1

Speech Indexing

Roeland Ordelman[1], Franciska de Jong[1,2], and David van Leeuwen3

Human Media Interaction, University of Twente
TNO Information and Communication Technology,
TNO Defence, Security and Safety

1.1 Introduction

The amount of metadata attached to multimedia collections that can be used for searching is very much dependent on the available resources within the organisations that create or own the collections. Large national audiovisual institutions, such as Sound&Vision in The Netherlands[1], put a lot of effort in archiving their assets and they label collection items with at least titles, dates and short content descriptions (descriptive metadata, see chapter **??**). However, many organisations that create or own multimedia collections lack the resources to apply even the most basic form of archiving. Certain collections may become the stepchild of an archive—minimally managed, poorly preserved, and hardly accessible.

Although the saying "information is in the audio, video is for entertainment[2]" puts it somewhat strongly, it gives an impression of the potential for the deployment of audio and for the application of information extraction techniques to support multimedia information retrieval tasks. Especially the speech in audio is an important information source that, once transformed into text and/or enriched with linguistic annotation, can enable the conceptual querying of video content. The basic idea is to use automatic speech recognition technology to generate such a linguistic annotation or textual representation (see Figure 1.1) and to use this as (a source for) automatically created metadata that can be used for searching by applying standard text-based information retrieval techniques.

Next to the words spoken, also information about the speaker can be extracted from the speech waveform, referred to as speaker classification. Typical examples are the speaker's identity or gender, which can be useful for the detection of document structure (who is speaking when), or even the speaker's

[1] Sound&Vision:`http://www.beeldengeluid.nl/`
[2] Richard Schwartz (BBN Technologies) at the Multimedia Retrieval Video-Conference at the University of Twente in 1999

background (social, geographic, etc.) or emotional state. Apart from speech, often other clues in the audio can have added value, such as background noise sources, sounds, music, adverts, channel characteristics and bandwidth of transmission.

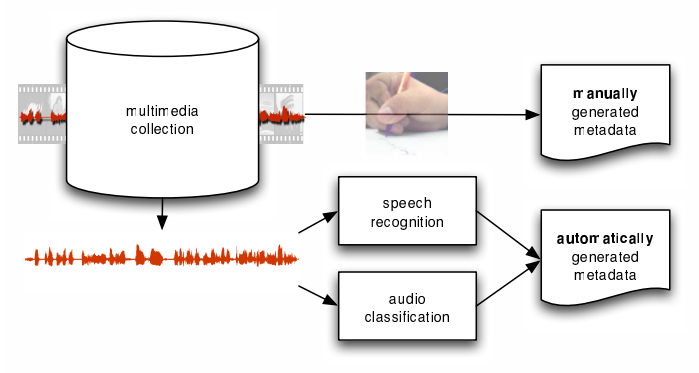

Fig. 1.1. Automatic metadata generation using the audio stream

This chapter will focus on the automatic extraction of information from the speech in multimedia documents. This approach is often referred to as speech indexing and it can be regarded as a subfield of audio indexing that also incorporates for example the analysis of music and sounds. If the objective of the recognition of the words spoken is to support retrieval, one commonly speaks of spoken document retrieval (SDR). If the objective is on the coupling of various media types the term media mining or even cross-media mining is used. Most attention in this chapter will go to SDR. The focus is less on searching (an index of) a multimedia database, but on enabling multiple views on the data by cross-linking all the available multifaceted information sources in a multimedia database. In section 1.6 cross-media mining will be discussed in more detail.

1.1.1 Relation To Other Chapters

Throughout this book, the searching process is generally described as trying to find a match between an information need, formulated in a query and represented in a query representation, and a collection of documents, represented in a document representation (often referred to as an index). From a user's perspective, using natural language to formulate an information need in a query, is the most evident choice although other modalities are thinkable (see also "Interaction" in Chapter**??**). However, with audiovisual content, the representation of a natural language query does not match the representation of the documents (images in pixels, audio in samples). The main focus

of this chapter (and also chapters**??**) is on solving this representation mismatch by converting the document collection to the natural language query representation.

