometric distributions for various values of p.



(a) Geometric distribution.

(b) Truncated geometric distribution.

web search tasks, where document collections are extremely large. Because documents are of finite length, the assumption made in RBP is less reasonable for the modelling of horizontal browsing. This is because users will likely have a limited amount of content to inspect within an individual document. Furthermore, users may be given an entry point that is located at the far right (left) end of a document, so that there is no extra content for them to inspect after (before) such point. In these circumstances users can save some time by stopping their search early, and this should therefore be then taken into account in our user model.

To properly account for these factors in the F and B states, the browsing process is modelled with a truncated geometric distribution which spreads the total probability mass across a closed interval $[a, b]$. The PDF of a truncated geometric distribution for the interval $[a, b]$ with support $i \geq 0$ can be written as shown in Equation 7.10.

$$P(I = i) = Tr(i; p, a, b) = \begin{cases} \frac{p^i(1-p)}{p^a - p^{b+1}} & \text{if } a \leq i \leq b \\ 0 & \text{otherwise} \end{cases} \tag{7.10}$$

Figures 7.4a and 7.4b show respectively the PDFs of a geometric distribution and its truncated counterpart for various values of the continuation probability p. Truncating the distribution at $b = 30$ is more appropriate if the entry point from which the user begins inspecting the document lies within 30 units of the end of such a document.

Deriving gain from horizontal browsing

Under the proposed model of horizontal browsing shown in Figure 7.3, one can calculate probabilities for different events of interest. One such event is that of a user finding the onset of some relevant region within a document, given an entry point suggested by a SCR system. In this respect, it is assumed that a user is satisfied if encountering the starting point of a relevant region, instead of just its middle or ending parts. Our evaluation framework then defines the gain a user may obtain by inspecting an entry point based on the probability they will actually find any relevant information from the examination of this entry point. Thus, instead of considering gain as something that can be either acquired or not acquired by the user, our framework measures potential gain probabilistically. Users are then more or less likely to find relevant content depending on their willingness to spend time browsing irrelevant material plus the quality of the entry point suggested by the system.

Let the entry point under evaluation be located at a central position 0 and the beginning of some relevant region, if any, located at some distance $r \geq 0$ from the entry point. An additional constraint for r is that it must be a valid offset within the document, that is, it cannot go beyond the boundaries of the document. This can be specified as $r \leq n$ with n being the ending offset of the document relative to the entry point. For instance, Figure 7.4b shows an example where the entry point is located within $n = 30$ time units from the ending of the document. According to our horizontal browsing model, the probability of a user finding r while being at state F by starting from the entry point suggested by the system, can be calculated based on the number of time units $I_F = i$ the user must listen to before reaching position r. Since I_F is geometrically distributed with probability $p = p_{sf}$, the probability that the user will reach position r can be calculated as shown in Equation 7.11.

$$\begin{aligned}
PFF(r) &= P(I_F \geq r) \\
&= \sum_{i=r}^{n} P(I_F = i) \\
&= \sum_{i=r}^{n} Tr(i; p_{sf}, 0, n) \\
&= \frac{1 - p_{sf}}{1 - p_{sf}^{n+1}} \sum_{i=r}^{n} p_{sf}^{i} \\
&= \frac{(1 - p_{sf})}{(1 - p_{sf}^{n+1})} \frac{(p_{sf}^{r} - p_{sf}^{n+1})}{(1 - p_{sf})} \\
&= \frac{p_{sf}^{r} - p_{sf}^{n+1}}{1 - p_{sf}^{n+1}} \qquad (7.11)
\end{aligned}$$

Thus, the probability of a user finding r from state F while browsing forwards in time, $PFF(r)$, is the sum of the probabilities corresponding to all possible events in which a user listens to at least the first r time units of speech material before transitioning onto state CB. For example, for the case of Figure 7.4b, $PFF(20)$ gives 0.087, 0.194, and 0.269 for p_{sf} equal to 0.90, 0.95, 0.975 respectively. Thus, users with higher persistence probability will have greater chances of reaching position r in the document and of successfully locating the onset of the relevant material. Under this definition of $PFF(r)$, the amount $1 - PFF(r)$ denotes the probability of the user not finding r while seeking forward. Additionally, for relevant onsets preceding the entry point, $r \leq 0$, it is assumed that $PFF(r) = 0$.

The analogous of $PFF(r)$ with $r \geq 0$ for when users commence their search by browsing backwards, from state B, is denoted as $PFB(r')$ with $r' \leq 0$, that is, with the relevant onset point having negative offset relative to the entry point. This probability can be calculated as shown in Equation 7.11, by taking $r = -r'$ and replacing p_{sf} by p_{sb}, and n by the number of positions that exist between the entry point and the beginning of the document. If the document does not contains any relevant information, r is taken to be infinite so that $PFF(r) = PFB(r') = 0$. Similarly to the forward browsing case, the value $1 - PFB(r')$ denotes the probability of the user not finding the relevant onset $r' \leq 0$ while seeking backwards. Also, if the relevant onset r' is located after the entry point so that $r' > 0$, then $PFB(r') = 0$.

So far, the probabilities of finding relevant content for when the user is in states F or B have been calculated in isolation from the rest of our horizontal browsing model. In order to calculate the probability of a user finding r by considering the complete model from Figure 7.3, the remainder of states and transition probabilities need to be taken into account. Note that, according to our model, a user may miss some relevant information contained in the document during the forward (backward) seeking but may still be able to find such a region if they move on to state CB (CF) and then continue seeking in the opposite direction.

The probability of a user finding a relevant onset r located after the entry point ($r \geq 0$) under the full model, $PF_a(r)$, can be calculated by summing the probabilities associated with all paths in the graph which contain at least r transitions from state F to itself. It can be shown that such a probability can be calculated as shown in Equation 7.12.

$$\begin{aligned} PF_a(r) &= p_f \; P(\text{finding } r \text{ from } F) + (1 - p_f) \; P(\text{finding } r \text{ from } B) \\ &= p_f \, PFF(r) + (1 - p_f) \, p_{cf} \, PFF(r) \\ &= PFF(r) \, [p_f + (1 - p_f) \, p_{cf}] \end{aligned} \tag{7.12}$$

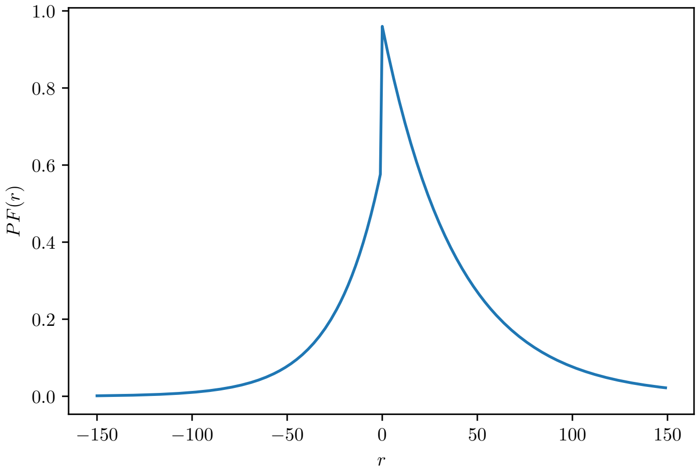
Analogously, the probability of finding a relevant point $r \leq 0$ appearing before the entry point under the full model is given by $PF_b(r) = PFB(r) \, [(1 - p_f) + p_f \, p_{cb}]$.

The probability of a user finding any relevant onset r from the start state S of the model, can then be expressed in terms of PF_a and PF_b as shown in Equation 7.13.

$$PF(r) = \begin{cases} PF_a(r) & \text{if } r \geq 0 \\ PF_b(-r) & \text{otherwise} \end{cases} \tag{7.13}$$

With the above definitions, the gain a user obtains from assessing an individual retrieval result in our evaluation framework is given by $PF(r)$, that is, the probability that the user will locate some relevant material by following the entry point suggested by the SCR system. Defined this way, gain has a direct interpretation: a gain equal to 0.60 for a

Figure 7.5: Probability of finding the onset of some relevant content r ($PF(r)$) if starting from an entry point located at 0 for an hypothetical user.



particular user, entry point, and document containing some relevant material, signifies that this user has a 60% chance of finding the onsets of such relevant content. Note that under this framework, users will always derive some gain from an entry point if the document pointed to by this entry point contains at least one relevant region. Such gain will be high if the entry point is likely to help the user to locate the relevant information and low otherwise. Figure 7.5 shows a plot of $PF(r)$ for an hypothetical user with $p_f = 0.8$, $p_{sf} = 0.975$, $p_{sb} = 0.96$, $p_{cf} = 0.8$, and $p_{cb} = 0.5$. This user would attain a gain close to 1.0 if the entry point returned by the system is within a couple of time units of a relevant region. Such a user would still be able to find the onset of an hypothetical relevant content if this is located before the entry point, but with lower probability than if located after.

Estimating horizontal browsing effort

The previous section discussed how gain can be calculated for a particular entry point and document according to our user model of horizontal browsing. For this definition of gain, highly persistent users, who are willing to spend large amounts of time browsing the contents of a document, are more likely to obtain a greater probability of gain than less patient users. In fact, a user with persistence probabilities equal to 1.0 will always be able to find any relevant content that a document may contain, independently of how far this content may be from the entry point where the user starts browsing. In these cases, users with maximal persistence will always locate a relevant region with probability equal to 1.0. Within our probabilistic model, the difference between persistent and impatient users lies in the amount of effort they are willing to invest in inspecting a search result. Extremely persistent users will always be able to find every single instance of relevant information at the cost of increased effort, while less patient users will make more effective use of their time by investing the least amount of effort possible to accrue the maximum amount of gain they can.

In order to account for both effort and gain, the following describes how horizontal effort can be estimated within our probabilistic model. As with gain, effort is estimated in a probabilistic manner, as the amount of effort a user is expected to invest while inspecting a search result. As in the TBG framework, our model represents effort based on the number of time units of content a user is expected to examine during horizontal browsing. Let I be a random variable representing the number of time units a user listens to before finishing the search. Expected effort can be calculated as the expected value that I will take under the constraints imposed by our model and the characteristics of the user. For a given document with boundaries $[a, b]$, where $a \leq 0$ and $b \geq 0$ given as relative offsets to some entry point produced by an SCR system, the expected effort can be calculated as the expectation $E[I]$ shown in Equation 7.14,

$$E[I] = \sum_{i=0}^{b-a} i \, P(I = i) \tag{7.14}$$

where $I = i$ denotes the event that a user consumes exactly i time units throughout the entire search process. Note that the summation in Equation 7.14 runs from $i = 0$ up to the maximum number of time units a user could consume in a document $(b - a)$. Conveniently, the event $I = i$ can be expressed in terms of the event I_F, representing a user who consumes exactly i units when starting from state F, and I_B for the analogous case when the user starts from state B. Based on these two events, the expectation $E[I]$ bounded to $[a, b]$ can be rewritten as in Equation 7.15.

$$\begin{aligned} E[I] &= p_f \, E[I_F] \; + \; (1 - p_f) \, E[I_B] \\ &= p_f \sum_{i=0}^{b-a} i \, P(I_F = i) + (1 - p_f) \sum_{i=0}^{b-a} i \, P(I_B = i) \end{aligned} \tag{7.15}$$

From within state F, users can consume $I_F = i$ time units by following two possible paths in the state transition system from Figure 7.3. For a hypothetical document with infinite boundaries, users can consume $j \leq i$ units browsing forward and then move on to the backward browsing state and consume $i - j$ units backwards. Analogously, if starting from state B, users can consume j units backwards first, and then the remaining $i - j$ units forward. In both cases, the total units consumed are $j + (i - j) = i$. The probability $P(I_F = i)$ for an unbounded document can be then obtained by summing the probabilities of all possible paths that connect states F with E, in which the total sum of units consumed at states F or B is exactly equal to $I_F = i$. The terms in this summation can be re-arranged into Equation 7.16,

$$\begin{aligned} P(I_{_F} = i) = {} & p_{sf}^{i}\,(1 - p_{sf})\,(1 - p_{cb}) \\ & + \left[\sum_{j=0}^{i} p_{sf}^{j}\, p_{sb}^{i-j} \right] (1 - p_{sf})\,(1 - p_{sb})\, p_{cb} \end{aligned} \tag{7.16}$$

which expresses $P(I_{_F} = i)$ as two main terms, one corresponding to paths that transition from CB to E (no backward continuation), and another term for paths from CB to E that pass through state B. While the left term in the equation captures cases where a user consumes i units by always browsing forward, the right term does this for cases where a user consumes j units forward and $i - j$ units backwards.

For a bounded document, users can only consume content up to the boundaries of the document. In particular, for a document with boundaries $[a, b]$, with $a \leq 0$ and $b \geq 0$ expressed as relative positions with respect to an entry point located at the origin 0, users can consume at most a and b units in the backward and forward directions respectively. Despite being reasonably simple, Equation 7.16 does not properly reflect $P(I_{_F} = i)$ for a bounded document, since many of the terms considered in the summation of Equation 7.16 are for values of i and $i - j$ which may be greater than b and a respectively. Although adapting Equation 7.16 to comply with the bounded condition would result in a complicated expression, when $P(I_{_F} = i)$ is considered in the context of the expectation $E[I_{_F}]$, terms can be re-arranged in a way that the summation in the right term of Equation 7.16 for the bounded condition can be expressed with the double bounded sum shown in Equation 7.17.

$$\begin{aligned} \sum_{i=0}^{b-a} \sum_{j=0}^{i} p_{sf}^{j}\, p_{sb}^{i-j} - \{\text{invalid terms}\} &= \left[\sum_{j=0}^{b} p_{sf}^{j} \right] \left[\sum_{l=0}^{-a} ps_{sb}^{l} \right] \\ &= \frac{(1 - p_{sf}^{b+1})}{(1 - p_{sf})} \, \frac{(1 - p_{sb}^{-a+1})}{(1 - p_{sb})} \end{aligned} \tag{7.17}$$

Applying a similar term arrangement procedure over this summation when it is multiplied by i, gives Equation 7.18,

$$\begin{aligned}
&\sum_{i=0}^{b-a}\sum_{j=0}^{i} i\, p_{sf}^{j}\, p_{sb}^{i-j} - \{\text{invalid terms}\} &(7.18)\\
&\quad = \sum_{j=0}^{b} p_{sf}^{j} \left[\sum_{l=0}^{-a} (l+j)\, ps_{sb}^{l}\right]\\
&\quad = \sum_{j=0}^{b} p_{sf}^{j} \left[\sum_{l=0}^{-a} l\, ps_{sb}^{l} \; + \; j \sum_{l=0}^{-a} ps_{sb}^{l}\right]\\
&\quad = \left[\sum_{j=0}^{b} p_{sf}^{j}\right] \left[\sum_{l=0}^{-a} l\, ps_{sb}^{l}\right] \; + \; \left[\sum_{j=0}^{b} j\, p_{sf}^{j}\right] \left[\sum_{l=0}^{-a} ps_{sb}^{l}\right]\\
&\quad = SUM
\end{aligned}$$

where each of the individual geometric series appearing in the last step of Equation 7.18 can be reduced according to Equations 7.19 and 7.20.

$$\sum_{i=0}^{n} p^{i} = \frac{1-p^{n+1}}{1-p} \tag{7.19}$$

$$\sum_{i=0}^{n} i\, p^{i} = \frac{p(1-p^{n+1})-(n+1)p^{n+1}(1-p)}{(1-p)^2} \tag{7.20}$$

The bounded summation from Equation 7.18 as well as Equations 7.19 and 7.20, can then be used to obtain an expression for $E[I_{F}]$ for the bounded case. This is shown in Equation 7.21,

$$\begin{aligned}
E[I_F] &= \sum_{i=0}^{b-a} i\, P(I_F = i) &(7.21)\\
&= \left[\sum_{i=0}^{b} i\, p_{sf}^{i}\right] \frac{(1-p_{sf})}{(1-p_{sf}^{b+1})}\,(1-p_{cb})\\
&\qquad + SUM \frac{(1-p_{sf})}{(1-p_{sf}^{b+1})}\,\frac{(1-p_{sb})}{(1-p_{sb}^{1-a})}\,p_{cb}\\
&= \; T_1 + T_2\, p_{cb}
\end{aligned}$$

where

$$T_1 = \frac{p_{sf}(1-p_{sf}^{b+1}) - (b+1)p_{sf}^{b+1}(1-p_{sf})}{(1-p_{sf})(1-p_{sf}^{b+1})}$$

$$T_2 = \frac{p_{sb}(1-p_{sb}^{1-a}) - (1-a)p_{sb}^{1-a}(1-p_{sb})}{(1-p_{sb})(1-p_{sb}^{1-a})}$$

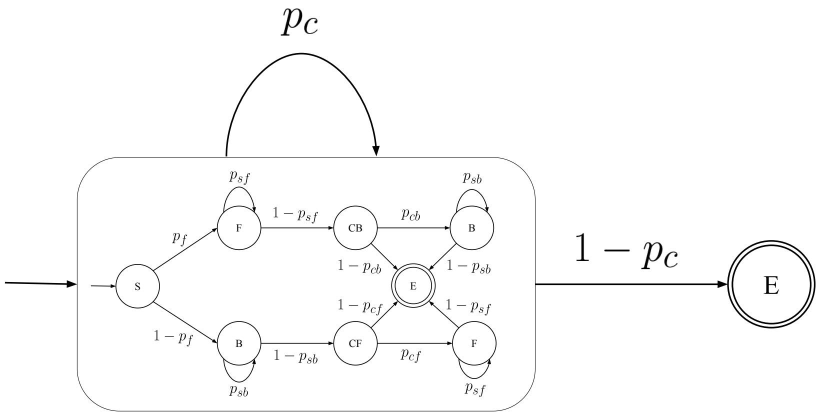
The expectation $E[I_B]$ that covers the case when users start searching in the backward direction, can be calculated in a similar manner than $E[I_F]$, by replacing I_F, p_{sf}, b, and $-a$ by I_B, p_{sb}, $-a$, and b, respectively in Equation 7.21. Finally, the expected effort for the complete transition system, $E[I]$, is then calculated as shown in Equation 7.15 based on $E[I_F]$ and $E[I_B]$.

The preceding derivation of $E[I]$ represents the expected amount of effort a user would invest in browsing horizontally in the forward and backward directions, until reaching the boundaries of the document. In reality, users are likely to stop browsing before reaching the ending of a document if they come across a relevant region. In order to account for this condition in the estimation of $E[I]$, the boundaries $[a, b]$ are defined so that if the document contains a relevant region at position r, then $b = r$ if $r \geq 0$ and $a = r$ otherwise. Given these definitions, the effort a user with parameters $p_f = 0.8$, $p_{sf} = 0.975$, $p_{sb} = 0.96$, $p_{cf} = 0.8$, and $p_{cb} = 0.5$, is expected to invest in auditioning an entry point when starting from position 0 within a document with boundaries $[-1000, 10]$ is 18.95 time units. For a user with zero probability of searching backwards ($p_f = 1$ and $p_{cb} = 0$) the effort reduces to 4.74, while for a user with zero probability of searching forward ($p_f = 0$ and $p_{cf} = 0$) the effort increases to 24. For an extremely persistent user who is willing to search in both backward and forward directions with equal probability ($p_f = 0.5$, $p_{cb} = p_{cf} = 1$) and with persistence $p_{sf} = p_{sb} = 0.9975$, the expected effort for inspecting the entry point within the interval $[-1000, 10]$ ascends to 315 time units.

7.3.2 Vertical browsing model

The previous section described our probabilistic model of horizontal browsing behaviour. For a hypothetical user that begins browsing a document by following an entry point suggested by a SCR system, the horizontal model can be used to produce an estimate of the gain the user would acquire by inspecting this result and the effort associated with this action. In particular, gain is calculated as the probability that the user finds any relevant region, while effort is calculated as the time the user is expected to invest in browsing content. This model provides all the necessary components to calculate horizontal gain and effort for a single search result only. In order to incorporate factors associated with vertical browsing, this section describes how our horizontal model can be augmented to consider gain and effort associated with the process of inspecting multiple SCR results,

Figure 7.6: Complete model of browsing behaviour. The entire horizontal model acts as a starting state which users can visit to audit the next search result in the ranking. E is the ending state, visited by users when finishing their vertical search.



arranged as an ordered list of entry points.

Our complete model augmented with vertical browsing components is illustrated in Figure 7.6. As can be seen from the state diagram, our model follows the underlying model of ranked-biased precision (RBP) (Moffat and Zobel, 2008), previously described in Section 2.1.3. Initially, users begin by inspecting the first element in the list, e_1. Here, the action of "inspecting" an entry point consists of performing horizontal browsing within a document. After examining e_1, users can opt for moving onto examining the next result in the list, e_2, with continuation probability p_c, or finishing their search with complementary probability. Under these assumptions, our evaluation procedure can be framed within the gain-discount framework, with gain derived from the probabilities computed by the horizontal browsing model and discount given by the continuation probabilities p_c as well as horizontal effort. The following sections give the details of how gain and effort are calculated in our extended model.

Deriving gain from vertical browsing

Let $e_1, e_2, \ldots, e_K$ be a ranked list of K search results produced for a query by a SCR system. Each of these results represent an entry point suggestion for horizontal browsing, expressed as an interval $[a, b]$, with $a \leq 0$ indicating the number of time units from the entry point relative to the beginning of the document, and $b \geq 0$ its analogue with respect to the ending of the document. Let $r_1, r_2, \ldots r_R$ be the list of starting points of regions in the collection of documents that are known to be relevant for this query.

Consider first a ranking with a single search result e_1. Such an entry point may be located within a document that does not contain any relevant region. If this is the case, the probability of the user finding a relevant starting point according to our horizontal model will be zero, and so the gain g_1 assigned to rank 1. By contrast, the document associated with e_1 may contain a subset of relevant regions which could effectively be found by the user. Within our evaluation framework, the assumption is that users will stop browsing horizontally and leave the document once they find a relevant region. Among the multiple regions that may be reachable by a user from e_1, there will be one with maximum probability of being found. In these circumstances, we assume that g_1 will be given by the maximum probability obtainable from considering all relevant regions that an entry point can lead to. Let $PF(r_i, e_1)$ be the probability that a user finds the relevant region r_i by following e_1 according to our horizontal browsing model. The gain at rank 1, can then be expressed as shown in Equation 7.22.

$$g_1 = PF(r_{max}, e_1), \qquad \text{where} \quad r_{max} = \arg\max_r PF(r, e_1) \tag{7.22}$$

When the ranking to be evaluated contains multiple entry points $e_1, e_2, \ldots, e_K$, users may be able to locate some relevant region r by following any of the entries that point to the document containing r. Some of these entries may in fact point to similar locations within the document and, depending on how close they are from each other, be treated as “near duplicates” by many of the evaluation measures described in Section 7.2. The strategy adopted by these measures for handling duplicate results consists of treating the top-ranked result, e_k, as relevant and all its lower-ranked duplicates, e_{k+j}, $j > 0$, as non-relevant. Note that this is based on the implicit assumption that users will be able to locate r from e_k independently of the quality of this entry point and its proximity to r. If an SCR system produces a perfect entry point at rank e_{k+1}, these measures will not reward this system appropriately, and would instead measure user gain and effort based on the potentially less optimal entry point e_k.

Instead of assuming that users are always able to find r from the best-ranked near duplicate, our model considers that the user can find r with some non-zero probability, which may be lower than 1. Near duplicates are then taken care of by adopting a “cascade” model, similar to that used in the expected reciprocal rank (ERR) measure (Chapelle et al., 2009). Recall from the description of ERR in Section 2.1.3 that the main idea of the cascade model is to assume that users would be less interested in examining documents further down in the ranking after having found a highly relevant document at previous ranks. However, within the context of our evaluation framework, the assumption made is that a user will be less interested in r when inspecting some ranked result, and thus derive less gain from it, if it is highly likely that the user will find r at previous ranks. Thus, if the user has low chances of finding r from result e_k, but high chances of finding it from e_{k+1}, then they are likely to derive more gain in finding r from e_{k+1} than from e_k. On the contrary, if r can be located from e_k with high probability, it is more likely that more gain will be derived from this result than from a result e_{k+1} with lower associated probability.

Based on this adaptation of the cascade model, the probability that a user locates r from e_k, after having examined all previous search results $e_1, \ldots, e_{k-1}$ in the ranking can be calculated as shown in Equation 7.23.

$$PF@k(r) = \prod_{i=1}^{k-1} (1 - PF(r, e_i))\, PF(r, e_k) \tag{7.23}$$

By defining $PF@k(r)$ this way, the simplifying assumption made is that the probability that a user finds r from the result e_k is independent from that of the user finding r from any other search result in the ranked list.

The gain at rank k for the general case of a ranked list of search results $e_1, \ldots, e_K$ and multiple possible relevant regions $r_1, \ldots, r_R$, can thus be computed as shown in Equation 7.24.

$$g_k = PF@k(r_{max}) \qquad \text{where} \quad r_{max} = \arg\max_r PF(r, e_k) \tag{7.24}$$

By taking the maximum possible r_{max} at rank k, our model assumes that among all relevant regions contained in the document associated with e_k, it is more likely that the user will encounter r_{max}. After finding r_{max}, the user is assumed to abandon horizontal search and to drive their attention back to the ranked list.