Solving the representation mismatch:

- convert the document collection to the natural language query representation (e.g., speech to text)
- adjust the query to the document representation (e.g., example image as query)
- convert both document and query representation to an intermediate representation (e.g., query and speech to sound units).

1.1.2 Outline

The remainder of this chapter starts with a brief introduction into speech recognition in general (section 1.2), and a detailed overview of the application of speech recognition technology in the context of multimedia indexing in a section on spoken document retrieval (section 1.3). Here, the synchronisation or time-alignment of collateral data sources, large vocabulary speech recognition, keyword spotting and SDR using sub-word unit representations will be discussed. In the section on robust speech recognition and retrieval (section 1.4), we zoom in on the optimisation of speech recognition performance in the context of spoken document retrieval, discussing query and document expansion, vocabulary optimisation, topic-based language models and acoustic adaptation. The chapter will be finalised by discussing audio segmentation (section 1.5) and a topic that links speech indexing to other modalities in the multimedia framework: cross-media mining (section 1.6).

1.2 Brief introduction into speech recognition

In the speech recognition process several steps can be distinguished. Recognition systems convert an acoustic signal into a sequence of words via a series of processes that are visualised in a very simplified manner in Figure 1.2.

1.2.1 Feature extraction

First the digitised acoustic signal is converted into a compact representation that captures the characteristics of the speech signal using spectral information. This step is usually referred to as feature extraction. A spectrum describes how the different frequency components in a waveform vary in time and this information is represented by vectors of features which are computed for example every 10 ms for a 16 ms overlapping time window. An LPC (Linear Predictive Coding) spectrum is an example of a smoothed spectrum. Often,

the spectral features are modified in one way or another in order to make them more consistent with how the human ear works (e.g., Mel-scale), and averaged over spectral bands. A commonly used feature set is based on a derivation of the spectrum—the cepstrum—that is computed by taking a Discrete Cosine Transform of a band-filtered spectrum. When the resulting Cepstral Coefficients are Mel-scaled we end up with the popular MFCC (Mel Frequency Cepstral Coefficients) feature set.

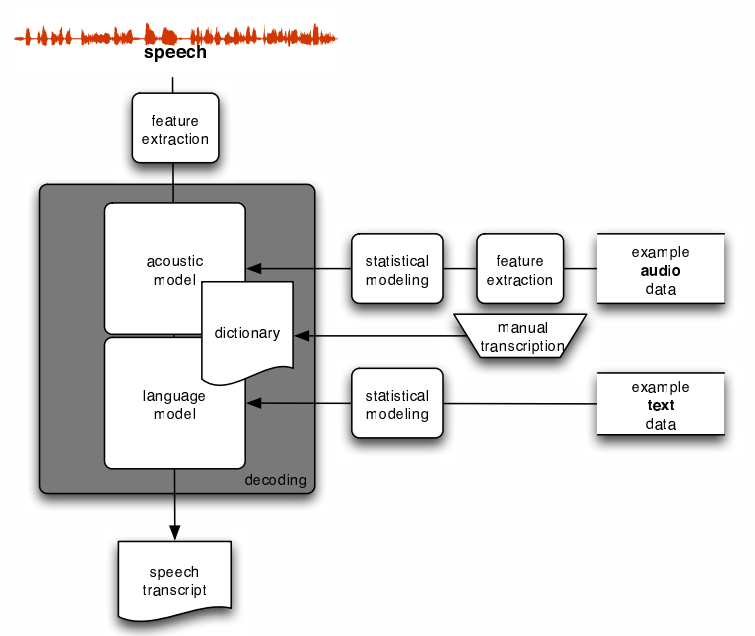

Fig. 1.2. Simplified overview of a speech recognition system: on the left the flow from speech to a speech transcription via feature extraction and a decoding stage that includes the acoustic model, the language model and dictionary; on the right the flow of required statistical (audio&text) and manual information.