Combining vertical and horizontal effort

In order for our evaluation measure to be completely defined under the gain-discount framework, concrete definitions for the discount factor d_k and normalisation $\mathcal{N}$ must be given. The discount factor within our model is calculated as in the ranked-biased precision (RBP) measure, based on the deepness of the ranks that users must reach in their quest for relevant information.

First, note that vertical effort can be seen as an increasing function of k, $eff(k)$, that increases every time the user moves from a rank k on to $k + 1$. Conversely, the inverse of the effort function, $eff(k)^{-1}$, is a function that decreases with k. The inverse of effort can alternatively be interpreted as the willingness of a user to continue up to a certain rank k. This interpretation is adopted by the model underlying RBP, in which the inverse of the effort acts as the factor d_k that discounts the gain users may acquire at each rank k. By adopting a similar interpretation for our vertical browsing model, the discount associated with vertical browsing for a rank $1 \leq k \leq K$ can be expressed as shown in Equation 7.25.

$$\frac{p_c^{k-1}\,(1 - p_c)}{1 - p_c^{K+1}} \tag{7.25}$$

Although Equation 7.25 could be used directly to complete our definition of d_k, such a discount would only account for vertical browsing effort, neglecting any horizontal effort that may be associated with the process of examining the result at rank k. Instead, the final discount factor associated with rank k adopted in our evaluation framework combines both types of effort. This is formally expressed in Equation 7.26,

$$d_k = \frac{p_c^{k-1}\,(1-p_c)}{(E[I_k]+1)\,(1-p_c^{K+1})} \tag{7.26}$$

where $E[I_k]$ is the expected horizontal browsing effort calculated as shown in Equation 7.15 for the entry point and document boundaries associated with result e_k. This estimate applies a geometric discount to model the attenuation of user interest as a function of k, and an additional measure $(E[I_k]+1)^{-1}$ which further applies a discount proportional to the extra cost users are expected to pay while browsing within the k-th result.

Given our definitions of g_k and d_k, the final evaluation measure induced by our user models, termed "no pain no gain" (NPNG), is given in Equation 7.27,

$$NPNG = \frac{1}{\mathcal{N}}\sum_{k=1}^{K} g_k\, d_k = \frac{1}{\mathcal{N}}\sum_{k=1}^{K} PF@k(r_{max})\,\frac{p_c^{k-1}\,(1-p_c)}{(E[I_k]+1)\,(1-p_c^{K+1})} \tag{7.27}$$

The last element that still needs to be defined to complete our evaluation measure is the normalisation factor $\mathcal{N} > 0$. In most traditional evaluation measures, this constant is commonly used to map the summation of discounted gain into a value between $[0,1]$, so that the value produced by the metric can be properly compared and averaged across different queries.

Usually, the major difference between queries that evaluation measures attempt to normalise for is the number of items that are known to be relevant in the ground truth for each query. For a query with R relevant items, our definition of g_k establishes that the maximum gain a user can accumulate is $\sum_k g_k = R$. Such a condition may occur if the ranking under consideration contains one entry point per relevant item and if such entries are perfectly aligned with the onsets of the relevant items. Such a ranking would also incur the minimum cumulative effort a user could possible invest in examining the results suggested by the retrieval system. Note that, according to our horizontal browsing model, some users may begin their search by exploring regions that precede the entry points returned by the system. Therefore, under the assumptions made by our model, some users may still have to invest non-zero effort even when given a perfect retrieval output. To account for these factors in the NPNG measure, $\mathcal{N}$ can be defined as the maximum NPNG that a particular user can obtain from a perfect ranking. For a query with relevant items $r_1, \ldots, r_R$, the normalisation factor of NPNG can be calculated as shown in Equation 7.28

$$\mathcal{N} = \sum_{j=1}^{R} \frac{p_c^{j-1}\,(1-p_c)}{(E_{min}[I_j]+1)\,(1-p_c^{K+1})} \tag{7.28}$$

where $E_{min}[I_j]$ is the minimum effort that the user is expected to invest in the process of finding the relevant onset r_j when given the best entry point possible, that is $e = r_j$.

$E_{min}[I_j]$ can be calculated by using Equation 7.15, and by setting the document boundaries to $a = -r_j$ and b to the number of positions between r_j and the end of the document. Because the quantity $E_{min}[I_j]$ is greater for longer documents which have higher values of a and b, the offsets $r_1, \ldots, r_R$ need to be iterated in ascending order in the summation of Equation 7.28. Such an order ensures that the smallest geometric discounts dominated by the quantity p_c^{j-1} are applied against the greatest values of $E_{min}[I_j]$ and that the summation of Equation 7.28 will be maximised.

7.3.3 Summary

This section described the development of a novel evaluation framework for SCR that attempts to model the behaviour of users when searching for relevant information in a ranked list of entry points to documents. Within this framework, a measure called NPNG was instantiated that proposes a simple browsing model, where users are assumed to invest vertical effort while scanning the ranked list of results from top to bottom, and pay some extra effort when carrying out horizontal browsing within each search result by moving in the backward and/or forward directions.

The NPNG measure solves many of the limitations of existing evaluation measures for SCR. Unlike gAP, in which vertical effort is underestimated and overweighted by horizontal effort, NPNG can consider a wide range of user models, capturing the behaviour of some users who may be unwilling to examine results below certain positions in the ranking and instead prefer investing their time and effort in horizontal exploration. Additionally, NPNG properly accounts for the effort that users may incur in examining entry points that do not lead to any relevant material which is undervalued in the case of gAP.

In contrast to the majority of two-sided evaluation measures, NPNG does not evaluate the quality of the retrieved end points. In this way, it avoids some of the biases that overlap-based two-sided measures have with respect to passage length. Unlike two-sided measures, NPNG implements a flexible and more realistic user model which properly accounts for cases in which users may still be able to find relevant material even when the retrieved passage does not overlap with the relevant content.

7.4 Cross-evaluation of content structuring methods for SCR

This section presents a large-scale comparison of different content structuring approaches in the context of a SCR task for collections of unstructured documents. Content structuring methods are evaluated in terms of their ability to provide increased SCR effectiveness, measured in terms of NPNG for a diverse range of users with different browsing habits. Within this experimental setup, a comparison of evaluation measures for SCR is carried out in order to gain further insight into the advantages that NPNG may offer over other evaluation measures.

7.4.1 Task, collections, and evaluation measures

Structuring approaches were evaluated in the context of an SCR task, where the goal was to produce a ranked list of document offsets pointing to locations where relevant content may be found by users. A particular SCR method was then represented by a content structuring approach and a ranking method. The former was used to determine entry points or passage candidates, whereas the latter was used to rank these candidates in order of relevance to a given query. For the sake of comparing our NPNG measure against other one-sided as well as two-sided measures, the SCR methods under evaluation were designed to return both starting and ending offsets (passages) for every search result included in a ranking. Also, SCR methods were designed to represent passage boundaries with time offsets, measured in seconds, relative to the beginning of the speech document to which a passage belongs.

Since the focus is on comparing among structuring approaches in a unstructured retrieval setup, the test collection used for this purpose must contain relevance assessments indicating where relevant regions occur within the spoken documents. In addition, the ground truth must be free from any “boundary” bias that may have been introduced by the dataset creators when collecting relevance assessments via pooling. As discussed in Section 7.1.1, the majority of relevant passages available for the SH14 and SAVA query sets of the BBC2 collection are either 60, 90, or 120 seconds in length, with boundaries determined automatically by the SCR systems that contributed to the pool. Because there is a potential risk of favouring segmentation approaches that produce passages with similar boundaries to these, the experiments described in this section were not conducted with the SH14 and SAVA topics. The SH13 query set was also discarded because of the small number of topics and relevant passages it contains.

Despite being collected via a pooling procedure, the relevant passages available for the SD2, SQD1, and SQD2 topics of the SDPWS2 collection were manually corrected by human assessors and thus represent a less biased test collection that may be more suitable for the evaluation of content structuring methods. Among the different types of transcripts available for the SDPWS2 collection, the experiments reported in this section were conducted with the ASR transcripts produced by the Julius ASR system described in Section 4.2.2, under the MATCH condition of acoustic and language models. Text transcripts were processed as in the experiments from Chapter 6.3 by applying the MeCab tokeniser and by only keeping nouns and verbs in the transcripts not present in a standard stop word list for Japanese. All content segmentation methods considered were applied to these processed versions of the transcripts.

Evaluation measures

In order to better understand how different structuring methods may affect retrieval effectiveness, as well as to set a basis for comparing our NPNG measure, the different SCR approaches were evaluated by using a wide array of evaluation measures. The subsequent experiments report retrieval effectiveness calculated by using the following measures:

Within-segment precision ($SegP$), recall ($SegR$) and F1 ($SegF_1$) These two-sided measures were calculated based on the amount of overlap between the top 100 retrieved passages and relevant passages, without applying any ranked-based discount. Approaches that return long and short passages will tend to obtain higher values of $SegR$ and $SegP$ respectively, while those that produce passages similar to those in the ground truth will acquire high values of $SegF_1$.

Generalised average precision (gAP) The generalised average precision measure was described in Section 7.2.1 and shown in Equation 7.2. Our implementation of gAP used $G = 10$ seconds as the granularity parameter and the standard triangular reward function from Figure 7.1a.

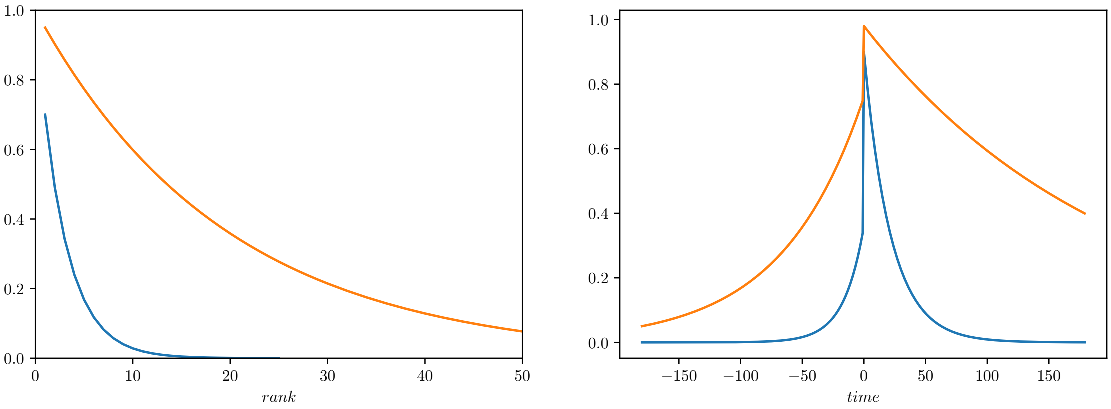
Overlap average precision (oAP) The overlap average precision metric, was described in Sections 5.3.1 and 7.2.2. As discussed previously, oAP tends to favour approaches that return long passages since the measure completely neglects horizontal user effort.

Utterance average precision (uAP) The uAP measure was proposed by Akiba et al. (2011), and was described in Section 7.2.2. Our implementation of uAP first maps passage boundaries onto their corresponding slide units from the documents of the SDPWS2 collection. Given a time-offset t this mapping selects the ID of the slide unit whose span covers position t.

Average segment precision (sAP) and distance weighted precision ($dwsAP$) These are the measures proposed by Eskevich et al. (2012c) that calculate gain by quantifying the within-passage precision and apply an inverse ranked-based discount. The distance weighted counterpart, $dwsAP$, applies the same penalty as gAP to results whose entry points are far away from the onsets of a relevant region. For $dwsAP$, the same reward function as gAP was used with $G = 10$. Our implementation of these measures is the one described in (Eskevich et al., 2012c), in which the normalisation factor $\mathcal{N} = R$ is the number of passages in the ranked list that contain some relevant material. Because the inverse of this normalisation factor will be higher for lower values of R, the sAP and $dwsAP$ measures will tend to favour SCR methods that return a small number of search results.

Average interpolated segment precision ($AiSP$) This is the enhanced measure we proposed in (Racca and Jones, 2015a), described in Section 7.2.3 and inspired by sAP, that assumes a more realistic user model of browsing behaviour based on time-traces. $AiSP$ considers that users browse through a time-trace created by flattening the original ranking

Figure 7.7: A tolerant (orange) and intolerant (blue) group of users for NPNG. The left plot shows the probability of reaching a certain rank. The right plot shows the probability of consuming a certain amount of seconds of speech material.



of passages into a ranking of individual time units (in seconds). The model underlying *AiSP* assumes that users can browse back and forth looking for relevant material, but never beyond the boundaries of the retrieved passage. Gain is defined as the total time the user spends listening to relevant material divided by the total time spent. Discount is applied within the offsets in the time-trace (time positions), and average precision is calculated at fixed recall points.

No-pain no-gain (NPNG) In an ideal scenario, the parameters of NPNG would be estimated based on analysis of the search logs of an SCR system. In the absence of such data, the SCR methods under study were evaluated for four user models in NPNG, each representing a particular profile of "persistence" with respect to vertical and horizontal browsing. Two persistence profiles were considered: one representing tolerant users willing to invest large amounts of time in either vertical or horizontal search; and a second one representing impatient users who would prefer to spend less time searching at the risk of not finding much relevant material. Figure 7.7 illustrates these two user profiles for vertical and horizontal browsing. The left plot shows the probability of reaching a certain rank, while the right plots the probability of listening to speech up to or for more than a certain amount of time. There are thus four possible user combinations: $v\uparrow h\uparrow$ high vertical and high horizontal patience, $v_i\downarrow h_i\downarrow$ low vertical and low horizontal patience, and the remainder combinations $v\uparrow h\downarrow$ and $v\downarrow h\uparrow$ which interleave the low and high persistent profiles. Table 7.1 shows the specific persistence probabilities used to instantiate each of these user models in NPNG.

For evaluating a particular SCR approach over a set of queries, each query was evaluated independently with an evaluation measure, and then these results averaged using an arithmetic mean as in MAP. All measures, with exception to NPNG, implement the traditional strategy for handling near duplicates in the ranked list, where users can derive gain at most once for a relevant region in the ground truth. Subsequent results pointing to or

Table 7.1: Persistence probabilities for tolerance and intolerant users

Model	p_c	p_f	p_{sf}	p_{sb}	p_{cb}	p_{cf}
$v{\uparrow}h{\uparrow}$	.95	.80	.995	.985	.90	.70
$v{\downarrow}h{\downarrow}$	.70	.80	.955	.940	.50	.20
$v{\uparrow}h{\downarrow}$	.95	.80	.955	.940	.50	.20
$v{\downarrow}h{\uparrow}$	.70	.80	.995	.985	.90	.70

overlapping with a relevant region that has already been consumed at previous ranks are treated as non-relevant results in our implementations. Only the top 300 highest-ranked results produced by each SCR method for a query were evaluated.

7.4.2 Comparison of content structuring methods

This section describes the experimental setup and the results of our comparison experiments. To allow for a fair comparison of different structuring techniques, the same retrieval function was used for scoring passages produced by these structuring methods. Exceptions to this were the positional and HMM based methods described later, which by design need special retrieval configurations. The retrieval function chosen for the remainder of the methods was the BM25 function from Equation 2.9, extended with the exponential IDF factor as described in Section 6.3.2. BM25 parameters were set to $b = 0.42$, $k_1 = 2$, $k_3 = 31$, and $d = 1.4$, which result from averaging the optimal parameters obtained with BM25 for the query sets SD2, SQD1, and SQD2 in the experiments reported in Section 6.3.2. Because different passage collections would result from applying different structuring methods to a collection of documents, the collection statistics needed for BM25 scoring, particularly the N, n_t for term t, $docl$, and $avedl$ values, were calculated for each structuring method separately based on the collection of passages produced by each method.

Structuring methods

The following list of content structuring approaches were evaluated in terms of NPNG and the rest of the measures listed in Section 7.4.1. Detailed descriptions for the majority of these methods were provided in Sections 2.3 and 2.4.

Full-document (DOC) A trivial structuring approach is not to perform any structuring at all. This method segments a document into a single long passage that covers the entire contents of the document.

Inter-pausal units (IPU) This method divides a spoken document into its constituent utterances, denoted in the SDPWS2 collection as inter-pausal units (IPUs). While IPUs may contain too few terms to provide optimal SCR effectiveness, they may still be useful to detect short regions of content containing query phrases, which may represent good entry points to return as search results for a query.

Slide-group units (SLIDE) This method segments a speech recording into its slide-group passages. Recall from Section 4.2.2 that slide-groups were created manually by human assessors who clustered together slides describing an heterogeneous topic in an SDPWS2 recording. Thus, SLIDE units represent, a priori, a possible “ideal” collection of retrieval units for the SDPWS2 collection.

Oracle units (ORACLE) For a given query, this approach divides a document into the passages that are known to be relevant for this query. Non-relevant content, that is, content excerpts which are not covered by any relevant region, are fragmented into SLIDE units by default. Note that this approach is query dependent as it produces a different passage segmentation for different queries.

Sliding windows (WIN) The most traditional structuring method for SCR and passage retrieval tasks, whereby documents are fragmented into arbitrary passages of fixed or variable length by moving a window across the contents of the document and extracting a passage every time the window is right-shifted a certain number of steps. Several variations of this approach were considered, using different values for the length and step parameters, as well as two passage consolidation strategies, filtering and recombination, for eliminating near-duplicate results in the ranked lists. The filtering strategy discards the passage at rank k if it overlaps with any of the passages ranked at $j < k$. The merging strategy combines adjacent or overlapping passages into a single passage, which is then given the rank of its highest scoring passage. In what follows, windowing approaches are denoted by WIN-L-X, with L denoting the length of the window used and X being either F or M for the filtering or merging strategies respectively.

Multi-windows (MWIN) This is the method proposed by Kaszkiel and Zobel (1997) for producing arbitrary variable length passages by sliding windows of different lengths over the documents. The final collection of passages is then comprised of the union of all passages produced by each independent sliding window procedure. For the experiments with MWIN, the windows chosen to produce passages had lengths ranging from 30 up to 480 seconds in increments of 30. Because many passages in the union overlap, the filtering strategy described above for WIN was applied to ranked list produced by the MWIN method to remove near duplicate results.

Text Tilling (TT), C99, Utiyama and Isahara (UI), MinCut (MC), and BayesSeg (BS) These correspond to the text segmentation algorithms developed respectively by Hearst (1997), Choi (2000), Utiyama and Isahara (2001), Malioutov and Barzilay (2006), and Eisenstein and Barzilay (2008), described previously in Section 2.3. Each of these algorithms was used to produce a collection of passages by processing every ASR transcript of the SDPWS2 collection. Since some of these methods assume the input to be arranged as a list of sentences, IPUs were used to construct sentences from the words appearing in a speech transcript. For each IPU in a transcript, a sentence was constructed with the words appearing in each IPU. The implementation of TT used is the one included in the NLTK toolkit v3.2.2 (Bird, 2006)[1]. For C99, the original implementation of the algorithm by Choi was used. For UI, MC, and the BS algorithms, the implementations released by Eisenstein and Barzilay[2] were used. Default parameters were used for the C99, UI, MC, and BS algorithms. Since MC requires the number of desired segments per document, this was set to the number of segments produced by the BS algorithm for each document.

TT variations (TT-k, TT-k-ED and TT-k-EC) Recall that TT forms blocks of text by grouping k adjacent pseudo-sentences, which are in turn formed by sequences of w consecutive terms. The TT algorithm was configured to produce passages of different sizes by varying the k in the argument in the NLTK implementation, which specifies the number of pseudo-sentences considered by TT in each text block. In the results reported in this section, these alternative configurations are denoted by TT-k. In all cases, the number of words per pseudo-sentence was set to $w = 20$. Additionally, in order to gain further insight into the differences that may exist between windowing and lexical cohesion approaches, experiments were conducted with a modified version of TT that produces overlapping passages of similar length. For this, the passages produced by TT-k were extended so that they acquired a fixed length value of l. Two alternatives for computing l were explored: In TT-k-ED, all passages from document d were extended to have length equal to that of the longest passage in d. In TT-k-EC, all passages from the collection were extended to the length of the longest passage in that collection.

Positional models (PM, PS-PM, and NS-PM) Whereas all approaches listed above produce a static segmentation of the document collection, the PM and S-PM techniques infer retrieval units dynamically, depending on the contents of the query. For a given query, a positional model was used to calculate a relevance score for every position within a document, based on the pseudo-frequency counts that every query term appearing in the document propagates to this position. As described in Section 6.2.2, every occurrence of a term t contributes to the pseudo-frequency of t at some position p by the distance that exist between each term occurrence and p. The pseudo-frequency values of all terms at every position p were then multiplied by their respective inverse document frequencies and finally summed in order to obtain a BM25 score for every position within the document. These relevance scores thus form a density contour over document positions, with peaks and valleys found at locations where there is a high, respectively low, density of highly discriminative query terms. A similar technique to that implemented by TextTilling was

[1] `http://www.nltk.org/`

[2] `https://github.com/jacobeisenstein/bayes-seg`

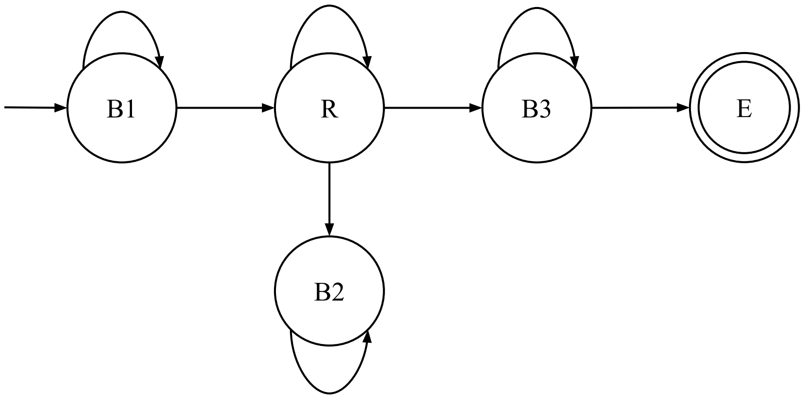
then used to find the locations of prominent peaks. The locations of prominent peaks for each document in the collection were next ranked based on their associated relevance scores and returned as a list of one-sided results for the query. Two implementations of this technique were considered for modelling term propagation, one that uses the symmetric Gaussian kernel (PM) from Equation 6.2 and another that uses the skewed Gaussian kernel from Equation 6.3 with either positive (PS-PM) or negative (NS-PM) skewness. In both cases, the spread parameter σ was set to 100 seconds, and the skewness parameter α to positive or minus 0.07. To allow for the application of two-sided evaluation measures, the results produced by these methods were converted into passages of 50 seconds length. Because positional models calculate term frequencies in a substantially different manner than standard BM25 applied to passages, term weights in the position models were computed with BM25 parameters set to $b = 0.02$, $k_1 = 4.16$, $k_3 = 83.3$, and $d = 1.37$, which performed significantly better than default values for this approach in the experiments from Chapter 6.

Cover units (COVER) This method is inspired by the work of Clarke et al. (2000a), and performs a query-dependent segmentation of a document based on "cover sets". A cover set is defined as the shortest passage that captures a certain number of unique query terms. In our implementation of this method, a new passage is generated for every pair of query terms appearing in a document. The boundaries of a cover passage are defined as the positions of these occurrences within the document. Applied naively, this process would generate n^2 passages for a document containing n occurrences of query terms. In order to keep the number of passages low, our implementation only generates cover passages by considering only the top 3 terms from the query with the highest inverse document frequency. Overlapping passages in the final ranking of cover passages were removed with the same filtering strategy used in the sliding window methods.

Divisive clustering (DIV) This method adopts a top-down recursive approach to progressively divide a document into smaller units for a particular query. At each step, the algorithm selects a position at which to split an item into a left and a right part. The criteria for selecting a splitting position p seeks to maximise the relevance scores of any of the individual parts that would result if such an item is split at p. Once a maximal splitting position is found, the score of its associated left and right parts is compared against that of the item. If a part scores higher than the item, then the algorithm is applied recursively to this part. Otherwise, the algorithm stops and returns the history of breaking positions found so far from which passages are then generated.

Hidden Markov Model (HMM) This method replicates the HMM approach used by Jiang and Zhai (2006) at TREC HARD 2004. The observations of the HMM correspond to words, while the set of hidden states corresponds to either: a background state, assumed

Figure 7.8: 5-state HMM structure proposed by Jiang and Zhai (2006) for passage retrieval. The states B1, B2, and B3 are background states, while R and E are the relevant and ending states respectively.



to generate non-query words; and a relevant state, assumed to emit query-related words. Emission probabilities of background states over non-query words are given by a maximum-likelihood language model, estimated from the entire collection of documents. The output probabilities at the relevant state are instead given by a relevance language model (RLM), estimated by interpolating a maximum-likelihood query and collection LMs with Jelinek-Mercer smoothing, and a smoothing factor equal to 0.5. A segmentation for a document can be inferred from the most likely sequence of states in the HMM that generates the document. In particular, contiguous spans of words generated by the relevant state can then be extracted as hypothetical relevant passages to be retrieved. Our implementation adopts the 5-state HMM structure used by Jiang and Zhai, with three background states, a relevant state inbetween background states, and an ending state to mark the ending of the document. This HMM structure is shown in Figure 7.8. The transition probabilities in this HMM were set to those reported in Jiang and Zhai (2006), which the authors obtained by training their HMM with TREC HARD data.

The structuring methods considered may be divided into two broad categories. A category of static segmentation approaches, which produce a collection of query-independent segments. And a group of dynamic or flexible approaches, which define segments based on the contents of the query. Note that since some of the dynamic approaches considered do not produce passages at indexing time, it is unclear how the collection statistics needed for BM25 scoring should be calculated for these methods. The approach taken in this study was to re-use the collection statistics derived from SLIDE units to calculate term weights for the dynamic structuring methods under analysis.

Table 7.2 provides statistics about the passages that are produced by static structuring methods. IPUs and DOC units represent the shortest and longest static units considered, while the rest of segmentation methods produced passages with lengths varying between 15 and 400 seconds. When used with default parameters, most lexical cohesion approaches produced passages with similar characteristics, with the exception of UI which tended to

Table 7.2: Collection statistics for passages produced by static structuring methods for the SDPWS2 collection. Each column specifies: the total number of passages in the collection (N), their average length in seconds (*avel*), their standard length deviation (*stdl*), and the amount of overlap between adjacent passages (*over*).

Method	N	*avel*	*stdl*	*over*
IPU	37,757	2.5	2.0	0%
SLIDE	2,334	47.7	46.3	0%
DOC	98	1,202.6	145.8	0%
WIN-15-5	23,650	14.9	1.0	66%
WIN-30-15	7,921	29.6	2.9	49%
WIN-60-30	3,986	58.4	7.9	49%
WIN-90-45	2,669	86.7	13.6	49%
WIN-120-60	2,019	113.9	21.4	48%
WIN-500-250	511	413.7	140.9	46%
MWIN	27,402	130.4	118.7	49%
TT-5	4,592	25.7	10.6	0%
TT-10	2,908	40.6	20.6	0%
TT-15	2,005	58.9	27.7	0%
TT-20	1,578	74.8	36.6	0%

Method	N	*avel*	*stdl*	*over*
TT-5-ED	4,592	66.9	22.5	59%
TT-10-ED	2,908	93.2	24.6	55%
TT-15-ED	2,005	119.9	26.9	50%
TT-20-ED	1,578	150.3	36.1	49%
TT-5-EC	4,592	143.9	2.1	82%
TT-10-EC	2,908	233.8	5.5	82%
TT-15-EC	2,005	233.8	4.5	74%
TT-20-EC	1,578	337.6	9.0	77%
MC	1,708	69.1	79.2	0%
C99	1,495	78.9	67.2	0%
BS	1,708	105.2	162.4	0%
UI	599	197.0	102.6	0%

produce longer passages. Table 7.2 also shows how extending TT passages to a fixed length (TT-k-ED and TT-k-EC) increases the average passage length and overlap, and reduces length variability among passages.

Table 7.3 presents length statistics for passages returned at rank 300 or lower by the retrieval methods under study. For structuring methods that produce overlapping passages and that adopt a filtering or merging post-processing strategy, retrieval methods return 6 or less passages per document in the top 300 results. The reason for this is that passage rankings are generated by only considering the passages from the top 50 full-documents ranked by BM25 for the SDPWS2 collection. The number of passages per document decreases from 6 to about 3 and 1.7 for methods that produce longer passages or with high overlap. Among dynamic structuring methods, the COVER and DIV methods tend to produce longer passages relative to windowing methods. Compared to the standard PM method, positional models that use a skewed Gaussian kernel (NS-PM and PS-PM) tend to produce more peaky density contours, which results in more results retrieved per document. Also, the fact that the *avel* figures from Table 7.2 are greater than those from Table 7.3 evidences that BM25 tends to assign increased scores to passages that are longer than the average in each passage collection.

Results of comparison experiments

Table 7.4 presents effectiveness values of all structuring methods under study, as calculated by the traditional evaluation measures considered and NPNG for the four user models described previously. Each effectiveness value was obtained by averaging the values produced by a measure across the 220 queries that comprise the SD2, SQD1, and SQD2 topic sets. The superscript of an effectiveness value indicates the its relative rank with respect to

Table 7.3: Length statistics of passages retrieved by BM25, PMs, or HMM based methods for the SDPWS2 queries for each structuring technique. The columns indicate: the average *depth* of the ranked lists; the average number of returned passages per document and query (*aveN*); the average length of returned passages in seconds (*avel*); and their standard length deviation (*stdl*).

Method	*depth*	*aveN*	*avel*	*stdl*
IPU	300.0	6.3	3.8	1.3
DOC	46.7	1.0	1210.7	0.0
SLIDE	300.0	6.0	80.9	38.2
ORACLE	300.0	6.0	81.1	38.7
WIN-60-30-F	160.8	3.4	59.4	0.8
WIN-90-45-F	156.4	3.2	88.6	1.8
WIN-120-60-F	153.5	3.1	117.3	3.6
WIN-500-250-F	126.2	2.6	416.8	107.7
WIN-15-5-M	122.3	3.1	21.8	3.2
WIN-30-15-M	147.7	3.3	44.2	8.4
WIN-60-30-M	131.0	2.8	96.6	20.8
WIN-90-45-M	120.9	2.5	155.1	34.5
MWIN-F	18.0	1.3	229.3	15.3
TT-5	300.0	6.1	28.6	9.0
TT-15	300.0	6.0	67.8	22.0
TT-10	300.0	6.0	48.9	16.4
TT-20	300.0	6.0	86.1	30.6

Method	*depth*	*aveN*	*avel*	*stdl*
TT-5-ED-F	123.1	2.7	66.1	0.0
TT-15-ED-F	134.3	2.8	119.8	0.1
TT-10-ED-F	128.7	2.7	93.5	0.1
TT-20-ED-F	132.8	2.7	152.5	0.1
TT-5-EC-F	62.6	1.8	143.9	0.1
TT-15-EC-F	77.7	1.8	233.9	0.1
TT-10-EC-F	60.2	1.6	233.9	0.1
TT-20-EC-F	70.1	1.6	337.7	0.2
C99	300.0	6.0	123.3	62.3
UI	293.8	6.0	205.3	85.6
MC	300.0	6.0	138.2	66.8
BS	300.0	6.0	247.0	150.3
PM	165.1	3.3	50.0	0.0
NS-PM	280.4	5.6	50.0	0.0
PS-PM	279.6	5.6	50.0	0.0
COVER	10.7	1.2	231.6	14.0
DIV	299.0	6.0	133.3	95.3
HMM	113.7	2.3	60.1	0.0

others from the same column, and therefore shows how structuring methods would rank relative to other methods if sorted by a specific evaluation measure. The last column in the table shows the average rank obtained by each structuring method across all evaluation measures.

The following observations can be drawn from the results shown in Table 7.4.

O1: As expected, the ORACLE method attains the highest effectiveness values overall, across the majority of the evaluation measures considered. Thus, despite ORACLE passages being highly variable in terms of length, the problems associated with length normalisation in BM25 may be out-weighted by the benefits of using passages that capture the exact boundaries of the relevant regions.

O2: Returning full-documents instead of passages (DOC) as search results decreases performance, according to effectiveness measures that penalise horizontal user effort, such as gAP, uAP, sAP, $dwsAP$, $AiSP$, and NPNG.

O3: Among lexical-cohesion based methods, TT performed generally more effectively than C99, UI, MC, and BS across most evaluation measures. Extending TT passages so that they have the length of the longest passage in a document prior to retrieval (TT-k-ED) seems to be beneficial, as long as the resulting passages do not become excessively long.

O4: Within sliding window approaches, it is not entirely clear if any of the configurations evaluated performs better than the rest, as the most effective configuration varies between measures. A trend that arises from these results though is that methods producing long windows, and generally long passages, tend to acquire lower effectiveness scores.

O5: Among positional models, using a skewed Gaussian kernel with negative skewness (NS-PM) performed generally better than using the symmetric (PM) and positive skewed (PS-PM) kernels. Using a negative skewed kernel produces the effect of propagating relevance scores back in time. Because the relevant scores of future positions within a document get propagated to positions in the past, past positions then become better estimators of the relevance of subsequent content and, consequently, better entry point candidates.

O6: Within the group of non-positional dynamic structuring methods, divisive clustering (DIV) performed effectively in terms of vertical ranking quality as indicated by the measures oAP, uAP, and NPNG with $v{\downarrow}h{\uparrow}$, while not producing accurate entry points as shown by gAP and NPNG with $h{\downarrow}$. The methods COVER and HMM performed poorly overall.

O7: As shown by gAP, and NPNG with $h{\downarrow}$, adopting a segmentation strategy that produces a large number of relatively short overlapping passages, combined with post filtering or merging as in WIN-60-30-F and WIN-15-5-M, results in more accurate entry points that can reduce horizontal effort dramatically relative to other methods. However, considering small ranking units appears to be detrimental for ranking quality, as methods that produce longer units tend to perform better for measures that value vertical effort more highly than horizontal effort.

O8: After ORACLE, it is unclear which segmentation method performs best overall, as there is not a single method that performs substantially better than the rest across all or the majority of the evaluation measures. Rather, different segmentation methods qualify as "second-best" under the scope of different evaluation measures. Whether a segmentation method acquires a high effectiveness score under a measure seems to depend strongly on the assumptions made about which features characterise an "ideal" ranked list of result and the assumed model of user behaviour. For instance, because SLIDE units tend to align well with the beginnings of relevant regions in the SDPWS2 collection, the SLIDE method ranks in second place for the uAP measure, as well as for measures that heavily favour accurate entry points, such as gAP and NPNG with $h{\downarrow}$. The fact that the COVER method is the second-best under the sAP and $dwsAP$ measures most likely arises due to the few number of results returned by this method for each query, which can minimise the normalisation factor of sAP and $dwsAP$. Returning short IPUs performs best in terms of $AiSP$, possibly because $AiSP$ calculates horizontal effort as the sum of the lengths of the passages in the ranked list, which is minimal for IPUs when compared to other retrieval units. While WIN-60-30-F ranked as the second-best method according to NPNG for patient users ($v{\uparrow}h{\uparrow}$), less patient users ($v{\downarrow}h{\uparrow}$) would most likely derive more gain from the MWIN-F, DIV, and TT-5-ED-F methods.

Observation O8 makes it clear that no single evaluation measure provides sufficient information to properly determine the relative effectiveness of one SCR method compared to others. The conclusions drawn from using a single evaluation measure will most likely be biased towards the particular type of SCR methods favoured by the chosen measure. Using a group of evaluation measures instead provides useful information about the different strengths and weaknesses of each individual method. However, methods tend to be ranked inconsistently by the different measures, which makes it hard to devise an overall winner. Although computing the average rank may seem to provide a useful indicator about the overall average performance of an SCR approach, it is not obvious whether all measures should be treated equally by such an average, neither how to combine directly the effectiveness scores from multiple measures.

A single evaluation measure represents a single specific type of user, with a particular type of behaviour. At the same time, different structuring approaches induce SCR systems that can produce search results with different characteristics. In the same way a specific user may find more benefit from using one specific SCR system over some another, a specific SCR system may better satisfy one type of user over one another. In this regard, the results from Table 7.4 suggest that:

1. SCR methods based on short passages produced by sliding windows are the most effective for highly patient users ($v{\uparrow}h{\uparrow}$), and may even perform better for these users than methods based on manually generated passages (SLIDE);

2. Methods that consider longer ranking units such as those produced by MWIN-F and DIV are most effective for patient users who only inspect the top results in the ranked lists ($v{\downarrow}h{\uparrow}$), as longer units help to improve ranking quality overall;

3. Impatient users who are not willing to invest significant amounts of time in browsing speech content are best satisfied with SCR methods that produce highly accurate entry points close to the beginning of relevant regions, such as manual (SLIDE) and short units (WIN-15-5-M TT-5).

Recall from the experiments presented in Chapter 6, that the ranking of SLIDE units can be improved if units are re-ranked based on the relevance scores of their documents. Such a re-ranking strategy could potentially improve the quality of the rankings produced by every SCR method in general, and may bring the performance of short units closer to that obtained with long units for users of type $v{\downarrow}h{\uparrow}$. In order to determine the extent of these improvements, the ranked lists induced from the structuring methods were re-ranked based on the document score interpolation (DSI) technique described in Section 6.2.1, and then re-evaluated.

Table 7.5 shows the resulting effectiveness scores. For measures that are able to capture ranking quality, the DSI technique provided increased effectiveness scores for the majority of SCR methods. These improvements were particularly notorious for SCR methods that return short passages, such as IPU, WIN-15-5-M, WIN-30-15-M, and WIN-60-30-F. In particular, by adopting these modifications, methods that rely on short units can be seen to be as effective as those that use longer units for patient users who perform shallow vertical searches ($v{\downarrow}h{\uparrow}$).

Thus, the content structuring method that seems most effective across all types of users considered is one that considers relatively short overlapping passages and both document as well as passage level relevance scores to perform the ranking. While short units permit to locate focused regions with a high density of query terms that are useful as entry points suggestions, the re-ranking of these units based on document relevance scores improves the ranking of potentially relevant regions at top positions in the ranked list and increases robustness to ASR errors.

7.5 Summary

This chapter presented a detailed investigation of evaluation measures in the context of an unstructured SCR task and conducted a wide ranging comparison of different structuring approaches for SCR.

Designing evaluation measures for SCR tasks where documents lack of a clear structure requires the consideration of multiple factors which do not need to be accounted for when evaluating straightforward document retrieval systems. Under the scope of traditional pooling strategies, the collection of relevance assessments requires a manual post-processing of the results from the pools to adjust the boundaries of relevant passages and reduce the amount of bias in the resulting ground truth. Also, special considerations regarding the way results will be presented to users in the search results page must be accounted for. For example, the type of browsing and audio playback interface that will be used, as well as the kind of information per result (thumbnails, starting and ending points, density contour, etc) that will be shown to users. Under a standard presentation layout, with results arranged as a ranked list of audio pointers, SCR systems must seek to minimise the effort users are required to invest in inspecting the ranked list of results (vertical browsing) and navigating within the contents of each document (horizontal browsing), while maximising the amount of relevant material presented.

Several evaluation measures for SCR and other unstructured retrieval tasks have been developed in the past, most of which can be seen as a normalised accumulation of gain over ranks discounted by some decreasing function of the ranks. The majority of these measures assume a particular deterministic model of user behaviour in which the factors

Table 7.4: Retrieval effectiveness calculated by traditional evaluation measures and NPNG for various structuring methods over the SDPWS2 queries.

Method	$SegF_1$	$SegP$	$SegR$	gAP	oAP	uAP	sAP	$dwsAP$	$AiSP$	NPNG				*Rank*
										$v{\uparrow}h{\uparrow}$	$v{\downarrow}h{\downarrow}$	$v{\uparrow}h{\downarrow}$	$v{\downarrow}h{\uparrow}$	
DOC	$.191^{35}$	$.118^{35}$	$.977^{1}$	$.037^{35}$	$.305^{17}$	$.137^{34}$	$.098^{35}$	$.029^{35}$	$.147^{35}$	$.149^{33}$	$.037^{33}$	$.035^{35}$	$.130^{31}$	35
IPU	$.231^{34}$	$.827^{2}$	$.161^{35}$	$.149^{20}$	$.149^{35}$	$.143^{33}$	$.181^{15}$	$.144^{7}$	$.325^{1}$	$.192^{29}$	$.063^{13}$	$.108^{6}$	$.110^{33}$	21
SLIDE	$.575^{2}$	$.583^{3}$	$.755^{15}$	$.221^{2}$	$.276^{26}$	$.242^{2}$	$.189^{10}$	$.152^{5}$	$.264^{6}$	$.282^{7}$	$.113^{2}$	$.160^{2}$	$.188^{9}$	2
ORACLE	$.868^{1}$	$.868^{1}$	$.868^{5}$	$.281^{1}$	$.289^{23}$	$.313^{1}$	$.292^{1}$	$.292^{1}$	$.298^{2}$	$.421^{1}$	$.262^{1}$	$.351^{1}$	$.299^{1}$	1
WIN-60-30-F	$.470^{4}$	$.479^{10}$	$.641^{26}$	$.192^{4}$	$.288^{24}$	$.209^{5}$	$.183^{13}$	$.126^{10}$	$.265^{5}$	$.293^{2}$	$.064^{9}$	$.094^{8}$	$.190^{7}$	4
WIN-90-45-F	$.449^{9}$	$.405^{16}$	$.709^{21}$	$.181^{8}$	$.311^{15}$	$.213^{4}$	$.183^{14}$	$.115^{16}$	$.259^{9}$	$.290^{4}$	$.060^{14}$	$.083^{13}$	$.192^{4}$	5
WIN-120-60-F	$.428^{13}$	$.357^{20}$	$.761^{13}$	$.172^{13}$	$.314^{14}$	$.207^{8}$	$.171^{18}$	$.104^{20}$	$.249^{14}$	$.273^{11}$	$.054^{21}$	$.072^{18}$	$.188^{8}$	12
WIN-500-250-F	$.280^{33}$	$.186^{34}$	$.931^{2}$	$.056^{34}$	$.337^{2}$	$.164^{31}$	$.130^{31}$	$.050^{34}$	$.189^{33}$	$.204^{28}$	$.043^{30}$	$.050^{30}$	$.158^{28}$	33
WIN-15-5-M	$.442^{12}$	$.570^{5}$	$.493^{31}$	$.179^{10}$	$.234^{31}$	$.181^{28}$	$.202^{7}$	$.149^{6}$	$.262^{8}$	$.271^{13}$	$.074^{4}$	$.113^{4}$	$.178^{22}$	10
WIN-30-15-M	$.472^{3}$	$.468^{12}$	$.671^{23}$	$.171^{14}$	$.284^{25}$	$.188^{21}$	$.189^{11}$	$.123^{11}$	$.248^{16}$	$.291^{3}$	$.058^{19}$	$.088^{11}$	$.191^{5}$	7
WIN-60-30-M	$.421^{16}$	$.329^{23}$	$.831^{10}$	$.130^{24}$	$.324^{8}$	$.187^{22}$	$.165^{24}$	$.087^{26}$	$.219^{28}$	$.270^{15}$	$.051^{24}$	$.066^{22}$	$.186^{12}$	20
WIN-90-45-M	$.358^{29}$	$.260^{31}$	$.879^{4}$	$.095^{32}$	$.334^{4}$	$.183^{26}$	$.153^{28}$	$.066^{32}$	$.208^{31}$	$.245^{23}$	$.046^{29}$	$.056^{28}$	$.177^{25}$	31
MWIN-F	$.387^{24}$	$.297^{27}$	$.805^{11}$	$.111^{28}$	$.323^{10}$	$.194^{16}$	$.218^{5}$	$.119^{13}$	$.224^{25}$	$.245^{22}$	$.063^{12}$	$.071^{19}$	$.194^{2}$	15
TT-5	$.421^{15}$	$.577^{4}$	$.451^{34}$	$.188^{7}$	$.237^{30}$	$.182^{27}$	$.156^{26}$	$.116^{14}$	$.271^{3}$	$.269^{16}$	$.077^{3}$	$.119^{3}$	$.169^{27}$	14
TT-15	$.452^{8}$	$.441^{13}$	$.650^{24}$	$.178^{11}$	$.292^{22}$	$.204^{10}$	$.154^{27}$	$.100^{21}$	$.252^{12}$	$.289^{5}$	$.068^{7}$	$.099^{7}$	$.187^{11}$	9
TT-10	$.456^{7}$	$.510^{6}$	$.567^{28}$	$.190^{6}$	$.266^{27}$	$.201^{13}$	$.159^{25}$	$.113^{17}$	$.256^{11}$	$.277^{8}$	$.071^{5}$	$.110^{5}$	$.179^{19}$	8
TT-20	$.449^{10}$	$.415^{14}$	$.690^{22}$	$.180^{9}$	$.295^{21}$	$.208^{7}$	$.151^{29}$	$.098^{23}$	$.252^{13}$	$.277^{10}$	$.066^{8}$	$.094^{9}$	$.179^{20}$	13
TT-5-ED-F	$.466^{5}$	$.470^{11}$	$.646^{25}$	$.191^{5}$	$.295^{20}$	$.209^{6}$	$.199^{8}$	$.133^{9}$	$.267^{4}$	$.287^{6}$	$.064^{10}$	$.089^{10}$	$.191^{6}$	3
TT-15-ED-F	$.417^{18}$	$.354^{21}$	$.745^{17}$	$.152^{18}$	$.314^{13}$	$.197^{14}$	$.174^{17}$	$.100^{22}$	$.243^{17}$	$.270^{14}$	$.059^{18}$	$.076^{17}$	$.187^{10}$	16
TT-10-ED-F	$.445^{11}$	$.407^{15}$	$.711^{20}$	$.173^{12}$	$.301^{18}$	$.201^{12}$	$.188^{12}$	$.115^{15}$	$.248^{15}$	$.277^{9}$	$.059^{17}$	$.079^{15}$	$.186^{14}$	11
TT-20-ED-F	$.393^{23}$	$.324^{25}$	$.760^{14}$	$.137^{22}$	$.296^{19}$	$.185^{24}$	$.168^{20}$	$.096^{24}$	$.223^{26}$	$.252^{21}$	$.054^{20}$	$.066^{24}$	$.181^{18}$	25
TT-5-EC-F	$.405^{20}$	$.329^{24}$	$.775^{12}$	$.132^{23}$	$.318^{11}$	$.196^{15}$	$.192^{9}$	$.111^{18}$	$.235^{19}$	$.261^{19}$	$.050^{26}$	$.068^{20}$	$.183^{17}$	18
TT-15-EC-F	$.362^{27}$	$.267^{29}$	$.850^{6}$	$.097^{31}$	$.324^{9}$	$.183^{25}$	$.170^{19}$	$.087^{27}$	$.223^{27}$	$.239^{25}$	$.052^{23}$	$.063^{25}$	$.179^{21}$	28
TT-10-EC-F	$.361^{28}$	$.265^{30}$	$.841^{7}$	$.098^{30}$	$.327^{6}$	$.192^{18}$	$.175^{16}$	$.092^{25}$	$.225^{23}$	$.241^{24}$	$.053^{22}$	$.061^{26}$	$.184^{16}$	23
TT-20-EC-F	$.321^{32}$	$.226^{33}$	$.891^{3}$	$.083^{33}$	$.332^{5}$	$.185^{23}$	$.165^{23}$	$.083^{28}$	$.216^{29}$	$.223^{27}$	$.050^{25}$	$.055^{29}$	$.178^{24}$	29
C99	$.413^{19}$	$.360^{19}$	$.725^{19}$	$.147^{21}$	$.305^{16}$	$.180^{29}$	$.121^{33}$	$.071^{31}$	$.224^{24}$	$.273^{12}$	$.060^{15}$	$.080^{14}$	$.184^{15}$	22
UI	$.376^{25}$	$.284^{28}$	$.840^{8}$	$.125^{25}$	$.325^{7}$	$.189^{20}$	$.129^{32}$	$.072^{30}$	$.229^{22}$	$.257^{20}$	$.047^{28}$	$.068^{21}$	$.170^{26}$	26
MC	$.394^{22}$	$.329^{22}$	$.734^{18}$	$.151^{19}$	$.318^{12}$	$.193^{17}$	$.139^{30}$	$.079^{29}$	$.230^{21}$	$.267^{17}$	$.064^{11}$	$.078^{16}$	$.186^{13}$	19
BS	$.331^{31}$	$.245^{32}$	$.838^{9}$	$.103^{29}$	$.334^{3}$	$.176^{30}$	$.116^{34}$	$.057^{33}$	$.212^{30}$	$.233^{26}$	$.059^{16}$	$.066^{23}$	$.178^{23}$	30
PM	$.405^{21}$	$.486^{9}$	$.489^{32}$	$.166^{16}$	$.216^{33}$	$.191^{19}$	$.248^{3}$	$.162^{4}$	$.233^{20}$	$.144^{34}$	$.037^{34}$	$.044^{33}$	$.109^{34}$	26
NS-PM	$.427^{14}$	$.490^{8}$	$.509^{30}$	$.201^{3}$	$.244^{29}$	$.205^{9}$	$.227^{4}$	$.163^{3}$	$.263^{7}$	$.187^{30}$	$.049^{27}$	$.061^{27}$	$.135^{30}$	17
PS-PM	$.420^{17}$	$.497^{7}$	$.514^{29}$	$.166^{15}$	$.227^{32}$	$.202^{11}$	$.213^{6}$	$.142^{8}$	$.240^{18}$	$.140^{35}$	$.041^{32}$	$.050^{31}$	$.096^{35}$	24
COVER	$.365^{26}$	$.324^{26}$	$.623^{27}$	$.114^{27}$	$.246^{28}$	$.159^{32}$	$.268^{2}$	$.167^{2}$	$.201^{32}$	$.157^{32}$	$.042^{31}$	$.046^{32}$	$.139^{29}$	32
DIV	$.458^{6}$	$.384^{18}$	$.754^{16}$	$.163^{17}$	$.340^{1}$	$.216^{3}$	$.166^{22}$	$.108^{19}$	$.259^{10}$	$.265^{18}$	$.068^{6}$	$.084^{12}$	$.193^{3}$	6
HMM	$.353^{30}$	$.394^{17}$	$.459^{33}$	$.117^{26}$	$.149^{34}$	$.109^{35}$	$.167^{21}$	$.121^{12}$	$.162^{34}$	$.167^{31}$	$.036^{35}$	$.041^{34}$	$.119^{32}$	34

Table 7.5: Retrieval effectiveness calculated by traditional evaluation measures and NPNG for re-ranked lists of passages based on document scores.