1.2.2 Acoustic modelling

In the statistical speech recognition framework, the feature vectors are treated as acoustic observations (O). The task is formulated as to find the sequence of words W ($W = \{\omega_1, \omega_2, \ldots, \omega_N\}$) that are most likely to have been spoken on the basis of the the acoustic observations (O). The probability of a sentence being produced given some acoustic observations is typically expressed as

$P(W|O)$. The most likely sentence ($\hat{W}$) is found by computing $P(W|O)$ for all possible sentences and choosing the one with the highest probability:

$$\hat{W} = \arg\max_{W} P(W|O) \tag{1.1}$$

Using Bayes' rule, the conditional probability of a sentence W being spoken, assuming that certain acoustic observations O were made, can be expressed as:

$$P(W|O) = \frac{P(O|W) \cdot P(W)}{P(O)} \tag{1.2}$$

where $P(O|W)$ is the likelihood that specific acoustic observations are made given a sentence W, and $P(W)$ the prior probability of the sequence of words W, which can be estimated using a statistical language model. $P(O)$ is the probability of observing the given speech input. As for the computation of the most probable sentence given a certain speech input, $P(O)$ is the same for all possible W, $P(O)$ may be regarded as a normalisation factor that can well be removed from the computation:

$$\hat{W} = \arg\max P(W|O) = \arg\max \overbrace{P(O|W)}^{AM} \cdot \overbrace{P(W)}^{LM} \tag{1.3}$$

The most popular approach in speech recognition is to use hidden Markov models (HMMs) to compute the acoustic model probabilities $P(O|W)$ and language model probabilities $P(W)$. An HMM is a stochastic automaton that consists of a set of connected states, each having a transition probability and an output or emission probability associated with it (probability density functions, see below). The transition probabilities model the transitions from one state to the other. The output probabilities model the observation likelihoods of an observation being generated from a particular state. In HMM speech recognition, the problem of finding $P(O|W)$ can be expressed as finding $P(O|M)$, the likelihood that the observations O were generated by a sequence of word HMM models (M) that are associated with a sentence W. The word models are in turn composed of sub-word unit models, typically models based on the smallest unit in speech, the phone. In other words, the calculation of $P(O|W)$ involves the computation of the likelihood that the observations O are generated by a particular set of HMM states. The usual HMM training approach is to construct probability density functions (PDFs) that model the likelihood of HMM states emitting a particular observation. These PDFs are typically Gaussians or mixtures of Gaussians. The parameters of the PDFs are typically estimated so as to model the training data. The Viterbi algorithm or alternatively a best-first search algorithm (stack decoding or A$*$-search), is then used to find the best path through the network of HMMs given the observations. See for example [**?**] for a detailed survey of HMMs and search algorithms in speech recognition.

1.2.3 Language modelling

The prior probability of a sentence $p(W)$ in speech recognition can typically be estimated using statistical n-gram models. Using the chain rule of probability, $p(W)$ can be formally expressed as:

$$p(W) = \prod_{i=1}^{n} p(\omega_i|\omega_1, \ldots, \omega_{i-1}) \tag{1.4}$$

where $p(\omega_i|\omega_1, \ldots, \omega_{i-1})$ is the probability that the word ω_i was spoken, immediately following the preceding word sequence $\omega_1, \ldots, \omega_{i-1}$, that is referred to as the history of the word ω_i. However, computing the probability of a word given a long history of words is not feasible. Theoretically, it depends on the entire past history of a discourse. The n-gram language model attempts to provide an adequate approximation of $P(\omega_i)$ by referring to the Markov assumption that the probability of a future event can be predicted by looking at its immediate past. n-gram language models therefore use the previous $n-1$ words (typically one or two words) as an approximation of the entire history. That this approximation is reasonably adequate can be derived from the fact that n-gram language models were introduced in speech recognition in the 1970's and still remain state-of-the-art. For a two-word history, trigram models can be generated by reformulating equation 1.4 as:

$$p(\omega) \approx p(\omega_0) \cdot p(\omega_1|\omega_0) \cdot \prod_{i=2}^{n} p(\omega_i|\omega_{i-2}, \omega_{i-1}) \tag{1.5}$$