Method	*SegF*$_1$	*SegP*	*SegR*	*gAP*	*oAP*	*uAP*	*sAP*	*dwsAP*	*AiSP*	NPNG				*Rank*
										$v{\uparrow}h{\uparrow}$	$v{\downarrow}h{\downarrow}$	$v{\uparrow}h{\downarrow}$	$v{\downarrow}h{\uparrow}$	
DOC	$.191^{35}$	$.118^{35}$	$.977^{1}$	$.037^{35}$	$.305^{17}$	$.137^{34}$	$.098^{35}$	$.029^{35}$	$.147^{35}$	$.149^{34}$	$.037^{34}$	$.035^{35}$	$.130^{33}$	35
IPU	$.247^{34}$	$.849^{2}$	$.173^{35}$	$.178^{13}$	$.188^{34}$	$.189^{29}$	$.264^{3}$	$.196^{2}$	$.356^{1}$	$.238^{27}$	$.072^{6}$	$.144^{3}$	$.133^{32}$	15
SLIDE	$.577^{2}$	$.587^{3}$	$.753^{16}$	$.222^{2}$	$.284^{26}$	$.267^{2}$	$.219^{11}$	$.170^{5}$	$.276^{7}$	$.298^{8}$	$.115^{2}$	$.177^{2}$	$.191^{9}$	2
ORACLE	$.873^{1}$	$.873^{1}$	$.873^{5}$	$.282^{1}$	$.301^{23}$	$.323^{1}$	$.304^{1}$	$.304^{1}$	$.311^{2}$	$.431^{1}$	$.264^{1}$	$.380^{1}$	$.295^{1}$	1
WIN-60-30-F	$.471^{4}$	$.477^{10}$	$.638^{26}$	$.196^{4}$	$.297^{25}$	$.227^{4}$	$.217^{12}$	$.143^{12}$	$.277^{6}$	$.308^{3}$	$.068^{9}$	$.102^{8}$	$.194^{6}$	5
WIN-90-45-F	$.462^{7}$	$.414^{14}$	$.719^{20}$	$.181^{12}$	$.317^{14}$	$.223^{8}$	$.205^{17}$	$.124^{16}$	$.264^{13}$	$.303^{5}$	$.062^{15}$	$.088^{14}$	$.192^{8}$	10
WIN-120-60-F	$.433^{14}$	$.361^{19}$	$.770^{13}$	$.177^{15}$	$.316^{15}$	$.217^{10}$	$.194^{20}$	$.117^{20}$	$.255^{16}$	$.280^{16}$	$.051^{23}$	$.073^{18}$	$.181^{19}$	14
WIN-500-250-F	$.279^{33}$	$.186^{34}$	$.929^{2}$	$.054^{34}$	$.333^{3}$	$.164^{33}$	$.132^{34}$	$.049^{34}$	$.182^{34}$	$.199^{29}$	$.039^{32}$	$.048^{31}$	$.148^{28}$	33
WIN-15-5-M	$.443^{12}$	$.577^{5}$	$.491^{32}$	$.194^{6}$	$.261^{29}$	$.211^{13}$	$.254^{5}$	$.177^{3}$	$.288^{4}$	$.299^{7}$	$.084^{3}$	$.128^{5}$	$.196^{3}$	3
WIN-30-15-M	$.477^{3}$	$.470^{11}$	$.673^{23}$	$.183^{10}$	$.306^{18}$	$.214^{11}$	$.232^{8}$	$.146^{10}$	$.269^{9}$	$.311^{2}$	$.066^{11}$	$.096^{10}$	$.206^{2}$	4
WIN-60-30-M	$.422^{17}$	$.332^{22}$	$.823^{10}$	$.135^{25}$	$.332^{4}$	$.199^{22}$	$.185^{27}$	$.094^{28}$	$.228^{24}$	$.278^{18}$	$.052^{22}$	$.067^{23}$	$.186^{14}$	22
WIN-90-45-M	$.358^{29}$	$.260^{31}$	$.877^{4}$	$.095^{32}$	$.336^{2}$	$.190^{28}$	$.166^{29}$	$.067^{33}$	$.212^{30}$	$.248^{22}$	$.046^{28}$	$.056^{28}$	$.174^{25}$	30
MWIN-F	$.387^{24}$	$.297^{27}$	$.805^{11}$	$.112^{28}$	$.323^{11}$	$.194^{26}$	$.220^{10}$	$.119^{19}$	$.223^{28}$	$.244^{23}$	$.063^{14}$	$.071^{21}$	$.190^{10}$	21
TT-5	$.434^{13}$	$.586^{4}$	$.468^{33}$	$.196^{5}$	$.251^{30}$	$.204^{18}$	$.211^{14}$	$.147^{8}$	$.291^{3}$	$.295^{10}$	$.082^{4}$	$.138^{4}$	$.183^{16}$	9
TT-15	$.450^{9}$	$.441^{13}$	$.644^{24}$	$.183^{11}$	$.303^{22}$	$.224^{6}$	$.194^{21}$	$.124^{15}$	$.266^{10}$	$.304^{4}$	$.071^{7}$	$.104^{7}$	$.192^{7}$	7
TT-10	$.470^{5}$	$.522^{6}$	$.576^{28}$	$.191^{8}$	$.276^{27}$	$.223^{7}$	$.205^{18}$	$.137^{14}$	$.270^{8}$	$.300^{6}$	$.073^{5}$	$.127^{6}$	$.183^{18}$	7
TT-20	$.444^{11}$	$.413^{15}$	$.686^{22}$	$.184^{9}$	$.305^{20}$	$.227^{3}$	$.188^{23}$	$.116^{21}$	$.265^{11}$	$.292^{11}$	$.064^{12}$	$.099^{9}$	$.180^{21}$	12
TT-5-ED-F	$.463^{6}$	$.467^{12}$	$.641^{25}$	$.192^{7}$	$.305^{19}$	$.225^{5}$	$.228^{9}$	$.145^{11}$	$.278^{5}$	$.298^{9}$	$.066^{10}$	$.096^{11}$	$.194^{5}$	6
TT-10-ED-F	$.450^{10}$	$.410^{16}$	$.713^{21}$	$.177^{14}$	$.305^{21}$	$.212^{12}$	$.212^{13}$	$.124^{17}$	$.258^{15}$	$.290^{12}$	$.059^{17}$	$.085^{15}$	$.186^{13}$	13
TT-15-ED-F	$.420^{18}$	$.357^{21}$	$.749^{17}$	$.155^{21}$	$.319^{13}$	$.209^{15}$	$.195^{19}$	$.110^{23}$	$.250^{17}$	$.281^{15}$	$.056^{19}$	$.078^{17}$	$.183^{17}$	16
TT-20-ED-F	$.394^{22}$	$.325^{24}$	$.765^{14}$	$.144^{22}$	$.299^{24}$	$.199^{25}$	$.191^{22}$	$.109^{24}$	$.233^{23}$	$.256^{21}$	$.054^{20}$	$.064^{24}$	$.181^{20}$	24
TT-5-EC-F	$.405^{21}$	$.329^{23}$	$.775^{12}$	$.138^{23}$	$.325^{10}$	$.204^{20}$	$.209^{15}$	$.120^{18}$	$.242^{19}$	$.267^{19}$	$.054^{21}$	$.072^{19}$	$.189^{12}$	16
TT-10-EC-F	$.361^{27}$	$.265^{29}$	$.841^{7}$	$.098^{30}$	$.329^{5}$	$.199^{23}$	$.185^{25}$	$.094^{26}$	$.226^{25}$	$.240^{25}$	$.049^{24}$	$.060^{25}$	$.179^{22}$	25
TT-15-EC-F	$.361^{27}$	$.265^{29}$	$.841^{7}$	$.098^{30}$	$.329^{5}$	$.199^{23}$	$.185^{25}$	$.094^{26}$	$.226^{25}$	$.240^{25}$	$.049^{24}$	$.060^{25}$	$.179^{22}$	25
TT-20-EC-F	$.321^{32}$	$.226^{33}$	$.891^{3}$	$.082^{33}$	$.328^{8}$	$.187^{31}$	$.164^{30}$	$.078^{31}$	$.211^{31}$	$.220^{28}$	$.049^{27}$	$.053^{29}$	$.171^{27}$	32
C99	$.410^{19}$	$.360^{20}$	$.727^{19}$	$.155^{20}$	$.307^{16}$	$.204^{19}$	$.154^{31}$	$.091^{29}$	$.237^{21}$	$.288^{13}$	$.060^{16}$	$.090^{13}$	$.184^{15}$	20
UI	$.378^{25}$	$.285^{28}$	$.849^{6}$	$.134^{26}$	$.327^{9}$	$.200^{21}$	$.149^{32}$	$.085^{30}$	$.226^{27}$	$.264^{20}$	$.049^{26}$	$.070^{22}$	$.173^{26}$	27
MC	$.389^{23}$	$.324^{25}$	$.737^{18}$	$.156^{19}$	$.321^{12}$	$.210^{14}$	$.167^{28}$	$.096^{25}$	$.241^{20}$	$.280^{17}$	$.064^{13}$	$.084^{16}$	$.190^{11}$	18
DP	$.338^{31}$	$.252^{32}$	$.838^{9}$	$.108^{29}$	$.329^{7}$	$.187^{30}$	$.137^{33}$	$.069^{32}$	$.214^{29}$	$.243^{24}$	$.057^{18}$	$.072^{20}$	$.174^{24}$	29
PM	$.407^{20}$	$.487^{9}$	$.491^{31}$	$.168^{16}$	$.217^{33}$	$.191^{27}$	$.259^{4}$	$.167^{6}$	$.234^{22}$	$.147^{35}$	$.039^{33}$	$.044^{34}$	$.116^{34}$	28
NS-PM	$.422^{16}$	$.489^{8}$	$.505^{30}$	$.201^{3}$	$.242^{31}$	$.209^{16}$	$.235^{6}$	$.161^{7}$	$.264^{12}$	$.191^{30}$	$.046^{29}$	$.059^{27}$	$.137^{30}$	19
PS-PM	$.424^{15}$	$.502^{7}$	$.517^{29}$	$.167^{17}$	$.229^{32}$	$.206^{17}$	$.232^{7}$	$.147^{9}$	$.245^{18}$	$.150^{33}$	$.043^{31}$	$.050^{30}$	$.105^{35}$	23
COVER	$.365^{26}$	$.324^{26}$	$.623^{27}$	$.120^{27}$	$.261^{28}$	$.167^{32}$	$.277^{2}$	$.172^{4}$	$.208^{32}$	$.158^{32}$	$.044^{30}$	$.047^{32}$	$.143^{29}$	31
DIV	$.457^{8}$	$.384^{18}$	$.757^{15}$	$.167^{18}$	$.345^{1}$	$.222^{9}$	$.187^{24}$	$.115^{22}$	$.259^{14}$	$.286^{14}$	$.070^{8}$	$.093^{12}$	$.194^{4}$	11
HMM	$.353^{30}$	$.393^{17}$	$.459^{34}$	$.136^{24}$	$.177^{35}$	$.138^{35}$	$.208^{16}$	$.139^{13}$	$.184^{33}$	$.190^{31}$	$.037^{35}$	$.045^{33}$	$.134^{31}$	34

that are important for user satisfaction are either not considered or otherwise given disproportionate importance by the metric. Recent evaluation measures proposed for IR tasks such as the U-measure and time-biased gain take a more user-centric approach to guide the calculation of effectiveness scores.

Inspired by these user-centric evaluation measures, Section 7.3 presented the development of a novel evaluation measure for SCR that models user behaviour more explicitly than traditional measures. Our new measure, termed no-pain no-gain (NPNG), was derived from a probabilistic finite-state transition system, with nodes representing the different possible states a user may be in while interacting with a list of SCR results. A major advantage of NPNG over existing measures is that it can model a wide range of user browsing behaviours and is thus not biased towards a particular type of user.

Section 7.4 presented the results of SCR experiments that compared a large number of content structuring methods in a unstructured SCR task. The effectiveness scores obtained for these methods under a large set of traditional evaluation measures showed that a single measure cannot be used to appropriately determine the relative differences that may exist between structuring approaches. Instantiating NPNG with a set of specific and interpretative user models facilitated the analysis of the various SCR methods considered and permitted us to identify some of their strengths and weaknesses. The most effective structuring method for SCR, according to the user models considered, consists of re-ranking relatively short overlapping passages based on the relevance score of their documents.

Chapter 8

Conclusions

This chapter summarises the main contributions of this thesis to the state-of-the-art in SCR. Concrete answers to the research questions previously stated in Chapter 1 are provided in Section 8.2. While potential directions for future work are further discussed in Section 8.3.

8.1 Summary of main contributions

The following summarises the main contributions of the investigation and experimental work presented in this thesis across Chapters 5, 6, and 7.

Chapter 5

This chapter described a series of experiments conducted to determine whether prosodic/acoustic information can aid in the identification of informative terms occurring in spoken documents and passages, and whether this additional evidence could be used in combination with lexical information to provide better estimates of the relative importance that terms should be given in the SCR ranking process. The focus was on analysing whether prosodic/acoustic derived features extracted at the term level can be effectively combined with standard term frequency statistics, such as TF and IDF estimates, to improve existing ranking functions for SCR. For this purpose, acoustic descriptors of pitch (F_0), energy (E_{rms}), loudness (E_l), and duration (D) were extracted from the speech collections, standardised based on speaker information, and aligned against every term occurrence from the speech transcripts. Prominence scores for individual terms were then derived based on this set of acoustic descriptors by aggregating individual acoustic features across the occurrence, passage, or document levels. Experiments were then conducted to study the usefulness of these acoustically motivated scores for term weighting in a document and a passage retrieval task.

Initial experiments presented in Section 5.3 evaluated the document and passage retrieval effectiveness of various heuristic-based retrieval functions that sought to combine prominence scores with lexical scores derived from term and document frequency estimates. Some drawbacks of these approaches were pointed out within the framework of probabilistic relevance for IR. Comparisons of these heuristic approaches against a strong lexical-based Okapi BM25 baseline indicated that the acoustically modified term weights did not provide any significant gains over this baseline. This was also the case when prominence scores were calculated with combinations of energy, pitch, and duration features, which were previously proven to be more useful than scores calculated from a single acoustic descriptor.

Despite these negative results, the experiments with heuristic functions demonstrated that prominence information does encode, albeit to a small degree, useful information about the relative importance of terms in speech. In particular, these experiments proved that passages and documents can be ranked more effectively in order of relevance if such ranking is based on prominence scores alone, than if based on sub-optimal lexical-based weights. Thus, while information about prominent terms can potentially be exploited to quantify the significance of terms in speech content, such information is rather weak when compared to the information that can be inferred from term frequency statistics.

The analysis and experiments presented in Section 5.4 provided further insights into the relationship that exists between prominent and informative terms as predicted by frequency statistics. In particular, correlation and regression analyses were conducted to study how term weights derived from acoustic and lexical information may be related. The results from these analyses showed that acoustically emphasised terms tend to be those that occur rarely in the collections, thus supporting the observation of previous research that “new” or “unpredictable” terms are more likely to be made prominent in speech (Prince, 1981; Hirschberg and Grosz, 1992; Bell et al., 2009; Röhr, 2013).

Further experiments with binary classifiers demonstrated that acoustic/prosodic information encodes meaningful information about terms, which can even be used to distinguish between terms occurrences appearing in relevant and non-relevant contexts. Yet, in this sense, acoustic information proved to be significantly less effective than lexical information, and not to provide any additional complementary information to lexical information. Additional experiments with a learning-to-rank approach also showed that prominence information can only improve upon term weights based on frequency statistics when these are poorly estimated.

The experiments described in this thesis continued with previous investigations conducted by Silipo and Crestani (2000), Chen et al. (2001), and Guinaudeau and Hirschberg (2011) on the utilisation of prosodic prominence information for speech retrieval applications. The work presented in this thesis significantly expands this previous work in several ways. First, it re-validates the analysis made by Silipo and Crestani (2000) over a higher number of test collections which are orders of magnitude larger than the collection they used and contain a significantly larger number of queries and high-quality relevance assessments. Besides this, the analysis by Silipo and Crestani was based on human annotations of stress levels, while ours was based on automatically extracted features thus representing a more practical scenario.

The results reported in this thesis contrast with those reported by Silipo and Crestani, who found substantially stronger correlations between prominent and informative words. These differences are most likely attributed to: (i) possible differences between manual and automatic estimations of acoustic prominence scores; (ii) the nature of the data under analysis (telephone versus TV and lectures) as well as the length of the documents considered (sentences versus full-documents); (iii) and the consideration of stop words and lack of word length normalisation in the estimation of scores. With respect to (i), the signal processing algorithms used in this work for feature extraction, as well as the techniques adopted for feature aggregations, are far from being perfect and should by no means be considered as a reliable replacement of manual annotations of prosodic prominence. Further aspects related to this issue are discussed in Section 8.3. With respect to (iii), the fact that Silipo and Crestani included stop words in their analysis plus that they calculated prominence scores based on the sum of syllables in a word may have overestimated the correlation levels observed by the authors. This is because stop words, which are associated with low IR scores, tend to be pronounced significantly less prominently than content words, which are frequently associated with high IR scores. Also, words with more than one syllable are generally associated with more specific concepts, which are in turn less frequently used in spoken language and thus associated with high inverse document frequencies.

The contradictions in the observed levels of correlation between the experiments presented in this thesis and those conducted by Silipo and Crestani seem to suggest that prominence information may be less useful at distinguishing fine differences between important and unimportant content bearing words. While Silipo and Crestani found strong correlations between prominence and IR scores when including stop words in their analysis, our experiments show this correlation to significantly weaken when excluding stop words and other non-content bearing words from the analysis.

This thesis also re-validated previous experiments conducted by Chen et al. (2001) and Guinaudeau and Hirschberg (2011), who reported mixed results in a topic tracking and SCR tasks under a French and Mandarin Chinese spoken collections of broadcast news. In this respect, our results are consistent with those reported by Chen et al., who concluded that acoustically motivated weights can only provide non-significant improvements in retrieval effectiveness. Based on the research findings from this thesis, it is likely that the improvements observed by Guinaudeau and Hirschberg (2011) from using automatic prosodic prominence scores are due to an improper estimation of lexical-based term weights, possibly caused by differences between the external corpus they used to estimate the document frequency statistics and the one where retrieval models where finally applied.

Chapter 6

This chapter studied whether contextualisation techniques, whereby a passage is ranked based on the query terms it contains and also on those appearing elsewhere in its document, can help alleviate the detrimental effects that ASR errors can have on standard text retrieval methods when used to rank short pre-defined spoken passages. Experiments were conducted to quantify the ranking effectiveness of various contextualisation techniques for SCR in different conditions of speech recognition errors in the transcripts of spoken queries and documents.

An important finding from these experiments is that the use of contextual evidence can substantially increase spoken passage retrieval effectiveness when transcripts contain a large, albeit realistic, number of speech recognition errors. These improvements are so substantial that standard text-based retrieval methods with contextualisation can sometimes achieve levels of retrieval quality for extremely noisy ASR transcripts that are similar to those obtained for high-quality transcripts when context is not used.

Although previous research has demonstrated that expansion techniques based on pseudo-relevance feedback can also provide an effective solution to tackle the presence of ASR errors, the technique is only effective when applied to parallel corpora, which may be difficult to obtain for some application domains and spoken collections. Instead, this thesis demonstrates that contextualisation techniques can provide an additional complementary solution to standard expansion techniques, that can be applied without the need of external data.

Another important finding drawn is that the use of context becomes increasingly beneficial for SCR as the amount of ASR errors in the speech transcripts increase. Further experiments with techniques that adapt the amount of context considered for a passage automatically depending on its transcription quality provided no substantial benefits in retrieval effectiveness over non-adaptive techniques, but points out useful directions for future research.

This chapter also explored problems associated with the estimation of reliable term weights from relatively small speech collection, as well as the additional complications that this may bring into standard text-based retrieval methods when queries are verbose and spoken spontaneously. Solutions were proposed to mitigate this issue, these are, the use of within-query term frequency (QF) and exponential inverse document frequency (EIDF), which were shown to be effective for retrieving short spoken passages effectively in these particular adverse conditions.

Chapter 7

This chapter presented a novel framework for designing evaluation measures for SCR, based on finite-state probabilistic automaton which explicitly models the browsing processes that users carry out when interacting with a SCR system. The NPNG is a particular evaluation measure defined under this framework which estimates the accumulated gain discounted by estimates of vertical and horizontal browsing effort. Under the NPNG measure, and other traditional evaluation measures for SCR, this chapter then presented the results of a large study that evaluated the majority of content structuring strategies proposed in past SCR and tried to determine if there is a single "best" strategy.

A major contribution of the work presented in this chapter is the detailed analysis of current evaluation measures for SCR that is presented under the view of the gain-discount and gain-effort framework. This analysis permitted us to identify the limitations and biases of these measures, which served for the subsequent analysis of performance of structuring methods and inspired the development of the NPNG framework. The evaluation framework upon which the NPNG measure is based constitutes the second main contribution of the work described in Chapter 7.

The comparison experiments conducted helped us to understand several of the strengths and weaknesses of the different content structuring techniques that have been proposed in past SCR research, and confirmed the various biases that are present in traditional evaluation measures. These experiments also confirmed that a single evaluation measure is not sufficient for determining the relative performance of SCR systems, and that several measures capturing different user preferences and behaviours are instead needed.

8.2 Research questions revisited

This section returns to the research questions stated in Section 1.2, to discuss some of the concrete answers that this PhD dissertation has helped to provide.

RQ-1: *Can information about which prosodic units are made prominent in speech be combined with lexical information to derive improved term weighting schemes and retrieval functions that could enhance SCR effectiveness?*

The experiments and analysis conducted in Chapter 5 suggest that even when differences between important and non-important terms can be effectively distinguished by looking at different prosodic indicators in the speech signal, the information conveyed by these descriptors does not seem to provide any additional knowledge about terms to complement lexical information and help to improve the quality of classical lexical-based term weighting functions. Simple collection statistics based on the distribution of terms within and across documents appear to be sufficient to capture much of the information that the additional prosodic indicators explored in this thesis are able to express, at least when the latter are extracted automatically by using standard signal processing algorithms. When lexical-based weights are absent or poorly estimated, a combination of lexical and acoustic derived weights may provide improved retrieval effectiveness. Under these circumstances, it is difficult to envisage how prosodic information at the word level could be best integrated into existing lexical-based retrieval models to improve their retrieval capabilities.

RQ-2: *Can contextualisation techniques increase the robustness of standard text retrieval approaches to ASR errors when the retrieval units are made from short fragments of speech transcripts?*

The experiments described in Chapter 6 demonstrate that contextualisation techniques can indeed provide increased ranking effectiveness in SCR when pre-defined spoken passages are transcribed with extremely high word error rates (WERs). This is evidenced by the fact that retrieval effectiveness decreases more slowly as the number of ASR errors increase in the speech transcripts when passages are contextualised in the relevance scoring process. Standard text retrieval approaches that do not use contextualisation can suffer greatly from the deletion and substitution of query terms if retrieval passages are short. In these circumstances, relevant regions containing a high density of query terms relative to others become harder to distinguish from non-relevant regions as the volume of query term occurrences in the former reduces dramatically. Contextualisation techniques are particularly effective in these conditions since considering a passage within the context of its document can help to increase the volume of query term occurrences seen in the region to levels that can make it distinguishable from non-relevant sections. Considering context at different granularity levels, for instance local context around a passage and global context from its document, tends to provide the greatest robustness to ASR errors. Another observed trend is that standard text retrieval methods tend to perform more robustly against ASR errors when relying on increasing amounts of context, or equivalently, when longer pseudo-passages are considered.

RQ-3-A: *Can existing evaluation measures for SCR estimate levels of user satisfaction appropriately?*

The critical review of existing evaluation measures for SCR presented in Chapter 7 highlighted the various limitations that these measures have and showed that they are often unable to appropriately capture all aspects that users may consider important about the quality of a ranked list of SCR results. Factors such as the time a user may take to process the search results as well as the effects that advanced visualisation interfaces or VCR controls may have on this process, are currently beyond the scope of current evaluation measures. Within the dimensions of user satisfaction that current measures try to account for, are the amount of effort that users need to invest in vertical and horizontal browsing and the amount of relevant content that users can access to by inspecting the search results. The numeric estimates for these dimensions that existing measures try to calculate are often based on a set of simple deterministic rules and assumptions about user behaviour which result in scores that are unlikely to correlate well with true levels of user satisfaction for all types of users. Further, these estimates are commonly combined disproportionately across dimensions so that some dimensions tend to be over-emphasised by the measure. Consequently, the user satisfaction estimates that these measures calculate can only be considered appropriate for a very specific type of user.

RQ-3-B: *Can enhanced evaluation measures be developed to address the shortcomings of existing evaluation measures for SCR?*

The evaluation framework described in Chapter 7 provides a more explicit model of how users may process a ranked list of SCR results. Unlike existing evaluation measures for SCR which represent a specific type of user, the proposed evaluation framework is general enough to allow for the instantiation of different user models, each of which represents a particular type of user, with a specific type of browsing behaviour and set of preferences as to which factors may be more or less important for their satisfaction. The increased flexibility of this framework permits us to analyse the effectiveness of SCR systems on a per-user basis, and it is thus more useful to identify the limitations of SCR methods. In addition, the proposed framework can be instantiated with parameters learnt from real user interactions. Lastly, the proposed framework can be adapted to represent VCR-like controls more explicitly as well as specific browsing strategies that may be implemented by users when interacting with a playback tool.

RQ-3-C: *Which content structuring techniques are most effective in SCR in terms of maximising user satisfaction?*

The results of the experiments described in Section 7.4 suggest that there is no a single practical structuring technique that would provide maximal user satisfaction for every possible user. On one hand, structuring techniques that induce long passages, relative to the length of the regions known to contain relevant material, can provide improved ranking effectiveness by returning fewer search results and positioning passages with relevant content on top of irrelevant ones, thus reducing the amount of vertical browsing effort that users need to invest. On the other hand, techniques that induce short overlapping passages with arbitrary starting points are able to detect the locations of query phrases in a document, which can serve as useful indicators of where the relevant content may appear in the document. Returning these locations as search results can therefore reduce the amount of horizontal browsing effort that users need to invest. By contextualising passages based on the relevance scores of their container documents, structuring techniques that are effective at identifying accurate entry points can benefit from the improvements in ranking quality and stability seen from using long ranking units. This contextualised structuring technique based on fixed length overlapping windows can be even more effective than a technique based on manually defined units for patient users. Yet, there is still much to be gained from improving the automatic detection of entry points.

8.3 Future work

This section describes potential directions for future work based on the research described in this thesis.

Improved extraction and integration of prosodic prominence information

In order to obtain a prominence score associated with a single term, which may have multiple word occurrences in a single speech file, the techniques used in this thesis calculated multi-scale aggregations over low-level descriptors of energy, loudness, pitch, and duration estimates of the words, at different levels of content granularity. Aggregations were first applied within words and then across sets of words associated with the same indexing term to obtain term-level descriptors of prosodic prominence. The aggregation functions involved in this process consisted of simple descriptive statistics, such as averages and extremes. This multi-stage aggregation process is nicely illustrated by the diagrams from Figures 5.3, 5.6a, and 5.6b.

Given the high variability and complexity of the speech signal, a significant limitation of the multi-stage aggregation approach used in this thesis is that it is not immune to estimation errors that may be present in the low-level descriptors, especially when selecting extreme values (maximums and minimums) from within low-level contours. Considering the hierarchical multi-level nature of the feature extraction process, where features at higher-levels are "pooled" from those in lower-levels, there is scope for the application of deep neural networks (DNNs) to this problem. In particular, convolutional neural networks (CNNs) (LeCun et al., 1995) seem to suit these needs perfectly as this type of network architecture is frequently designed to process variable-length sequential data in a hierarchical fashion. A popular architecture consists of several convolutional layers followed by a max or average "pooling" operation.

CNNs have been demonstrated to be capable of learning effective high-quality representations from highly complex sequential data, such as images, video and audio, that result useful for a number of machine learning tasks, such as image and video classification (Krizhevsky et al., 2012; Karpathy et al., 2014), emotion detection (Ghayoumi and Bansal, 2016), and speech recognition (Abdel-Hamid et al., 2014). In relation to the extraction of prosodic prominence information, CNNs could be used for learning useful feature representations for spoken words given a set of low-level prosodic contours. These extracted features could then be fed into a learning-to-rank model to determine their usefulness. Besides the LambdaMART models explored in this thesis, other learning-to-rank models based on neural networks would better suit this integration. In particular, the recently proposed family of deep structured semantic models (DSSMs) (Huang et al., 2013; Shen et al., 2014b,a), based on several neural network architectures have already been successfully applied to document retrieval tasks.

Dealing with ASR errors in the speech transcripts

The experimental work presented in this thesis demonstrates that by considering longer versions of short spoken passages expanded with the contents from its document in the retrieval process, text retrieval methods can be made more robust to ASR errors. In addition to considering the terms appearing in the 1-best hypothesis of the ASR, contextualisation could be based on additional terms appearing in the N-best lists of each utterance, or in other lattice-based output representations produced by the ASR. Exploiting the term information from N-best lists has also been shown useful for reducing the effects of ASR errors in the past (Siegler et al., 1997; Tsuge et al., 2011), and could be integrated well with the contextualisation techniques described in this thesis.

In a positional-based technique, for instance, terms from the N-best hypotheses could be treated as appearing in the outskirts of the passage, with distances given by their recognition probabilities. Terms appearing in an hypothesis further down in the N-best rank could be positioned further away from the passage to reduce their contribution to the relevance scores in relation to more likely correct recognition hypothesis from the passage itself and its surroundings. An effective integration approach would therefore need to determine the right balance between 1-best context extracted from neighbouring passages and N-best context extracted from the contextualised passage. While 1-best context is less likely to contain ASR errors, it may be less topically related to the contextualised passage than this passage's N-best context, so finding the right trade-off is critical.

Besides contextualising with multiple ASR hypothesis, there is also scope to apply contextualisation techniques in combination with more standard expansion techniques based on pseudo-relevance feedback (PRF). Previous research in SCR has repeatedly shown that PRF-based expansion using external corpora can be also effective at reducing the effects produced by ASR errors (Singhal et al., 1999; Abberley et al., 1998; Gauvain et al., 1999; Johnson et al., 2000; Renals and Abberley, 2000). In this respect, an interesting research direction is to explore the extent to which contextualisation and PRF-based expansion techniques can complement each other to provide levels of robustness that would not be possible by using either technique in isolation.

Improved evaluation of SCR techniques

Our evaluation model could be extended to represent more realistic user models. In particular, instead of just assuming a single forward or backward horizontal browsing mode of “straight” playback, additional states could be added to represent other browsing strategies which users may adopt when using VCR-like controls. Besides modelling the fact that users may be less interested in relevant results that have already been seen at previous ranks, the NPNG model could be generalised to account for the fact that users will be less interested in assessing material they have already seen. If users are presented with a search result that points to a document they have previously browsed, it is likely that they will either skip this item or spend substantially less time on it. Such an event could be further conditioned on whether the user has previously encountered some relevant material in the document, as this may also affect their browsing behaviour.

Advancing the state-of-the-art in SCR evaluation necessitates the design and implementation of user studies to provide a better understanding of how users interact with SCR systems in practice. Such studies would provide invaluable information which would serve both for refining the structure and states of the NPNG model, as well as for estimating the parameters of the model based on data from real users.

In the experiments with structuring approaches presented in Chapter 7, NPNG was instantiated with four user models. Alternatively, the measure could be instantiated with a population of user models, as proposed by Carterette et al. (2011) and Clarke and Smucker (2014). The population of NPNG measures instantiated with each of these users could then be used to obtain a distribution of effectiveness values for a query for each of the SCR methods under evaluation. Finally, such distributions could be analysed to study how effectiveness might vary across users (effect sizes). In occasions, an SCR method with a lower average effectiveness and variance may be preferable over one with greater average effectiveness but greater variance, since this may indicate the former method to perform more consistently across users.

Improved jump-in time point detection

The experiments described in Chapter 7 evaluated SCR techniques which sought to determine the location within a spoken document where users should begin playback of the speech content when seeking for relevant information. In this regard, the basic approach implemented by these techniques consisted of segmenting the speech content into short passages, and then to identify passages associated with high relevance scores with respect to the query. While such a technique was shown to provide improved entry point accuracy compared to using longer passages or dynamically constructed passages in the experiments described in Chapter 7, future research should investigate alternative methods for jump-in point detection that could provide further reductions in horizontal browsing effort.

A possible direction for further research is to investigate whether prosodic/acoustic information could be useful for the task of entry point detection. In this respect, previous research in automatic topic segmentation has demonstrated that prosodic information can provide useful cues about the location of topic shifts in speech. Thus, a supervised approach whereby a machine learning model is trained with examples of entry points pointing to content that is relevant to some query may be worth exploring in future research. In addition to exploiting prosodic structure, prosodic prominence information at the word or sentence level could be potentially used to bias entry points to the locations of query terms that are detected to provide "novel" or "new" information to the contents being discussed in the speech recording.

Creation of test collections for SCR research

This thesis sought answers to the research questions proposed by carrying out an empirical investigation of SCR methods over two test collections: the BBC collection of broadcast TV content, and the SDPWS collection of lecture recordings. Despite the findings these experiments led to, the BBC and SDPWS collections present several limitations.

The BBC collection is an audiovisual collection where information may be conveyed by using the visual and spoken modalities. Even when the critical information may most frequently be conveyed through speech, the visual content should not be disregarded by retrieval methods if the objective is to improve the relevance of search results. The importance of visual information is noticeable in cases in which query-creators selected keywords with an aim to identify visual concepts that were relevant to their information needs. The fact that some of these topics may implicitly rely on visual information suggest that BBC topics may not be most appropriate for the investigation of prosody-based enhanced SCR methods, which will likely not have any effect on the ranking for visually-driven topics.

Since video was not recorded during the creation of the SDPWS collection, this collection is a speech-only collection and as such it may present a more suitable test-bed for the evaluation of the prosody-based enhanced methods studied in this thesis. Despite this, the size of the SDPWS collection is orders of magnitude smaller than most test collection used in IR research. Using small collections for IR research can have a negative effect on the reliability of the experiments conducted and the generalisability of their results.

For all reasons explained above, future research in SCR must invest in the creation of large collections of speech recordings that would permit a more direct study of the problematics investigated in this thesis. Such test collection should ideally have the following characteristics: (i) it should contain a relatively large number of spoken documents; (ii) if containing audiovisual documents, the information of importance should be within the audio track; (iii) it should contain a large number of topics targeting the spoken information contained in the documents; and (iv) it should contain examples of relevant passages with precise time-boundaries to allow for the investigation of content structuring and jump-in point detection approaches.

Appendices

Appendix A

List of publications

The following list contains the publications derived from or related to this PhD:

1. Racca, D. N. and Jones, G. J. F. (2016a). DCU at the NTCIR-12 SpokenQuery&Doc-2 task. In *Proceedings of the 12th NTCIR Conference on Evaluation of Information Access Technologies (NTCIR-12)*, pages 180–185, Tokyo, Japan

2. Racca, D. N. and Jones, G. J. F. (2016b). On the effectiveness of contextualisation techniques in spoken query spoken content retrieval. In *Proceedings of the 39th Annual International Conference on Research and Development in Information Retrieval (ACM SIGIR 2016)*, pages 933–936, Pisa, Italy

3. Racca, D. N. and Jones, G. J. F. (2015a). Evaluating Search and Hyperlinking: An example of the design, test, refine cycle for metric development. In *Proceedings of the MediaEval 2015 Multimedia Benchmark Workshop*, Wurzen, Germany

4. Racca, D. N. and Jones, G. J. F. (2015b). Incorporating prosodic prominence evidence into term weights for spoken content retrieval. In *Proceedings of the 16th International Conference on Spoken Language Processing (INTERSPEECH 2015)*, pages 1378–1382, Dresden, Germany

5. Racca, D. N., Eskevich, M., and Jones, G. J. F. (2014). DCU search runs at MediaEval 2014 Search and Hyperlinking. In *Working Notes Proceedings of the MediaEval 2014 Multimedia Benchmark Workshop*, Barcelona, Catalunya, Spain

6. Racca, D. N. and Jones, G. J. F. (2014). DCU at the NTCIR-11 SpokenQuery&Doc task. In *Proceedings of the 11th NTCIR Conference on Evaluation of Information Access Technologies (NTCIR-11)*, pages 376–383, Tokyo, Japan

7. Eskevich, M., Aly, R., Racca, D. N., Chen, S., and Jones, G. J. F. (2015). SAVA at MediaEval 2015: Search and anchoring in video archives. In *Working Notes Proceedings of the MediaEval 2015 Multimedia Benchmark Workshop*, Wurzen, Germany

8. Eskevich, M., Aly, R., Racca, D. N., Ordelman, R., Chen, S., and Jones, G. J. F. (2014). The Search and Hyperlinking task at MediaEval 2014. In *Working Notes Proceedings of the MediaEval 2014 Multimedia Benchmark Workshop*, Barcelona, Spain

9. Chen, S., Curtis, K., Racca, D. N., Zhou, L., Jones, G. J. F., and O'Connor, N. E. (2015). DCU ADAPT@ TRECVID 2015: Video hyperlinking task. In *Proceedings of TRECVID 2015*, Gaithersburg, MD, USA

Appendix B

Index Similarity Metrics

This appendix describes the various index similarity metrics used in this thesis, specifically, unique term error rate (UTER), term error rate (TER) (Johnson et al., 1999a), binary index accuracy (BIA), and ranked index accuracy (RIA) (van der Werff and Heeren, 2007). These metrics are calculated for each automatic transcript index against the reference index. The reference index is the index that results from indexing the ASR transcripts of a speech collection, while the reference index is the one that results from indexing the reference transcripts.

TER (Johnson et al., 1999a) is calculated as the sum of the absolute term frequency differences between the reference and hypothesised documents, divided by the length of the reference document, as shown in Equation B.1,

$$TER = \frac{\sum_i |ref_i - hyp_i|}{\sum_i ref_i} \tag{B.1}$$

where ref_i and hyp_i denote the frequency counts of term i in the reference and hypothesis transcripts respectively. Thus, a TER of 0 indicates that the ASR index is an exact representation of the reference index, whereas a TER of 1 indicates that there are as many recognition errors as term occurrences contained in the reference document. As opposed to TER, UTER disregards term counts and puts more weight on presence and absence errors which may be arguably more problematic for SCR applications. UTER can be calculated as shown in Equation B.1, where ref_i and hyp_i are binary values indicating the presence (1) or absence (0) of term i in the reference or hypothesis documents respectively.

An issue with measures such as TER and UTER is that they can acquire values higher than 1 if the hypothesis contains a large number of insertion errors. BIA (van der Werff and Heeren, 2007) solves this problem by calculating the product between the fraction of unique terms from the reference found in the hypothesis document (recall) and the fraction of unique terms from the hypothesis found in the reference document (precision).

BIA can then be calculated as shown in Equation B.2,

$$BIA = \frac{|ref \cap hyp|}{|ref|} \frac{|ref \cap hyp|}{|hyp|} \tag{B.2}$$

where ref and hyp denote the set of terms contained by the reference and hypothesis transcripts. Finally, RIA extends BIA to consider term and document frequencies, and is calculated as the cosine similarity between the normalised TF-IDF vector representations of the reference and hypothesised documents, shown in Equation B.3.

$$RIA = \frac{\boldsymbol{ref} \cdot \boldsymbol{hyp}}{|\boldsymbol{ref}|\,|\boldsymbol{hyp}|} \tag{B.3}$$

By constrast to other measures, RIA considers the relative importance of terms as assigned by a retrieval system, effectively down-weighting the contribution of highly frequent terms that are commonly less useful for retrieval applications.

Appendix C

LambdaMART

LambdaMART is based on LambdaRank, which is in turn based on RankNet (Burges et al., 2011). All these models see the ranking problem as a classification task, where the goal is to determine the order in which a pair of documents (d_i, d_j) should be ranked for a given query. The cost function that these methods seek to optimise during training is designed to capture the magnitude of the errors present in a given ranking of documents, while to be differentiable so that stochastic gradient descent optimisation can be used to adjust the model's parameters.

If $s_i = f(d_i)$ and $s_j = f(d_j)$ denote respectively the scores produced by a ranking function $f(x)$ for documents d_i and d_j with $s_i > s_j$, then the pairwise error in RankNet is calculated as shown in Equation C.1,

$$E = \frac{1}{2}(1 - S_{ij})\sigma(s_i - s_j) + \log(1 + e^{-\sigma(s_i - s_j)}) \tag{C.1}$$

where σ is a scaling constant, and S_{ij} is either 1, -1, or 0, if d_i is deemed, respectively, more relevant than d_j, less relevant than d_j, or equally relevant than d_j. RankNet tries to minimise the sum of pairwise errors that are present in a ranked list of results. Hence, the cost or error associated with a relevant document ranked lower than non-relevant documents increments with the depth at which the relevant document is ranked. In the original implementation of RankNet, the underlying machine learning model used was a neural network, whose weights w_k were updated via gradient descent optimisation. The gradients of a pairwise error with respect to the model's weights, $\frac{\partial E}{\partial w_k}$, can be expressed as the difference in the gradients of the documents' scores multiplied by a scalar which reflects the magnitude of the pairwise error, as shown in Equation C.2.

$$\frac{\partial E}{\partial w_k} = \sigma \left(\frac{1}{2}(1 - S_{ij}) - \frac{1}{1 + e^{\sigma(s_i - s_j)}} \right) \left(\frac{\partial s_i}{\partial w_k} - \frac{\partial s_j}{\partial w_k} \right) = \lambda_{ij} \left(\frac{\partial s_i}{\partial w_k} - \frac{\partial s_j}{\partial w_k} \right) \tag{C.2}$$

While minimising the number and magnitude of pairwise errors is likely to produce a model that can improve the quality of a given ranked list, the cost function used in RankNet still gives increased importance to relevant documents located at the lower-ends of the ranking. This goes against most traditional IR evaluation measures, such as MAP or NDCG, which pay more importance to documents ranked at the top of the ranked list. LambdaRank tackles this problem by modifying Equation C.2 to be more sensitive to the amount of change in the IR measure under consideration that results from swapping the ranks of two erroneously ordered documents. Specifically, LambdaRank redefines λ_{ij} in Equation C.2, to the form shown in Equation C.3,

$$\lambda_{ij} = \frac{-\sigma}{1 + e^{\sigma(s_i - s_j)}} |\Delta_{MAP}| \tag{C.3}$$

where $|\Delta_{MAP}|$ denotes the difference in MAP (or any other IR measure) that results from swapping the ranks of d_i and d_j in the ranked list for a query. Finally, for a given document d_k, the summation of pairwise lambdas involving d_k in a ranked list is calculated as shown in Equation C.4.

$$\lambda_k = \sum_j \lambda_{kj} - \sum_i \lambda_{ik} \tag{C.4}$$

LambdaMART combines ideas from both LambdaRank and Gradient Boosted Regression Trees (GBRT) (Friedman, 2001) also known as Multiple Additive Regression Trees (MART). The latter are based on the more general Gradient Boosting (Friedman, 2001) framework for training an ensemble $F(\vec{x}_i)$ of base-learner models in an iterative fashion, so that they minimise an arbitrary differentiable loss function $L(y_i, F(\vec{x}_i))$ given training data $\{(\vec{x}_i, y_i)\}_{i=1}^{M}$. Given an initial base-learner model f_0, the gradient boosting algorithm augments the ensemble at iteration N with a new base-learner model f_N that seeks to correct the mistakes made by the rest of the models in the ensemble. The output produced by the ensemble for input $\vec{x}_i$ at iteration N, $F_N(\vec{x}_i)$, is the weighted average of the outputs of its base models shown in Equation C.5.

$$F_N(\vec{x}_i) = \sum_{n=0}^{N} \alpha_n f_n(\vec{x}_i) \tag{C.5}$$

The next base-learner of the ensemble f_{N+1} is then constructed so that the overall error of the ensemble decreases when including f_{N+1}. This corresponds to finding a base-learner that is maximally correlated with the negative of the gradient of the ensemble's error. In the case of MART, where regression trees are used as base-learners, the next regression tree f_{N+1} is then trained such that its predictions are strongly correlated with the amounts $\hat{y}_i$ for input vectors $\vec{x}_i$, defined as shown in Equation C.6.

$$\hat{y}_i = -\frac{\partial L(y_i, F_N(\vec{x}_i))}{\partial F_N(\vec{x}_i)} \tag{C.6}$$

LambdaMART is a MART model in which the regression trees in the ensemble are trained with the lambda values, λ_i, as targets. In other words, a LambdaMART model is a MART model in which $\hat{y}_i = \lambda_i$, and λ_i is defined as in Equation C.4.

Appendix D

Coordinate Ascent Optimisation

D.1 Line Search

Let $p = \langle p_1, p_2, \ldots, p_n \rangle$ be the vector of parameters that we want to optimise and $\theta = \langle \theta_1, \theta_2, \ldots, \theta_n \rangle$ an initial parameter configuration. For a particular parameter p_i that can accept values in some interval $\alpha = [x, y]$, a line search is performed by evaluating the objective function at M distinct values of p_i while the values of the rest of the parameters are kept fixed. The M values are sampled equidistant in α and initially centred around θ_i. At each subsequent iteration of the algorithm, the size of the search interval α is reduced by a factor $0 < r < 1$ and the value of p_i that best maximises the objective function so far is chosen as the next point for centring the following M samples that are taken from α. This procedure is repeated for p_i until: (i) the size of α becomes smaller than some ϵ; (ii) a maximum number *maxit* of iterations have been performed; or (iii) the optimal value of p_i remains the same after *minit* iterations. In our implementation of line search, we set $M = 20$, $maxit = 30$ and $minit = 5$. Additionally, we set $\epsilon = 0.01$ and $r = 0.8$ for parameters that can take values in $\mathbb{R}$ while for those that can only take values in $\mathbb{N}$ we set $\epsilon = 1$ and reduce the size of α by 1 at every iteration. In order to reduce the size of the search space for parameters in $\mathbb{R}$ we truncate their values to two decimal positions.

D.2 Promising Directions

A line search can be performed for every parameter in p to obtain an optimal configuration of values θ^*. The vector from θ to θ^* suggests a "promising" direction in the multidimensional parameter space, so we further perform an additional line search on this direction by modifying the values of all the parameters linearly from θ_i to θ_i^*. By doing this, we hope to explore interesting regions of the parameter space which may led us to find even better parameter configurations. The process of performing n one-dimensional line searches plus one final multi-dimensional line search in the promising direction is commonly referred to as an epoch. In our implementation, we perform up to a maximum of 10 epochs and stop searching when the process results in the same parameter configuration in two consecutive epochs.

Appendix E

Results of experiments with binary classifiers

Table E.1 presents the results of the SCR experiments described in 5.4.2, and originally reported in Racca and Jones (2015b), with a modified BM25 function that incorporates the predictions of a binary classifier trained to classify between "relevant" and "non-relevant" occurrences of query terms given acoustic features. The table reports mean average precision (MAP) for SCR results produced by a standard BM25 function (BM25), and the modified BM25 function that incorporates the classifier's predictions (PROS). The results of these experiments are for the manual (MAN) and MATCH document transcripts of the SDPWD2 collection and query sets SD2 and SQD1. MAP figures in bold in the table show statistically significant differences based on a paired t-test ($p < 0.05$).

Table E.1: Retrieval results of SCR experiments with a modified BM25 function that incorporates the predictions of a binary classifier trained with acoustic features to classify between "relevant" and "non-relevant" occurrences of query terms. These results were originally reported in Racca and Jones (2015b).

Transcript	**Query set**		b, k_1 set	**PROS**	**BM25**
	Train	Test	to best in	MAP	MAP
MAN	SD2	SQD1	train	**.200**	.156
			test	.234	.192
	SQD1	SD2	train	.305	**.428**
			test	.442	.445
MATCH	SD2	SQD1	train	.111	.109
			test	.134	.129
	SQD1	SD2	train	.248	.242
			test	.275	.266

Bibliography

Abberley, D., Kirby, D., Renals, S., and Robinson, T. (1999a). The THISL broadcast news retrieval system. In *Proceedings of the ESCA ETRW Workshop on Accessing Information in Spoken Audio*, pages 14–19, Cambridge, UK.

Abberley, D., Renals, S., and Cook, G. (1998). Retrieval of broadcast news documents with the THISL system. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 1998)*, pages 3781–3784, Seattle, Washington, USA.

Abberley, D., Renals, S., Ellis, D., and Robinson, T. (1999b). The THISL SDR system at TREC-8. In *NIST Special Publication 500-246: Proceedings of the 8th Text REtrieval Conference (TREC-8)*, pages 699–706, Gaithersburg, Maryland.

Abdel-Hamid, O., Mohamed, A.-r., Jiang, H., Deng, L., Penn, G., and Yu, D. (2014). Convolutional neural networks for speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(10):1533–1545.

Abdul-Jaleel, N., Allan, J., Croft, W. B., Diaz, F., Larkey, L., Li, X., Smucker, M. D., and Wade, C. (2004). UMass at TREC 2004: Novelty and HARD. In *NIST Special Publication: SP 500-26. Proceedings of the 13th Text Retrieval Conference (TREC 2004)*, Gaithersburg, Maryland.

Akiba, T., Aikawa, K., Itoh, Y., Kawahara, T., Nanjo, H., Nishizaki, H., Yasuda, N., Yamashita, Y., and Itou, K. (2008). Test collections for spoken document retrieval from lecture audio data. In *Proceedings of the 6th International Conference on Language Resources and Evaluation, (LREC 2008)*, pages 1572–1577, Marrakech, (Morocco).

Akiba, T., Nishizaki, H., Aikawa, K., Hu, X., Itoh, Y., Kawahara, T., Nakagawa, S., and Nanjo, H. (2013a). Overview of the NTCIR-10 SpokenDoc-2 task. In *Proceedings of the 10th NTCIR Conference on Evaluation of Information Access Technologies (NTCIR-10)*, pages 573–587, Tokyo, Japan.

Akiba, T., Nishizaki, H., Aikawa, K., Kawahara, T., and Tomoko, M. (2011). Overview of the IR for spoken documents task in NTCIR-9 workshop. In *Proceedings of the 9th*

NTCIR Workshop Meeting on Evaluation of Information Access Technologies, Information Retrieval, Question Answering, and Cross-Lingual Information Access (NTCIR-9), pages 223–235, Tokyo, Japan.

Akiba, T., Nishizaki, H., Nanjo, H., and Jones, G. J. F. (2014). Overview of the NTCIR-11 SpokenQuery&Doc task. In *Proceedings of the 11th NTCIR Conference on Evaluation of Information Access Technologies (NTCIR-11)*, pages 350–364, Tokyo, Japan.

Akiba, T., Nishizaki, H., Nanjo, H., and Jones, G. J. F. (2016). Overview of the NTCIR-12 SpokenQuery&Doc-2 task. In *Proceedings of the 12th NTCIR Conference on Evaluation of Information Access Technologies (NTCIR-12)*, pages 167–179, Tokyo, Japan.

Akiba, T., Takigami, T., Ohno, T., and Kase, K. (2013b). DTW-distance-ordered spoken term detection and STD-based spoken content retrieval: Experiments at NTCIR-10 SpokenDoc-2. In *Proceedings of the 10th NTCIR Conference on Evaluation of Information Access Technologies (NTCIR-10)*, pages 618–625, Tokyo, Japan.

Alink, W. and Cornacchia, R. (2011). Out-of-the-box strategy for Rich Speech Retrieval MediaEval 2011. In *Working Notes Proceedings of the MediaEval 2011 Multimedia Benchmark Workshop*, Pisa, Italy.

Allan, J. (2001). Perspectives on information retrieval and speech. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (ACM SIGIR 2001), Workshop on Information Retrieval Techniques for Speech Applications*, pages 1–10, New Orleans, LA, USA.

Allan, J. (2003). HARD track overview in TREC 2003: High accuracy retrieval from documents. In *NIST Special Publication: 500-255 Proceedings of the 12th Text REtrieval Conference (TREC-2003)*, pages 24–37, Gaithersburg, Maryland.

Allan, J. (2004). HARD track overview in TREC 2004: High accuracy retrieval from documents. In *NIST Special Publication: SP 500-26. Proceedings of the 13th Text Retrieval Conference (TREC 2004)*, Gaithersburg, Maryland.

Aly, R., Eskevich, M., Ordelman, R., and Jones, G. J. F. (2013a). Adapting binary information retrieval evaluation metrics for segment-based retrieval tasks. Technical report, University of Twente.

Aly, R., Ordelman, R. J., Eskevich, M., Jones, G. J. F., and Chen, S. (2013b). Linking inside a video collection: what and how to measure? In *Proceedings of the 22nd International Conference on World Wide Web Companion*, pages 457–460, Rio de Janeiro, Brazil.

Aly, R., Verschoor, T., and Ordelman, R. (2011). UTwente does Rich Speech Retrieval at MediaEval 2011. In *Working Notes Proceedings of the MediaEval 2011 Multimedia Benchmark Workshop*, Pisa, Italy.

Arvola, P., Kekäläinen, J., and Junkkari, M. (2011). Contextualization models for XML retrieval. *Information Processing & Management*, 47(5):762–776.

Asahara, M. and Matsumoto, Y. (2003). IPADIC version 2.7. 0 user's manual. *Computational Linguistics Laboratory. Graduate School of Information Science. Nara Institute of Science and Technology.*

Association, I. P. (1999). *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet.* Cambridge University Press.

Atal, B. S. and Hanauer, S. L. (1971). Speech analysis and synthesis by linear prediction of the speech wave. *The Journal of the Acoustical Society of America*, 50(2B):637–655.

Awad, G., Butt, A., Fiscus, J., Joy, D., Delgado, A., Michel, M., Smeaton, A. F., Graham, Y., Kraaij, W., Quénot, G., et al. (2017). TRECVID 2017: Evaluating ad-hoc and instance video search, events detection, video captioning and hyperlinking. In *Proceedings of the 2017 TRECVID Workshop.*

Awad, G., Fiscus, J., Michel, M., Joy, D., Kraaij, W., Smeaton, A. F., Quénot, G., Eskevich, M., Aly, R., and Ordelman, R. (2016). TRECVID 2016: Evaluating video search, video event detection, localization, and hyperlinking. In *Proceedings of the 2016 TRECVID Workshop.*

Bartell, B. T., Cottrell, G. W., and Belew, R. K. (1994). Automatic combination of multiple ranked retrieval systems. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (ACM SIGIR 1994)*, pages 173–181, Dublin, Ireland.

Beeferman, D., Berger, A., and Lafferty, J. (1997). Text segmentation using exponential models. In *Second Conference on Empirical Methods in Natural Language Processing*, pages 35–46, Providence, Rhode Island.

Belkin, N. J., Kantor, P., Fox, E. A., and Shaw, J. A. (1995). Combining the evidence of multiple query representations for information retrieval. *Information Processing & Management*, 31(3):431–448.

Bell, A., Brenier, J. M., Gregory, M., Girand, C., and Jurafsky, D. (2009). Predictability effects on durations of content and function words in conversational English. *Journal of Memory and Language*, 60(1):92–111.

Bell, P., Gales, M., Hain, T., Kilgour, J., Lanchantin, P., Liu, X., McParland, A., Renals, S., Saz, O., Wester, M., et al. (2015). The MGB Challenge: Evaluating multi-genre broadcast media recognition. In *Proceedings of the 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU 2015)*, pages 687–693, Scottsdale, AZ, USA.

Bengio, Y., Ducharme, R., Vincent, P., and Jauvin, C. (2003). A neural probabilistic language model. *Journal of Machine Learning Research*, 3:1137–1155.

Bird, S. (2006). NLTK: the natural language toolkit. In *Proceedings of the COLING/ACL on Interactive presentation sessions*, pages 69–72, Sydney, Australia.

Bogert, B. P., Healy, M. J., and Tukey, J. W. (1963). The quefrency analysis of time series for echoes: Cepstrum, pseudo-autocovariance, cross-cepstrum and saphe cracking. In *Proceedings of the symposium on time series analysis*, volume 15, pages 209–243.

Boytsov, L., Belova, A., and Westfall, P. (2013). Deciding on an adjustment for multiplicity in IR experiments. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval (ACM SIGIR 2013)*, pages 403–412, Dublin, Ireland.

Brodersen, K. H., Ong, C. S., Stephan, K. E., and Buhmann, J. M. (2010). The balanced accuracy and its posterior distribution. In *Proceedings of the 20th International Conference on Pattern Recognition (ICPR 2010)*, pages 3121–3124, Istanbul, Turkey.

Brown, M. G., Foote, J. T., Jones, G. J. F., Spärck Jones, K., and Young, S. J. (1995). Automatic content-based retrieval of broadcast news. In *Proceedings of the 3rd ACM International Conference on Multimedia*, pages 35–43, Dallas, Texas, USA.

Brown, M. G., Foote, J. T., Jones, G. J. F., Spärck Jones, K., and Young, S. J. (1996). Open-vocabulary speech indexing for voice and video mail retrieval. In *Proceedings of the 4th ACM International Conference on Multimedia*, pages 307–316, Boston, Massachusetts, USA.

Buckley, C., Salton, G., Allan, J., and Singhal, A. (1994). Automatic query expansion using SMART: TREC-3. In *NIST Special Publication: 500-226 Overview of the 3rd Text REtrieval Conference (TREC-3)*, pages 69–80, Gaithersburg, Maryland.

Burges, C., Shaked, T., Renshaw, E., Lazier, A., Deeds, M., Hamilton, N., and Hullender, G. (2005). Learning to rank using gradient descent. In *Proceedings of the 22nd International Conference on Machine Learning*, pages 89–96, Bonn, Germany.

Burges, C., Svore, K., Bennett, P., Pastusiak, A., and Wu, Q. (2011). Learning to rank using an ensemble of lambda-gradient models. In *Proceedings of the Yahoo! Learning to Rank Challenge*, pages 25–35, Yahoo! Labs. Sunnyvale, CA, USA.

Burges, C. J. (2010). From RankNet to LambdaRank to LambdaMART: An overview. Technical report, Microsoft Research.

Büttcher, S., Clarke, C. L. A., and Lushman, B. (2006). Term proximity scoring for ad-hoc retrieval on very large text collections. In *Proceedings of the 29th Annual International*

ACM SIGIR Conference on Research and Development in Information Retrieval (ACM SIGIR 2006), pages 621–622, Seattle, WA, USA.

Callan, J. P. (1994). Passage-level evidence in document retrieval. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (ACM SIGIR 1994)*, pages 302–310, Dublin, Ireland.

Carmel, D., Maarek, Y. S., Mandelbrod, M., Mass, Y., and Soffer, A. (2003). Searching XML documents via XML fragments. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (ACM SIGIR 2003)*, pages 151–158, Toronto, ON, Canada.

Carmel, D., Shtok, A., and Kurland, O. (2013). Position-based contextualization for passage retrieval. In *Proceedings of the 22Nd ACM International Conference on Information & Knowledge Management*, pages 1241–1244, San Francisco, CA, USA.

Carterette, B. (2011). System effectiveness, user models, and user utility: A conceptual framework for investigation. In *Proceedings of the 34th International ACM SIGIR conference on Research and Development in Information Retrieval (ACM SIGIR 2011)*, pages 903–912, Beijing, China.

Carterette, B., Kanoulas, E., and Yilmaz, E. (2011). Simulating simple user behavior for system effectiveness evaluation. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, pages 611–620, Glasgow, Scotland, UK.

Chapelle, O. and Chang, Y. (2011). Yahoo! learning to rank challenge overview. In *Proceedings of the Yahoo! Learning to Rank Challenge*, pages 1–24, Yahoo! Labs. Sunnyvale, CA, USA.

Chapelle, O., Metlzer, D., Zhang, Y., and Grinspan, P. (2009). Expected reciprocal rank for graded relevance. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, pages 621–630, Hong Kong, China.

Chatzichristofis, S. A. and Arampatzis, A. (2010). Late fusion of compact composite descriptors for retrieval from heterogeneous image databases. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval (ACM SIGIR 2010)*, pages 825–826, Geneva, Switzerland.

Chelba, C., Hazen, T. J., and Saraclar, M. (2008). Retrieval and browsing of spoken content. *IEEE Signal Processing Magazine*, 25(3).

Chen, B., Wang, H.-M., and Lee, L.-S. (2001). Improved spoken document retrieval by exploring extra acoustic and linguistic cues. In *Proceedings of the 7th European Conference on Speech Communication and Technology*, pages 299–302, Aalborg, Denmark.

Chen, F. R. and Withgott, M. (1992). The use of emphasis to automatically summarize a spoken discourse. In *Proceedings of the 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 1992)*, pages 229–232, San Francisco, CA, USA.

Chen, G., Yilmaz, O., Trmal, J., Povey, D., and Khudanpur, S. (2013). Using proxies for oov keywords in the keyword search task. In *Proceedings of the 2013 IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU 2013)*, pages 416–421, Olomouc, Czech Republic.

Chen, K., Hasegawa-Johnson, M., Cohen, A., Borys, S., Kim, S.-S., Cole, J., and Choi, J.-Y. (2006). Prosody dependent speech recognition on radio news corpus of American English. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(1):232–245.

Chen, S., Curtis, K., Racca, D. N., Zhou, L., Jones, G. J. F., and O'Connor, N. E. (2015). DCU ADAPT@ TRECVID 2015: Video hyperlinking task. In *Proceedings of TRECVID 2015*, Gaithersburg, MD, USA.

Chen, S. F. and Goodman, J. (1996). An empirical study of smoothing techniques for language modeling. In *Proceedings of the 34th Annual Meeting on Association for Computational Linguistics*, pages 310–318, Santa Cruz, CA, USA.

Chen, S.-H., Yang, J.-H., Chiang, C.-Y., Liu, M.-C., and Wang, Y.-R. (2012). A new prosody-assisted Mandarin ASR system. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(6):1669–1684.

Chiu, C., Tripathi, A., Chou, K., Co, C., Jaitly, N., Jaunzeikare, D., Kannan, A., Nguyen, P., Sak, H., Sankar, A., Tansuwan, J., Wan, N., Wu, Y., and Zhang, X. (2017). Speech recognition for medical conversations. *arXiv preprint arXiv:1711.07274*.

Choi, F. Y. (2000). Advances in domain independent linear text segmentation. In *Proceedings of the 1st North American Chapter of the Association for Computational Linguistics Conference*, pages 26–33, Seattle, Washington.

Chowdhury, A., McCabe, M. C., Grossman, D., and Frieder, O. (2002). Document normalization revisited. In *Proceedings of the 25th Annual International Conference on Research and Development in Information Retrieval (ACM SIGIR 2002)*, pages 381–382, Tampere, Finland.

Christodoulides, G. and Avanzi, M. (2014). An evaluation of machine learning methods for prominence detection in French. In *Proceedings of the 15th Annual Conference of the International Speech Communication Association (INTERSPEECH 2014)*, pages 116–119, Singapore.

Cieri, C., Graff, D., Liberman, M., Martey, N., Strassel, S., et al. (1999). The TDT-2 text and speech corpus. In *Proceedings of the DARPA Broadcast News Workshop*, pages 57–60, Herndon, Virginia.

Clarke, C. L., Cormack, G. V., Kisman, D. I., and Lynam, T. R. (2000a). Question answering by passage selection (MultiText experiments for TREC-9). In *NIST Special Publication 500-249: Proceedings of the 9th Text REtrieval Conference (TREC-9)*, pages 673–683, Gaithersburg, Maryland.

Clarke, C. L., Cormack, G. V., and Tudhope, E. A. (2000b). Relevance ranking for one to three term queries. *Information Processing & Management*, 36(2):291–311.

Clarke, C. L. and Smucker, M. D. (2014). Time well spent. In *Proceedings of the 5th Information Interaction in Context Symposium*, pages 205–214, Regensburg, Germany.

Cleverdon, C. (1962). *Report on the Testing and Analysis of an Investigation into the Comparative Efficiency of Indexing Systems*, volume 1. College of Aeronautics.

Cleverdon, C. W., Mills, J., and Keen, E. M. (1966). Factors determining the performance of indexing systems. *ASLIB Cranfield Research Project*, 1.

Cobârzan, C. and Schoeffmann, K. (2014). How do users search with basic HTML5 video players? In *Proceedings of the 20th Anniversary International Conference on Multimedia Modeling (MMM 2014)*, pages 109–120, Dublin, Ireland.

Collins-Thompson, K., Macdonald, C., Bennett, P., Diaz, F., and Voorhees, E. M. (2015). TREC 2014 web track overview. In *NIST Special Publication: SP 500-308. Proceedings of the 23th Text REtrieval Conference (TREC 2014)*, Gaithersburg, Maryland.

Cortes, C. and Vapnik, V. (1995). Support vector machine. *Machine Learning*, 20(3):273–297.

Craswell, N., Robertson, S., Zaragoza, H., and Taylor, M. (2005). Relevance weighting for query independent evidence. In *Proceedings of the 28th Annual International Conference on Research and Development in Information Retrieval (ACM SIGIR 2005)*, pages 416–423, Salvador, Brazil.

Crestani, F. (2001). Towards the use of prosodic information for spoken document retrieval. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (ACM SIGIR 2001)*, pages 420–421, New Orleans, USA.

Crestani, F., Sanderson, M., Theophylactou, M., and Lalmas, M. (1997). Short queries, natural language and spoken documents retrieval: Experiments at Glasgow University. In *NIST Special Publication: 500-240. Proceedings of the 6th Text REtrieval Conference (TREC-6)*, pages 667–686, Gaithersburg, Maryland.

Crockford, C. and Agius, H. (2006). An empirical investigation into user navigation of digital video using the VCR-like control set. *International Journal of Human-Computer Studies*, 64(4):340–355.

Cutler, A., Dahan, D., and Van Donselaar, W. (1997). Prosody in the comprehension of spoken language: A literature review. *Language and Speech*, 40(2):141–201.

Davis, S. and Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(4):357–366.

de Kretser, O. and Moffat, A. (1999). Effective document presentation with a locality-based similarity heuristic. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (ACM SIGIR 1999)*, pages 113–120, Berkeley, California, USA.

De Vries, A. P., Kazai, G., and Lalmas, M. (2004). Tolerance to irrelevance: A user-effort oriented evaluation of retrieval systems without predefined retrieval unit. In *Proceedings of the 7th International Conference on Computer-Assisted Information Retrieval (Recherche d'Information et ses Applications) - (RIAO 2004)*, pages 463–473, Avignon, France.

Deng, L., Yu, D., et al. (2014). Deep learning: Methods and applications. *Foundations and Trends in Signal Processing*, 7(3–4):197–387.

Eisenstein, J. and Barzilay, R. (2008). Bayesian unsupervised topic segmentation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2008)*, pages 334–343, Waikiki, Honolulu, Hawaii.

Enarvi, S., Smit, P., Virpioja, S., and Kurimo, M. (2017). Automatic speech recognition with very large conversational Finnish and Estonian vocabularies. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(11):2085–2097.

Eskevich, M. (2014). *Towards effective retrieval of spontaneous conversational spoken content*. PhD thesis, Dublin City University.

Eskevich, M., Aly, R., Ordelman, R., Chen, S., and Jones, G. J. F. (2013a). The Search and Hyperlinking Task at MediaEval 2013. In *Working Notes Proceedings of the MediaEval 2013 Multimedia Benchmark Workshop*, Barcelona, Catalunya, Spain.

Eskevich, M., Aly, R., Ordelman, R., Chen, S., and Jones, G. J. F. (2013b). The Search and Hyperlinking task at MediaEval 2013. In *Working Notes Proceedings of the MediaEval 2013 Multimedia Benchmark Workshop*, Barcelona, Spain.

Eskevich, M., Aly, R., Racca, D. N., Chen, S., and Jones, G. J. F. (2015). SAVA at MediaEval 2015: Search and anchoring in video archives. In *Working Notes Proceedings of the MediaEval 2015 Multimedia Benchmark Workshop*, Wurzen, Germany.

Eskevich, M., Aly, R., Racca, D. N., Ordelman, R., Chen, S., and Jones, G. J. F. (2014). The Search and Hyperlinking task at MediaEval 2014. In *Working Notes Proceedings of the MediaEval 2014 Multimedia Benchmark Workshop*, Barcelona, Spain.

Eskevich, M. and Jones, G. J. F. (2011a). DCU at MediaEval 2011: Rich Speech Retrieval (RSR). In *Working Notes Proceedings of the MediaEval 2011 Multimedia Benchmark Workshop*, Pisa, Italy.

Eskevich, M. and Jones, G. J. F. (2011b). DCU at the NTCIR-9 SpokenDoc passage retrieval task. In *Proceedings of the 9th NTCIR Workshop Meeting on Evaluation of Information Access Technologies, Information Retrieval, Question Answering, and Cross-Lingual Information Access (NTCIR-9)*, pages 257–260, Tokyo, Japan.

Eskevich, M. and Jones, G. J. F. (2013a). DCU at NTCIR-10 SpokenDoc2 passage retrieval task. In *Proceedings of the 10th NTCIR Conference on Evaluation of Information Access Technologies (NTCIR-10)*, pages 640–607, Tokyo, Japan.

Eskevich, M. and Jones, G. J. F. (2013b). Time-based segmentation and use of jump-in points in DCU search runs at the Search and Hyperlinking Task at MediaEval 2013. In *Working Notes Proceedings of the MediaEval 2013 Multimedia Benchmark Workshop*, Barcelona, Catalunya, Spain.

Eskevich, M., Jones, G. J. F., Aly, R., Ordelman, R. J., Chen, S., Nadeem, D., Guinaudeau, C., Gravier, G., Sébillot, P., de Nies, T., Debevere, P., Van de Walle, R., Galušcáková, P., Pecina, P., and Larson, M. (2013c). Multimedia information seeking through search and hyperlinking. In *Proceedings of the 3rd ACM Conference on International Conference on Multimedia Retrieval*, pages 287–294, Dallas,Texas, USA.

Eskevich, M., Jones, G. J. F., Chen, S., Aly, R., Ordelman, R., and Larson, M. (2012a). Search and Hyperlinking task at MediaEval 2012. In *Working Notes Proceedings of the MediaEval 2012 Multimedia Benchmark Workshop*, Pisa, Italy.

Eskevich, M., Jones, G. J. F., Wartena, C., Larson, M., Aly, R., Verschoor, T., and Ordelman, R. (2012b). Comparing retrieval effectiveness of alternative content segmentation methods for internet video search. In *Proceedings of the 10th International Workshop on Content-Based Multimedia Indexing (CBMI 2012)*, pages 223–228, Annecy, France.

Eskevich, M., Magdy, W., and Jones, G. J. F. (2012c). New metrics for meaningful evaluation of informally structured speech retrieval. In *Proceedings of the 34th European Conference on Information Retrieval Research (ECIR 2012)*, pages 170–181, Barcelona, Spain.

Eyben, F. (2016). *Real-time speech and music classification by large audio feature space extraction*. Springer.

Eyben, F., Weninger, F., Gross, F., and Schuller, B. (2013). Recent developments in openSMILE, the Munich open-source multimedia feature extractor. In *Proceedings of the 21st ACM International Conference on Multimedia (MM 2013)*, pages 835–838, Barcelona, Spain.

Fox, E. A. and Shaw, J. A. (1993). Combination of multiple searches. In *NIST Special Publication: 500-215 The Second Text REtrieval Conference (TREC-2)*, pages 243–252, Gaithersburg, Maryland.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232.

Fuhr, N., Gövert, N., Kazai, G., and Lalmas, M. (2002). INEX: Initiative for the evaluation of XML retrieval. In *Proceedings of the ACM SIGIR Workshop on XML and Information Retrieval*, pages 1–9, Tampere, Finland.

Fuhr, N. and Lalmas, M. (2007). Advances in XML retrieval: The INEX initiative. In *Proceedings of the 2006 International Workshop on Research Issues in Digital Libraries*, Kolkata, India.

Furui, S. (2007). Recent advances in automatic speech summarization. In *Proceedings of the (RIAO 2007) Conference on Large Scale Semantic Access to Content (Text, Image, Video, and Sound)*, pages 90–101, Pittsburgh, Pennsylvania, USA.

Gales, M. J. (1998). Maximum likelihood linear transformations for HMM-based speech recognition. *Computer Speech & Language*, 12(2):75–98.

Galuščáková, P., Pecina, P., and Hajič, J. (2012). Penalty functions for evaluation measures of unsegmented speech retrieval. In *Proceedings of the 3rd International Conference of the Cross-Language Evaluation Forum for European Languages (CLEF 2012): Information Access Evaluation. Multilinguality, Multimodality, and Visual Analytics*, pages 100–111, Rome, Italy.

Galuščáková, P. and Pecina, P. (2012). CUNI at Mediaeval 2012 Search and Hyperlinking task. In *Working Notes Proceedings of the MediaEval 2011 Multimedia Benchmark Workshop*, Pisa, Italy.

Galuščáková, P. and Pecina, P. (2014a). CUNI at MediaEval 2014 Search and Hyperlinking task: Search task experiments. In *Working Notes Proceedings of the MediaEval 2014 Multimedia Benchmark Workshop*, Barcelona, Spain.

Galuščáková, P. and Pecina, P. (2014b). Experiments with segmentation strategies for passage retrieval in audio-visual documents. In *Proceedings of the 4th ACM International Conference on Multimedia Retrieval (ICMR 2014)*, pages 217–224, Glasgow, Scotland.

Garofolo, J. S., Auzanne, C. G. P., and Voorhees, E. M. (2000). The TREC spoken document retrieval track: A success story. In *Proceedings of the RIAO Conference on Content-Based Multimedia Information Access*, pages 1–20, Paris, France.

Garofolo, J. S., Voorhees, E. M., Auzanne, C. G., Stanford, V. M., and Lund, B. A. (1998). 1998 TREC-7 spoken document retrieval track overview and results. In *NIST Special Publication: 500-242. Proceedings of the 7th Text REtrieval Conference (TREC-7)*, pages 79–90, Gaithersburg, Maryland.

Gauvain, J.-L. (2010). The Quaero program: Multilingual and multimedia technologies. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT) 2010*, Paris, France.

Gauvain, J.-L., de Kercadio, Y., Lamel, L., and Adda, G. (1999). The LIMSI SDR system for TREC-8. In *NIST Special Publication 500-246: Proceedings of the 8th Text REtrieval Conference (TREC-8)*, pages 475–482, Gaithersburg, Maryland.

Gauvain, J.-L., Lamel, L., and Adda, G. (1998). Partitioning and transcription of broadcast news. In *Proceedings of the 5th International Conference on Spoken Language Processing (ICSLP 1998)*, pages 1335–1338, Sydney, Australia.

Gauvain, J.-L., Lamel, L., and Adda, G. (2002). The LIMSI broadcast news transcription system. *Speech Communication*, 37(1):89–108.

Gauvain, J.-L., Lamel, L., Barras, C., Adda, G., and de Kercadio, Y. (2000). The LIMSI SDR system for TREC-9. In *NIST Special Publication 500-249: Proceedings of the 9th Text REtrieval Conference (TREC-9)*, pages 335–360, Gaithersburg, Maryland.

Ghayoumi, M. and Bansal, A. K. (2016). Emotion in robots using convolutional neural networks. In *Proceedings of the 8th International Conference on Social Robotics (ICSR 2016)*, pages 285–295, Kansas City, USA.

Glavitsch, U. and Schäuble, P. (1992). A system for retrieving speech documents. In *Proceedings of the 15th Annual International Conference on Research and Development in Information Retrieval (ACM SIGIR 1992)*, pages 168–176, Copenhagen, Denmark.

Godfrey, J. J., Holliman, E. C., and McDaniel, J. (1992). Switchboard: Telephone speech corpus for research and development. In *Proceedings of the 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 1992)*, pages 517–520, San Francisco, CA, USA.

Goldwater, S., Jurafsky, D., and Manning, C. D. (2010). Which words are hard to recognize? prosodic, lexical, and disfluency factors that increase speech recognition error rates. *Speech Communication*, 52(3):181–200.

Gövert, N., Abolhassani, M., Fuhr, N., and Großjohann, K. (2002). Content-oriented XML retrieval with HyRex. In *Proceedings of the First Workshop of the INitiative for the Evaluation of XML Retrieval (INEX)*, pages 26–32, Schloss Dagstuhl, Germany.

Guinaudeau, C. and Hirschberg, J. (2011). Accounting for prosodic information to improve ASR-based topic tracking for TV broadcast news. In *Proceedings of the 12th International Conference on Spoken Language Processing (INTERSPEECH 2011)*, pages 1401–1404, Florence, Italy.

Haeb-Umbach, R. and Ney, H. (1992). Linear discriminant analysis for improved large vocabulary continuous speech recognition. In *Proceedings of the 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 1992)*, pages 13–16, San Francisco, CA, USA.

Harman, D. (1993). Overview of the second text retrieval conference TREC-2. In *NIST Special Publication: 500-215 The Second Text REtrieval Conference (TREC-2)*, pages 1–20, Gaithersburg, Maryland.

Harter, S. P. (1975). A probabilistic approach to automatic keyword indexing. Part II. An algorithm for probabilistic indexing. *Journal of the Association for Information Science and Technology*, 26(5):280–289.

Hearst, M. A. (1993). TextTiling: A quantitative approach to discourse segmentation. Technical report, University of California, Berkeley.

Hearst, M. A. (1994). *Context and structure in automated full-text information access*. PhD thesis, University of California, Berkeley.

Hearst, M. A. (1997). TextTiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1):33–64.

Hearst, M. A. and Plaunt, C. (1993). Subtopic structuring for full-length document access. In *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (ACM SIGIR 1993)*, pages 59–68, New York, NY, USA.

Hiemstra, D., Ordelman, R., Aly, R., van der Werff, L., and de Jong, F. (2006). Speech retrieval experiments using XML information retrieval. In *Working Notes of the 7th Workshop of the Cross-Language Evaluation Forum (CLEF 2006): Evaluation of Multilingual and Multi-modal Information Retrieval*, pages 770–777, Alicante, Spain.

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A.-r., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N., et al. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6):82–97.

Hinton, G. E., Osindero, S., and Teh, Y.-W. (2006). A fast learning algorithm for deep belief nets. *Neural Computation*, 18(7):1527–1554.

Hirschberg, J. (2002). Communication and prosody: Functional aspects of prosody. *Speech Communication*, 36(1):31–43.

Hirschberg, J. and Grosz, B. (1992). Intonational features of local and global discourse. In *Proceedings of the Workshop on Speech and Natural Language*, pages 441–446, Harriman, New York.

Huang, P.-S., He, X., Gao, J., Deng, L., Acero, A., and Heck, L. (2013). Learning deep structured semantic models for web search using clickthrough data. In *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management*, pages 2333–2338, San Francisco, CA, USA.

Huang, X., Acero, A., Hon, H.-W., and Foreword By-Reddy, R. (2001). *Spoken language processing: A guide to theory, algorithm, and system development*. Prentice hall PTR.

Huang, X., Huang, Y. R., Zhong, M., and Wen, M. (2004). York University at TREC 2004: HARD and genomics tracks. In *NIST Special Publication: SP 500-26. Proceedings of the 13th Text Retrieval Conference (TREC 2004)*, Gaithersburg, Maryland.

Ilienko, A. (2013). Continuous counterparts of Poisson and binomial distributions and their properties. *arXiv preprint arXiv:1303.5990*.

Ircing, P. and Müller, L. (2007). Attempts to search Czech spontaneous spoken interviews - the University of West Bohemia at CLEF 2007 CL-SR track. In *Working Notes of the 8th Workshop of the Cross-Language Evaluation Forum (CLEF 2007): Advances in Multilingual and Multimodal Information Retrieval*, Budapest, Hungary.

Itou, K., Takeda, K., Takezawa, T., Matsuoka, T., Shikano, K., Kobayashi, T., Itahashi, S., and Yamamoto, M. (1999). JNAS: Japanese speech corpus for large vocabulary continuous speech recognition research. *The Journal of the Acoustical Society of Japan*, 20(3):199–206.

James, D. A. (1995). *The application of classical information retrieval techniques to spoken documents*. PhD thesis, University of Cambridge.

James, D. A. (1996). A system for unrestricted topic retrieval from radio news broadcasts. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 1996)*, pages 279–282, Atlanta, Georgia.

James, D. A. and Young, S. J. (1994). A fast lattice-based approach to vocabulary independent wordspotting. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 1994)*, pages 377–380, Adelaide, SA, Australia.

Järvelin, K. and Kekäläinen, J. (2002). Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446.

Jauhar, S. K., Chen, Y.-N., and Metze, F. (2013). Prosody-based unsupervised speech summarization with two-layer mutually reinforced random walk. In *Proceedings of the 6th International Joint Conference on Natural Language Processing (IJCNLP 2013)*, pages 648–654, Nagoya, Japan.

Jelinek, F. and Mercer, R. L. (1980). Interpolated estimation of Markov source parameters from sparse data. In *Proceedings of the International Workshop on Pattern Recognition in Practice*, pages 383–393, Amsterdam, Netherlands.

Jeon, J. H., Wang, W., and Liu, Y. (2011). N-best rescoring based on pitch-accent patterns. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, pages 732–741, Portland, Oregon.

Jiang, J. and Allan, J. (2016). Adaptive effort for search evaluation metrics. In *Proceedings of the 38th Annual European Conference on Information Retrieval (ECIR 2016)*, pages 187–199, Padua, Italy.

Jiang, J. and Zhai, C. (2004). UIUC in HARD 2004–passage retrieval using HMMs. In *NIST Special Publication: SP 500-26. Proceedings of the 13th Text Retrieval Conference (TREC 2004)*, Gaithersburg, Maryland.

Jiang, J. and Zhai, C. (2006). Extraction of coherent relevant passages using hidden Markov models. *ACM Transactions on Information Systems (TOIS)*, 24(3):295–319.

Johnson, S. E., Jourlin, P., Moore, G., Spärck Jones, K., and Woodland, P. C. (1999a). The Cambridge University spoken document retrieval system. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 1999)*, pages 49–52, Phoenix, AZ, USA.

Johnson, S. E., Jourlin, P., Spärck Jones, K., and Woodland, P. C. (2000). Spoken document retrieval for TREC-9 at Cambridge University. In *NIST Special Publication 500-249: Proceedings of the 9th Text REtrieval Conference (TREC-9)*, pages 117–126, Gaithersburg, Maryland.

Johnson, S. E., Woodland, P. C., Spärck Jones, K., and Jourlin, P. (1999b). Spoken document retrieval for TREC-8 at Cambridge University. In *NIST Special Publication*

500-246: Proceedings of the 8th Text REtrieval Conference (TREC-8), pages 197–206, Gaithersburg, Maryland.

Jones, G. J. F., Foote, J., Spärck Jones, K., and Young, S. (1997). The video mail retrieval project: experiences in retrieving spoken documents. In *Intelligent Multimedia Information Retrieval*, pages 191–214. MIT Press.

Jones, G. J. F., Foote, J. T., Spärck Jones, K., and Young, S. J. (1996). Retrieving spoken documents by combining multiple index sources. In *Proceedings of the 19th Annual International Conference on Research and Development in Information Retrieval (ACM SIGIR 1996)*, pages 30–38, Zurich, Switzerland.

Jones, G. J. F., Zhang, K., and Lam-Adesina, A. M. (2006). Dublin City University at CLEF 2006: Cross-language speech retrieval (CL-SR) experiments. In *Proceedings of the 7th Workshop of the Cross-Language Evaluation Forum (CLEF 2006): Evaluation of Multilingual and Multi-modal Information Retrieval*, pages 794–802, Alicante, Spain.

Kamps, J., De Rijke, M., and Sigurbjörnsson, B. (2004). Length normalization in XML retrieval. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (ACM SIGIR 2004)*, pages 80–87, Sheffield, United Kingdom.

Kamps, J., Pehcevski, J., Kazai, G., Lalmas, M., and Robertson, S. (2007). INEX 2007 evaluation measures. In *Proceedings of the 6th International Workshop of the Initiative for the Evaluation of XML Retrieval, (INEX 2007): Focused Access to XML Documents*, pages 24–33, Dagstuhl Castle, Germany.

Kaneko, T., Takigami, T., and Akiba, T. (2011). STD based on Hough transform and SDR using STD results: Experiments at NTCIR-9 SpokenDoc. In *Proceedings of the 9th NTCIR Workshop Meeting on Evaluation of Information Access Technologies, Information Retrieval, Question Answering, and Cross-Lingual Information Access (NTCIR-9)*, pages 264–270, Tokyo, Japan.

Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., and Fei-Fei, L. (2014). Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1725–1732, Columbus, Ohio.

Kaszkiel, M. and Zobel, J. (1997). Passage retrieval revisited. In *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (ACM SIGIR 1997)*, pages 178–185, Philadelphia, PA, USA.

Kaszkiel, M. and Zobel, J. (2001). Effective ranking with arbitrary passages. *Journal of the Association for Information Science and Technology*, 52(4):344–364.

Katz, S. (1987). Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 35(3):400–401.

Keikha, M., Park, J. H., Croft, W. B., and Sanderson, M. (2014). Retrieving passages and finding answers. In *Proceedings of the 2014 Australasian Document Computing Symposium*, pages 81–84, Melbourne, Australia.

Kekäläinen, J., Arvola, P., and Junkkari, M. (2009). *Contextualization*, pages 474–478. Springer US.

Kekäläinen, J. and Järvelin, K. (2002). Using graded relevance assessments in IR evaluation. *Journal of the American Society for Information Science and Technology*, 53(13):1120–1129.

Kießling, A. (1997). *Extraktion und Klassifikation prosodischer Merkmale in der automatischen Sprachverarbeitung*. Shaker.

Kim, J.-H. and Woodland, P. C. (2001). The use of prosody in a combined system for punctuation generation and speech recognition. In *Proceedings of the 7th European Conference on Speech Communication and Technology, Eurospeech 2001, Scandinavia*, pages 2757–2760, Aalborg, Denmark.

Kneser, R. and Ney, H. (1995). Improved backing-off for m-gram language modeling. In *Proceedings of the 1995 International Conference on Acoustics, Speech, and Signal Processing (ICASSP 1995)*, pages 181–184, Detroit, Michigan, USA.

Kolář, J., Shriberg, E., and Liu, Y. (2006). Using prosody for automatic sentence segmentation of multi-party meetings. In *Proceedings of the 9th International Conference on Text, Speech and Dialogue (TSD 2006)*, pages 629–636, Brno, Czech Republic.

Koumpis, K. and Renals, S. (2005). Automatic summarization of voicemail messages using lexical and prosodic features. *ACM Transactions on Speech and Language Processing (TSLP)*, 2(1).

Krifka, M. (2008). Basic notions of information structure. *Acta Linguistica Hungarica*, 55(3-4):243–276.

Krippendorff, K. (2011). Computing Krippendorff's Alpha-Reliability. *Departmental papers (ASC)*, pages 1–10.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1106–1114.

Kudo, T., Yamamoto, K., and Matsumoto, Y. (2004). Applying conditional random fields to Japanese morphological analysis. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP 2004)*, pages 230–237, Barcelona, Spain.

Lamel, L., Courcinous, S., Despres, J., Gauvain, J.-L., Josse, Y., Kilgour, K., Kraft, F., Le, V. B., Ney, H., Nußbaum-Thom, M., Oparin, I., Schlippe, T., Schlüter, R., Schultz, T., Fraga da Silva, T., Stüker, S., Sundermeyer, M., Vieru, B., Thang Vu, N., Waibel, A., and Woehrling, C. (2011). Speech recognition for machine translation in Quaero. In *International Workshop on Spoken Language Translation (IWSLT 2011)*, San Francisco, CA, USA.

Lamprier, S., Amghar, T., Levrat, B., and Saubion, F. (2008). Thematic segment retrieval revisited. In *Proceedings of the 13th International Conference on Artificial Intelligence: Methodology, Systems, and Applications (AIMSA 2008)*, pages 157–166, Varna, Bulgaria.

Lanchantin, P., Bell, P. J., Gales, M. J., Hain, T., Liu, X., Long, Y., Quinnell, J., Renals, S., Saz, O., Seigel, M. S., et al. (2013). Automatic transcription of multi-genre media archives. In *First Workshop on Speech, Language and Audio in Multimedia*, volume 1012, pages 26–31, Marseille, France. Cambridge University Press.

Larson, M., Eskevich, M., Ordelman, R., Kofler, C., Schmiedeke, S., and Jones, G. J. F. (2011). Overview of MediaEval 2011 Rich Speech Retrieval Task and genre tagging task. In *Working Notes Proceedings of the MediaEval 2011 Multimedia Benchmark Workshop*, Pisa, Italy.

Larson, M. and Jones, G. J. F. (2012a). Spoken content retrieval: A survey of techniques and technologies. *Foundations and Trends in Information Retrieval*, 5(4–5):235–422.

Larson, M. and Jones, G. J. F. (2012b). Spoken content retrieval: A survey of techniques and technologies. *Foundations and Trends in Information Retrieval*, 5(4–5):235–422.

Larson, M., Soleymani, M., Eskevich, M., Serdyukov, P., Ordelman, R., and Jones, G. J. F. (2012). The community and the crowd: Developing large-scale data collections for multimedia benchmarking. *IEEE Multimedia, Special Issue "Large-Scale Multimedia Data Collections"*, 19(3):15–23.

Lecorvé, G., Gravier, G., and Sébillot, P. (2008). An unsupervised web-based topic language model adaptation method. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2008)*, pages 5081–5084, Las Vegas, NV, USA.

LeCun, Y., Bengio, Y., et al. (1995). Convolutional networks for images, speech, and time series. *The Handbook of Brain Theory and Neural Networks*, 3361(10):255–258.

Lee, A. and Kawahara, T. (2009). Recent development of open-source speech recognition engine Julius. In *Proceedings of the 4th Annual Conference on Asia-Pacific Signal and Information Processing Association (APSIPA ASC 2009), 2009 Annual Summit and Conference*, pages 131–137, Sapporo, Japan.

Lee, A., Kawahara, T., and Doshita, S. (1998). An efficient two-pass search algorithm using word trellis index. In *Proceedings of the 5th International Conference on Spoken Language Processing (ICSLP 1998)*, pages 1831–1834, Sydney, Australia.

Lee, A., Kawahara, T., Takeda, K., Mimura, M., Yamada, A., Ito, A., Itou, K., and Shikano, K. (2002). Continuous Speech Recognition Consortium: an open repository for CSR tools and models. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC 2002)*, pages 1438–1441, Las Palmas, Canary Islands, Spain.

Lee, A., Shikano, K., and Kawahara, T. (2004). Real-time word confidence scoring using local posterior probabilities on tree trellis search. In *Proceedings of the 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2004)*, Montreal, Quebec, Canada.

Lee, L. and Rose, R. C. (1996). Speaker normalization using efficient frequency warping procedures. In *Proceedings of the 1996 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 1996)*, pages 353–356, Atlanta, GA, USA.

Lee, L.-s., Glass, J., Lee, H.-y., and Chan, C.-a. (2015). Spoken content retrieval—beyond cascading speech recognition with text retrieval. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(9):1389–1420.

Lehiste, I. (1973). Phonetic disambiguation of syntactic ambiguity. *Journal of the Acoustical Society of America*, 53(1):380–380.

Lehiste, I. B. (1970). *Suprasegmentals.* MIT Press Publication.

Levinson, S., Rabiner, L., and Sondhi, M. (1983). An introduction to the application of the theory of probabilistic functions of a Markov process to automatic speech recognition. *The Bell System Technical Journal*, 62(4):1035–1074.

Levow, G. (2007). University of Chicago at the CLEF 2007 cross-language speech retrieval track. In *Working Notes of the 8th Workshop of the Cross-Language Evaluation Forum (CLEF 2007): Advances in Multilingual and Multimodal Information Retrieval*, Budapest, Hungary.

Lileikyte, R., Lamel, L., and Gauvain, J.-L. (2015). Conversational telephone speech recognition for Lithuanian. In *Proceedings of the 3rd International Conference on Statistical Language and Speech Processing (SLSP 2015)*, pages 164–172, Budapest, Hungry. Springer International Publishing.

Liu, B. and Oard, D. W. (2006). One-sided measures for evaluating ranked retrieval effectiveness with spontaneous conversational speech. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (ACM SIGIR 2006)*, pages 673–674, Seattle, WA, USA.

Liu, Y., Shriberg, E., Stolcke, A., Hillard, D., Ostendorf, M., and Harper, M. (2006). Enriching speech recognition with automatic detection of sentence boundaries and disfluencies. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(5):1526–1540.

Luenberger, D. G. and Ye, Y. (1984). *Linear and nonlinear programming*, volume 2. Springer.

Luk, R. W., Leong, H. V., Dillon, T. S., Chan, A. T., Croft, W. B., and Allan, J. (2002). A survey in indexing and searching XML documents. *Journal of the Association for Information Science and Technology*, 53(6):415–437.

Lv, Y. and Zhai, C. (2009). Positional language models for information retrieval. In *Proceedings of the 32nd Annual International Conference on Research and Development in Information Retrieval (ACM SIGIR 2009)*, pages 299–306, Boston, MA, USA.

Maekawa, K. (2003). Corpus of Spontaneous Japanese: Its design and evaluation. In *Proceedings of the ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, Tokyo, Japan.

Maekawa, K., Kikuchi, H., and Tsukahara, W. (2004). Corpus of spontaneous Japanese: design, annotation and XML representation. In *Proceedings of the International Symposium on Large-Scale Knowledge Resources*, pages 19–24, Tokyo, Japan.

Maekawa, K., Koiso, H., Furui, S., and Isahara, H. (2000). Spontaneous speech corpus of Japanese. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC 2000)*, pages 947–952, Athens, Greece.

Maekawa, K., Yamazaki, M., Ogiso, T., Maruyama, T., Ogura, H., Kashino, W., Koiso, H., Yamaguchi, M., Tanaka, M., and Den, Y. (2014). Balanced corpus of contemporary written Japanese. *Language Resources and Evaluation*, 48(2):345–371.

Malioutov, I. and Barzilay, R. (2006). Minimum cut model for spoken lecture segmentation. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics (COLING ACL 2006)*, pages 25–32, Sydney, Australia.

Malioutov, I., Park, A., Barzilay, R., and Glass, J. (2007). Making sense of sound: Unsupervised topic segmentation over acoustic input. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pages 504–511, Prague, Czech Republic.

Mangu, L., Brill, E., and Stolcke, A. (1999). Finding consensus among words: lattice-based word error minimization. In *Proceedings of the 6th European Conference on Speech Communication and Technology, (Eurospeech 1999)*, Budapest, Hungary.

Manning, C. D., Raghavan, P., and Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.

Maskey, S. and Hirschberg, J. (2005). Comparing lexical, acoustic/prosodic, structural and discourse features for speech summarization. In *Proceedings of the 9th European Conference on Speech Communication and Technology, (INTERSPEECH 2005 - Eurospeech)*, pages 621–624, Lisbon, Portugal.

McGuinness, K., O'Connor, N. E., Aly, R., De Jong, F., Chatfield, K., Parkhi, O. M., Arandjelovic, R., Zisserman, A., Douze, M., and Schmid, C. (2013). The AXES PRO video search system. In *Proceedings of the 3rd ACM Conference on International Conference on Multimedia Retrieval*, pages 307–308, Dallas, TX, USA.

Mikolov, T., Karafiát, M., Burget, L., Cernockỳ, J., and Khudanpur, S. (2010). Recurrent neural network based language model. In *Proceedings of the 11th Annual Conference of the International Speech Communication Association (INTERSPEECH 2010)*, pages 1045–1048, Makuhari, Chiba, Japan.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Mishra, T., Sridhar, V. K. R., and Conkie, A. (2012). Word prominence detection using robust yet simple prosodic features. In *Proceedings of the 13th Annual Conference of the International Speech Communication Association (INTERSPEECH 2012)*, pages 1864–1867, Portland, OR, USA.

Mittendorf, E. and Schäuble, P. (1994). Document and passage retrieval based on hidden Markov models. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (ACM SIGIR 1994)*, pages 318–327, Dublin, Ireland.

Moffat, A. and Zobel, J. (2008). Rank-biased precision for measurement of retrieval effectiveness. *ACM Transactions on Information Systems (TOIS)*, 27(1):2.

Mulder, W. D., Bethard, S., and Moens, M.-F. (2015). A survey on the application of recurrent neural networks to statistical language modeling. *Computer Speech & Language*, 30(1):61–98.

Murata, M., Nagano, H., Hiramatsu, K., Kashino, K., and Satoh, S. (2016). Bayesian exponential inverse document frequency and region-of-interest effect for enhancing instance search accuracy. *IEICE Transactions on Information and Systems*, 99(9):2320–2331.

Murata, M., Nagano, H., Mukai, R., Kashino, K., and Satoh, S. (2014). BM25 with exponential IDF for instance search. *IEEE Transactions on Multimedia*, 16(6):1690–1699.

Muthusamy, Y. K., Cole, R. A., Oshika, B. T., Consortium, L. D., et al. (1992). The OGI multi-language telephone speech corpus. In *The Second International Conference on Spoken Language Processing, (ICSLP 1992)*, pages 895–898, Banff, Alberta, Canada.

Nanjo, H., Noritake, K., and Yoshimi, T. (2011). Spoken document retrieval experiments for SpokenDoc at Ryukoku University (RYSDT). In *Proceedings of the 9th NTCIR Workshop Meeting on Evaluation of Information Access Technologies, Information Retrieval, Question Answering, and Cross-Lingual Information Access (NTCIR-9)*, pages 275–280, Tokyo, Japan.

Nanjo, H., Yoshimi, T., Maeda, S., and Nishio, T. (2014). Spoken document retrieval experiments for SpokenQuery&Doc at Ryukoku University (RYSDT). In *Proceedings of the 11th NTCIR Conference on Evaluation of Information Access Technologies (NTCIR-11)*, pages 365–370, Tokyo, Japan.

Oard, D. W., Wang, J., Jones, G. J. F., White, R. W., Pecina, P., Soergel, D., Huang, X., and Shafran, I. (2006). Overview of the CLEF-2006 cross-language speech retrieval track. In *Proceedings of the 7th Workshop of the Cross-Language Evaluation Forum (CLEF 2006): Evaluation of Multilingual and Multi-modal Information Retrieval*, pages 744–758, Alicante, Spain.

Ogilvie, P. and Callan, J. (2004). Hierarchical language models for XML component retrieval. In *Proceedings of the 3rd International Workshop of the Initiative for the Evaluation of XML Retrieval on Advances in XML Information Retrieval (INEX 2004)*, pages 224–237, Dagstuhl Castle, Germany.

Ogilvie, P. and Callan, J. (2005). Parameter estimation for a simple hierarchical generative model for XML retrieval. In *Proceedings of the 4th International Workshop of the Initiative for the Evaluation of XML Retrieval on Advances in XML Information Retrieval and Evaluation (INEX 2005)*, pages 211–224, Dagstuhl Castle, Germany.

Ounis, I., Lioma, C., Macdonald, C., and Plachouras, V. (2007). Research directions in Terrier: A search engine for advanced retrieval on the web. *CEPIS Upgrade Journal*, 8(1):49–56.

Over, P., Fiscus, J., Sanders, G., Joy, D., Michel, M., Awad, G., Smeaton, A., Kraaij, W., and Quénot, G. (2014). TRECVID 2014: An overview of the goals, tasks, data, evaluation mechanisms and metrics. In *Proceedings of TRECVID 2014 Workshop*, Gaithersburg, USA.

Page, L., Brin, S., Motwani, R., and Winograd, T. (1999). The PageRank citation ranking: Bringing order to the Web. Technical report, Stanford InfoLab.

Pecina, P., Hoffmannová, P., Jones, G. J. F., Zhang, Y., and Oard, D. W. (2007a). Overview of the CLEF-2007 cross-language speech retrieval track. In *Proceedings of the 8th Workshop of the Cross-Language Evaluation Forum (CLEF 2007): Advances in Multilingual and Multimodal Information Retrieval*, pages 674–686, Budapest, Hungary.

Pecina, P., Hoffmannova, P., Jones, G. J. F., Zhang, Y., and Oard, D. W. (2007b). Overview of the CLEF 2007 cross-language speech retrieval track. In *Proceedings of the 8th Workshop of the Cross-Language Evaluation Forum (CLEF 2007): Advances in Multilingual and Multimodal Information Retrieval*, pages 674–686, Budapest, Hungary.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12(Oct):2825–2830.

Ponte, J. M. and Croft, W. B. (1998). A language modeling approach to information retrieval. In *Proceedings of the 21st Annual International Conference on Research and Development in Information Retrieval (ACM SIGIR 1998)*, pages 275–281, Melbourne, Australia.

Porter, M. F. (1980). An algorithm for suffix stripping. *Program*, 14(3):130–137.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., et al. (2011). The Kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society.

Preston, J., Hare, J., Samangooei, S., Davies, J., Jain, N., Dupplaw, D., and Lewis, P. H. (2013). A unified, modular and multimodal approach to search and hyperlinking video. In *Working Notes Proceedings of the MediaEval 2013 Multimedia Benchmark Workshop*, Barcelona, Catalunya, Spain.

Prince, E. F. (1981). Toward a taxonomy of given-new information. *Radical pragmatics*, 14:223–255.

Quinn, G. and Smeaton, A. (1999). Optimal parameters for segmenting a stream of audio into speech documents. In *Proceedings of the ESCA ETRW Workshop on Accessing Information in Spoken Audio*, pages 19–20, Cambridge, UK.

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the 1989 IEEE International Conference on Robotics and Automation*, pages 257–286.

Racca, D. N., Eskevich, M., and Jones, G. J. F. (2014). DCU search runs at MediaEval 2014 Search and Hyperlinking. In *Working Notes Proceedings of the MediaEval 2014 Multimedia Benchmark Workshop*, Barcelona, Catalunya, Spain.

Racca, D. N. and Jones, G. J. F. (2014). DCU at the NTCIR-11 SpokenQuery&Doc task. In *Proceedings of the 11th NTCIR Conference on Evaluation of Information Access Technologies (NTCIR-11)*, pages 376–383, Tokyo, Japan.

Racca, D. N. and Jones, G. J. F. (2015a). Evaluating Search and Hyperlinking: An example of the design, test, refine cycle for metric development. In *Proceedings of the MediaEval 2015 Multimedia Benchmark Workshop*, Wurzen, Germany.

Racca, D. N. and Jones, G. J. F. (2015b). Incorporating prosodic prominence evidence into term weights for spoken content retrieval. In *Proceedings of the 16th International Conference on Spoken Language Processing (INTERSPEECH 2015)*, pages 1378–1382, Dresden, Germany.

Racca, D. N. and Jones, G. J. F. (2016a). DCU at the NTCIR-12 SpokenQuery&Doc-2 task. In *Proceedings of the 12th NTCIR Conference on Evaluation of Information Access Technologies (NTCIR-12)*, pages 180–185, Tokyo, Japan.

Racca, D. N. and Jones, G. J. F. (2016b). On the effectiveness of contextualisation techniques in spoken query spoken content retrieval. In *Proceedings of the 39th Annual International Conference on Research and Development in Information Retrieval (ACM SIGIR 2016)*, pages 933–936, Pisa, Italy.

Renals, S. and Abberley, D. (2000). The THISL SDR system at TREC-9. In *NIST Special Publication 500-249: Proceedings of the 9th Text REtrieval Conference (TREC-9)*, pages 627–634, Gaithersburg, Maryland.

Reynar, J. C. (1998). *Topic segmentation: Algorithms and applications.* PhD thesis, University of Pennsylvania.

Robertson, S., Zaragoza, H., et al. (2009). The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389.

Robertson, S. E. (1977). The probability ranking principle in IR. *Journal of Documentation*, 33(4):294–304.

Robertson, S. E., Van Rijsbergen, C. J., and Porter, M. F. (1980). Probabilistic models of indexing and searching. In *Proceedings of the 3rd Annual ACM SIGIR Conference on Research and Development in Information Retrieval (ACM SIGIR 1980)*, pages 35–56, Cambridge, England.

Robertson, S. E. and Walker, S. (1994). Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval. In *Proceedings of the 17th Annual*

International Conference on Research and Development in Information Retrieval (ACM SIGIR 1994), pages 232–241, Dublin, Ireland.

Robertson, S. E., Walker, S., Jones, S., Hancock-Beaulieu, M. M., Gatford, M., et al. (1994). Okapi at TREC-3. In *NIST Special Publication: 500-226 Overview of the 3rd Text REtrieval Conference (TREC-3)*, pages 109–126, Gaithersburg, Maryland.

Röhr, C. T. (2013). Information status and prosody: Production and perception in German. *Interdisciplinary Studies on Information Structure*, 17:119.

Rose, R. C. (1991). Techniques for information retrieval from speech messages. *The Lincoln Laboratory Journal*, 4(1):45–60.

Rosenberg, A. (2010). AuToBI - a tool for automatic ToBI annotation. In *Proceedings of the 11th Annual Conference of the International Speech Communication Association (INTERSPEECH 2010)*, pages 146–149, Makuhari, Japan.

Rosenberg, A. (2012). Modeling intensity contours and the interaction between pitch and intensity to improve automatic prosodic event detection and classification. In *Proceedings of Spoken Language Technology Workshop (SLT 2012)*, pages 376–381, Miami, FL, USA.

Rousseau, A., Bougares, F., Deléglise, P., Schwenk, H., and Estève, Y. (2011). LIUM's systems for the IWSLT 2011 speech translation tasks. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT 2011)*, pages 79–85, San Francisco, CA, USA.

Rousseau, A., Deléglise, P., and Esteve, Y. (2012). TED-LIUM: an automatic speech recognition dedicated corpus. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, pages 125–129, Istanbul, Turkey.

Sahuguet, M., Huet, B., Červenková, B., Apostolidis, E., Mezaris, V., Stein, D., Eickeler, S., Garcia, J. R., Troncy, R., and Pikora, L. (2013). LinkedTV at MediaEval 2013 Search and Hyperlinking Task. In *Working Notes Proceedings of the MediaEval 2013 Multimedia Benchmark Workshop*, Barcelona, Spain.

Sakai, T. and Dou, Z. (2013). Summaries, ranked retrieval and sessions: A unified framework for information access evaluation. In *Proceedings of the 36th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (ACM SIGIR 2013)*, pages 473–482, Dublin, Ireland.

Salton, G. (1979). Mathematics and information retrieval. *Journal of Documentation*, 35(1):1–29.

Salton, G., Allan, J., and Buckley, C. (1993). Approaches to passage retrieval in full text information systems. In *Proceedings of the 16th Annual International Conference on*

Research and Development in Information Retrieval (ACM SIGIR 1993), pages 49–58, Pittsburgh, PA, USA.

Salton, G. and Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5):513–523.

Salton, G., Wong, A., and Yang, C.-S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620.

Sanderson, M. and Shou, X. M. (2007). Search of spoken documents retrieves well recognized transcripts. In *Proceedings of the 29th European Conference on Information Retrieval (ECIR 2007)*, pages 505–516, Rome, Italy.

Schaer, P. (2012). Better than their reputation? on the reliability of relevance assessments with students. In *Proceedings of the 3rd International Conference of the Cross-Language Evaluation Forum for European Languages (CLEF 2012): Information Access Evaluation. Multilinguality, Multimodality, and Visual Analytics*, pages 124–135, Rome, Italy.

Schaer, P., Mayr, P., Sünkler, S., and Lewandowski, D. (2016). How relevant is the long tail? a relevance assessment study on million short. In *Proceedings of the 7th International Conference of the CLEF Association: Experimental IR Meets Multilinguality, Multimodality, and Interaction (CLEF 2016)*, pages 227–233, Évora, Portugal.

Schmidt, K., Korner, T., Heinich, S., and Wilhelm, T. (2011). A two-step approach to video retrieval based on ASR transcriptions. In *Working Notes Proceedings of the MediaEval 2011 Multimedia Benchmark Workshop*, Pisa, Italy.

Schmiedeke, S., Xu, P., Ferrané, I., Eskevich, M., Kofler, C., Larson, M. A., Estève, Y., Lamel, L., Jones, G. J. F., and Sikora, T. (2013). Blip10000: a social video dataset containing SPUG content for tagging and retrieval. In *Proceedings of the 4th ACM Multimedia Systems Conference (MM Sys 2013)*, pages 96–101, Oslo, Norway.

Schouten, K., Aly, R., and Ordelman, R. (2013). Searching and hyperlinking using word importance segment boundaries in MediaEval 2013. In *Working Notes Proceedings of the MediaEval 2013 Multimedia Benchmark Workshop*, Barcelona, Catalunya, Spain.

Schuller, B., Steidl, S., Batliner, A., Bergelson, E., Krajewski, J., Janott, C., Amatuni, A., Casillas, M., Seidl, A., Soderstrom, M., et al. (2017). The INTERSPEECH 2017 computational paralinguistics challenge: Addressee, cold & snoring. In *Proceedings of the 18th International Conference on Spoken Language Processing (INTERSPEECH 2017)*, pages 3442–3446, Stockholm, Sweden.

Schwartz, R. and Austin, S. (1991). A comparison of several approximate algorithms for finding multiple (N-best) sentence hypotheses. In *Proceedings of the 1991 International Conference on Acoustics, Speech, and Signal Processing (ICASSP 1991)*, pages 701–704, Toronto, Canada.

Selkirk, E. O. (1984). *Phonology and syntax: the relationship between sound and structure.* MIT press.

Shao, Y., Hardmeier, C., Tiedemann, J., and Nivre, J. (2017). Character-based joint segmentation and POS tagging for Chinese using bidirectional RNN-CRF. *arXiv preprint arXiv:1704.01314*.

Shen, Y., He, X., Gao, J., Deng, L., and Mesnil, G. (2014a). A latent semantic model with convolutional-pooling structure for information retrieval. In *Proceedings of the 23rd ACM International Conference on Information and Knowledge Management (CIKM 2014)*, pages 101–110, Shanghai, China.

Shen, Y., He, X., Gao, J., Deng, L., and Mesnil, G. (2014b). Learning semantic representations using convolutional neural networks for web search. In *Proceedings of the 23rd International Conference on World Wide Web (WWW 2014)*, pages 373–374, Seoul, Republic of Korea.

Shi, J. and Malik, J. (2000). Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905.

Shiang, S.-R., Chou, P.-W., and Yu, L.-C. (2014). Spoken term detection and spoken content retrieval: evaluations on NTCIR-11 SpokenQuery&Doc task. In *Proceedings of the 11th NTCIR Conference on Evaluation of Information Access Technologies (NTCIR-11)*, pages 371–375, Tokyo, Japan.

Shigeyasu, K., Nanjo, H., and Yoshimi, T. (2009). A study of indexing units for Japanese spoken document retrieval. In *Proceedings of the 10th Western Pacific Acoustics Conference (WESPAC 2009)*, Beijing, China.

Shou, M. X., Sanderson, M., and Tuffs, N. (2003). The relationship of word error rate to document ranking. In *Proceedings of the AAAI Spring Symposium Intelligent Multimedia Knowledge Management Workshop, Technical Report SS-03-08*, pages 28–33, Palo Alto, California.

Shriberg, E., Stolcke, A., Hakkani-Tür, D., and Tür, G. (2000). Prosody-based automatic segmentation of speech into sentences and topics. *Speech Communication*, 32(1):127–154.

Siegler, M., Witbrock, M. J., Slattery, S. T., Seymore, K., Jones, R. E., and Hauptmann, A. G. (1997). Experiments in spoken document retrieval at CMU. In *NIST Special Publication: 500-240. Proceedings of the 6th Text REtrieval Conference (TREC-6)*, pages 291–302, Gaithersburg, Maryland.

Silipo, R. and Crestani, F. (2000). Prosodic stress and topic detection in spoken sentences. In *Proceedings of the 7th International Symposium on String Processing and Information Retrieval (SPIRE 2000)*, pages 243–252, A Coruña, Spain.

Simon, A.-R., Gravier, G., and Sébillot, P. (2015a). IRISA at MediaEval 2015: Search and anchoring in video archives task. In *Working Notes Proceedings of the MediaEval 2015 Multimedia Benchmark Workshop*, Wurzen, Germany.

Simon, A.-R., Sébillot, P., and Gravier, G. (2015b). Hierarchical topic structuring: from dense segmentation to topically focused fragments via burst analysis. In *Proceedings of the International Conference on Recent Advances on Natural Language Processing (RANLP 2015)*, pages 588–595, Hissar, Bulgaria.

Singhal, A., Buckley, C., and Mitra, M. (1996). Pivoted document length normalization. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (ACM SIGIR 1996)*, pages 21–29, Zurich, Switzerland.

Singhal, A., Choi, J., Hindle, D., Hirschberg, J., Pereira, F., and Whittaker, S. (1999). AT&T at TREC-7 SDR track. In *NIST Special Publication: 500-242. Proceedings of the 7th Text REtrieval Conference TREC-7*, pages 227–230, Gaithersburg, Maryland.

Singhal, A. and Pereira, F. (1999). Document expansion for speech retrieval. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (ACM SIGIR 1999)*, pages 34–41, Berkeley, California, USA.

Smeaton, A. F., Kelledy, F., and Quinn, G. (1997). Ad hoc retrieval using thresholds, WSTs for French mono-lingual retrieval, document-at-a-glance for high precision and triphone windows for spoken documents. In *NIST Special Publication 500-240: Proceedings of the 6th Text REtrieval Conference (TREC 6)*, pages 461–475, Gaithersburg, Maryland.

Smucker, M. D. and Clarke, C. L. (2012). Time-based calibration of effectiveness measures. In *Proceedings of the 35th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (ACM SIGIR 2012)*, pages 95–104, Portland, OR, USA.

Song, R., Yu, L., Wen, J.-R., and Hon, H.-W. (2011). A proximity probabilistic model for information retrieval. Technical report, Microsoft Research.

Soong, F. K. and Huang, E.-F. (1991). A tree-trellis based fast search for finding the n-best sentence hypotheses in continuous speech recognition. In *Proceedings of the 1991 International Conference on Acoustics, Speech, and Signal Processing (ICASSP 1991)*, pages 705–708, Toronto, Canada.

Spärck Jones, K. and van Rijsbergen C. J. (1975). Report on the need for and provision of an “ideal” information retrieval test collection. Technical report, Computer Laboratory, University of Cambridge.

Spärck Jones, K., Walker, S., and Robertson, S. E. (2000). A probabilistic model of information retrieval: Development and comparative experiments. Part 1 and 2. *Information Processing & Management*, 36(6):779–840.

Stanfill, C. and Waltz, D. L. (1992). Statistical methods, artificial intelligence, and information retrieval. *Text-based intelligent systems: Current research and practice in information extraction and retrieval*, pages 215–225.

Stoyanchev, S., Salletmayr, P., Yang, J., and Hirschberg, J. (2012). Localized detection of speech recognition errors. In *Proceedings of the Spoken Language Technology Workshop (SLT 2012), IEEE*, pages 25–30, Miami, USA.

Tax, N., Bockting, S., and Hiemstra, D. (2015). A cross-benchmark comparison of 87 learning to rank methods. *Information Processing & Management*, 51(6):757–772.

Taylor, M., Zaragoza, H., Craswell, N., Robertson, S., and Burges, C. (2006). Optimisation methods for ranking functions with multiple parameters. In *Proceedings of the 15th ACM International Conference on Information and Knowledge Management (CIKM 2006)*, pages 585–593, Arlington, Virginia, USA.

Terken, J. and Hermes, D. (2000). The perception of prosodic prominence. *Prosody: Theory and Experiment*, pages 89–127.

Terken, J. and Hirschberg, J. (1994). Deaccentuation of words representing 'given' information: Effects of persistence of grammatical function and surface position. *Language and Speech*, 37(2):125–145.

Tiedemann, J. and Mur, J. (2008). Simple is best: experiments with different document segmentation strategies for passage retrieval. In *Proceedings of the COLING 2nd Workshop on Information Retrieval for Question Answering*, pages 17–25, Manchester, UK.

Tsuge, S., Ohashi, H., Kitaoka, N., Takeda, K., and Kita, K. (2011). Spoken document retrieval method combining query expansion with continuous syllable recognition for NTCIR-SpokenDoc. In *Proceedings of the 9th NTCIR Workshop Meeting on Evaluation of Information Access Technologies, Information Retrieval, Question Answering, and Cross-Lingual Information Access (NTCIR-9)*, pages 249–256, Tokyo, Japan.

Tür, G., Hakkani-Tür, D., Stolcke, A., and Shriberg, E. (2001). Integrating prosodic and lexical cues for automatic topic segmentation. *Computational Linguistics*, 27(1):31–57.

Utiyama, M. and Isahara, H. (2001). A statistical model for domain-independent text segmentation. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL 2006)*, pages 499–506, Toulouse, France.

van der Werff, L. and Heeren, W. (2007). Evaluating ASR output for information retrieval. In *Proceedings of the 30th ACM SIGIR Workshop on Searching Spontaneous Conversational Speech Workshop*, pages 13–20, Amsterdam, Netherlands.

Varona Fernández, A., Nieto Nieto, S., Rodríguez Fuentes, L. J., Peñagarikano Badiola, M., Bordel García, G., and Díez Sánchez, M. (2011). A spoken document retrieval system for TV broadcast news in Spanish and Basque. *Procesamiento del Lenguage Natural*, 47:75–83.

Verma, M., Yilmaz, E., and Craswell, N. (2016). On obtaining effort based judgements for information retrieval. In *Proceedings of the 9th ACM International Conference on Web Search and Data Mining*, pages 277–286, San Francisco, CA, USA.

Vesel`y, K., Ghoshal, A., Burget, L., and Povey, D. (2013). Sequence-discriminative training of deep neural networks. In *Proceedings of the 14th Annual Conference of the International Speech Communication Association (INTERSPEECH 2013)*, pages 2345–2349, Lyon, France.

Viterbi, A. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, 13(2):260–269.

Voorhees, E., Garofolo, J., and Spärck Jones, K. (1997). The TREC-6 spoken document retrieval track overview and results. In *NIST Special Publication: 500-240. Proceedings of the 6th Text REtrieval Conference (TREC-6)*, pages 83–92, Gaithersburg, Maryland.

Voorhees, E. M. (2001). The TREC question answering track. *Natural Language Engineering*, 7(4):361–378.

Voorhees, E. M. and Buckley, C. (2002). The effect of topic set size on retrieval experiment error. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (ACM SIGIR 2002)*, pages 316–323, Tampere, Finland.

Voorhees, E. M. and Harman, D. K. (2005). *TREC: Experiment and Evaluation in Information Retrieval*. MIT Press Cambridge.

Wactlar, H. D., Christel, M. G., Gong, Y., and Hauptmann, A. G. (1999). Lessons learned from building a terabyte digital video library. *Computer*, 32(2):66–73.

Wactlar, H. D., Kanade, T., Smith, M. A., and Stevens, S. M. (1996). Intelligent access to digital video: Informedia project. *Computer*, 29(5):46–52.

Wade, C. and Allan, J. (2005). Passage retrieval and evaluation. Technical report, Massachusetts University Amherst Center for Intelligent Information Retrieval.

Wagner, M. and Watson, D. G. (2010). Experimental and theoretical advances in prosody: A review. *Language and Cognitive Processes*, 25(7-9):905–945.

Wang, J. and Oard, D. W. (2005). CLEF-2005 CL-SR at Maryland: Document and query expansion using side collections and thesauri. In *Proceedings of the 6th Workshop of the Cross-Language Evaluation Forum (CLEF 2005): Accessing Multilingual Information Repositories*, pages 800–809, Viena, Austria.

Ward, N. G. and Richart-Ruiz, K. A. (2013). Patterns of importance variation in spoken dialog. In *Proceedings of the 14th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL 2013)*, pages 107—111, Metz, France.

Ward, N. G. and Vega, A. (2012). A bottom-up exploration of the dimensions of dialog state in spoken interaction. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue SIGDIAL 2012 Conference*, pages 198–206, Seoul, South Korea.

Ward, N. G., Werner, S. D., Garcia, F., and Sanchis, E. (2015). A prosody-based vector-space model of dialog activity for information retrieval. *Speech Communication*, 68:85–96.

Ward, W. (1989). Understanding spontaneous speech. In *Proceedings of the Workshop on Speech and Natural Language*, pages 137–141, Cape Cod, Massachusetts, USA.

Wartena, C. (2012). Comparing segmentation strategies for efficient video passage retrieval. In *Proceedings of the 10th International Workshop on Content-Based Multimedia Indexing (CBMI 2012)*, pages 24–29, Annecy, France.

White, R. W., Oard, D. W., Jones, G. J. F., Soergel, D., and Huang, X. (2005). Overview of the CLEF-2005 cross-language speech retrieval track. In *Proceedings of the 6th Workshop of the Cross-Language Evaluation Forum (CLEF 2005): Accessing Multilingual Information Repositories*, pages 744–759, Viena, Austria.

Wilcox, L., Smith, I., and Bush, M. (1992). Wordspotting for voice editing and audio indexing. In *Proceedings of the Conference on Human Factors in Computing Systems (SIGCHI 1992)*, pages 655–656, Monterey, California, USA.

Woodland, P. C., Johnson, S. E., Jourlin, P., and Spärck Jones, K. (2000). Effects of out of vocabulary words in spoken document retrieval. In *Proceedings of the 23rd Annual International Conference on Research and Development in Information Retrieval (ACM SIGIR 2000)*, pages 372–374, Athens, Greece.

Wu, C., Karanasou, P., Gales, M. J., and Sim, K. C. (2016). Stimulated deep neural network for speech recognition. In *Proceedings of the 20th International Conference on Spoken Language Processing (INTERSPEECH 2016)*, pages 400–404, San Francisco, USA.

Xie, S., Hakkani-Tür, D., Favre, B., and Liu, Y. (2009). Integrating prosodic features in extractive meeting summarization. In *Proceedings of the 2009 IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU 2009)*, pages 387–391, Merano, Italy.

Xiong, W., Droppo, J., Huang, X., Seide, F., Seltzer, M., Stolcke, A., Yu, D., and Zweig, G. (2016). Achieving human parity in conversational speech recognition. *arXiv preprint arXiv:1610.05256*.

Young, S., Evermann, G., Hain, T., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., and Woodland, P. (2002). *The HTK Book (for HTK Version 3.2)*. Cambridge University Engineering Department.

Yu, D. and Deng, L. (2014). *Automatic speech recognition: A deep learning approach*. Springer.

Yu, D., Li, J., and Deng, L. (2011). Calibration of confidence measures in speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(8):2461–2473.

Zhai, C. (2001). Notes on the Lemur TFIDF model. Technical report, Carnegie Mellon University.

Zhang, H.-P., Liu, Q., Cheng, X.-Q., Zhang, H., and Yu, H.-K. (2003). Chinese lexical analysis using hierarchical hidden Markov model. In *Proceedings of the second SIGHAN Workshop on Chinese Language Processing*, pages 63–70, Sapporo, Japan.

Zhang, Y., Park, L. A., and Moffat, A. (2010). Click-based evidence for decaying weight distributions in search effectiveness metrics. *Information Retrieval*, 13(1):46–69.

Zobel, J. (1998). How reliable are the results of large-scale information retrieval experiments? In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1998)*, pages 307–314, Melbourne, Australia.

Zobel, J. and Moffat, A. (1998). Exploring the similarity space. In *Proceedings of the ACM SIGIR Forum*, volume 32, pages 18–34, New York, NY, USA.

Zobel, J. and Moffat, A. (2006). Inverted files for text search engines. *ACM Computing Surveys (CSUR)*, 38(2):6.