N-gram probability estimates can be computed using the relative frequencies, called maximum likelihood estimates (ML): the normalised counts of n-grams in a training corpus. For a trigram model that is:

$$p(\omega_3|\omega_1, \omega_2) = f(\omega_3|\omega_1, \omega_2) \doteq \frac{C(\omega_1, \omega_2, \omega_3)}{C(\omega_1, \omega_2)} \tag{1.6}$$

or in a generalised form:

$$p(\omega_i|\omega_{i-n+1}^{i-1}) = \frac{c(\omega_{i-n+1}^{i-1})}{\sum_{w_i} c(\omega_{i-n+1}^{i-1})}, \tag{1.7}$$

where we introduced the notation ω_i^j for the sequence of words $\omega_i, \omega_{i+1}, \ldots, \omega_j$. As even very large training corpora can never cover all possible n-grams for a language, it is possible that perfectly acceptable n-grams are not encountered in the training corpus. A language model based on equation 1.6 would assign a zero probability to such 'unseen' n-grams. So regardless of the evidence provided by the acoustic signal in favour of an n-gram not encountered in the training data, the n-gram will never be reproduced by the language model. Moreover, it is well-known that using relative frequencies as a way to estimate probabilities, produces poor estimates when the n-gram counts are small. To create a more uniform distribution, it is necessary to smooth these zero-probability and low-probability n-grams.

1.2.4 Dictionary

The speech recognition vocabulary is a list of all the words in the language model. It can be considered the model of pronunciation in the recognition system. By presenting for every word in the vocabulary a pronunciation, the pronunciation dictionary is the link between the acoustic model and the language model. Word pronunciations can be viewed as rules for the concatenation of phone models to arrive at the words contained in the language model. During decoding, the words in the dictionary are usually compactly represented by networks of phones, e.g., in the form of a lexical tree, where each path through the network represents a word. Using this network, a Viterbi based search can be performed to find the most probable path through the network.

To obtain word pronunciations for the large and dynamic speech recognition vocabularies, speech recognition developers usually deploy a large background pronunciation lexicon to enable a flexible generation of word pronunciations. When word pronunciations are not in the background lexicon, word transcripts can be manually generated, or produced by a grapheme-to-phoneme[3] converter (G2P) that uses rules or machine-learning techniques for pronunciation generation. As generating pronunciations manually is time consuming, a G2P converter is often indispensable, especially in the dynamic news domain that has a lot of proper names and names of cities and places that are often not included in background lexicons.

Background lexicons and especially G2P tools usually provide canonical word pronunciations only, according to a normative, "average" pronunciation of words. In practice however, words are pronounced in numerous variations in different degradations from the canonical pronunciation. Among others things this is due to age, gender or dialect (inter-speaker variability) and speaking style, speaking rate, co-articulation or emotional state of the speaker (intra-speaker variability). It has been estimated that in spontaneous speech around 40 % of the words is not pronounced according to the canonical representation. As such mismatches may occur both at acoustic modelling training stage and at the recognition stage, such variations result in a degradation of word accuracy of the speech recognition system. By incorporating pronunciation variations in the lexicon, the number of inaccurate phone-to-word mappings can be reduced.

1.2.5 Summary

Speech recognition is based on two important techniques, *modelling* and *search*. The task of *modelling* is to capture the acoustics of speech, the pronunciation of words and the sequence of words in a way that these models are general enough to describe the various sources of variability found in speech (speaker, coarticulation, word choice), whilst being specific enough to extract

[3] Also referred to as text-to-speech or letter-to-phone/sound

the linguistic information. The task of the *search* is to find the sequence of models that fit the observed speech data best in an efficient way.

1.3 Spoken Document Retrieval

To support spoken document retrieval, the crucial step is to create a textual representation of an audiovisual document by automatically annotating the speech in the data. The resulting representation can be regarded as automatically generated metadata that can either be used as a replacement for missing human-generated meta-data, or as an additional information source. Especially for smaller organisations that own multimedia collections, resources are often lacking to apply even a basic form of archiving. In such cases, applying spoken document retrieval is the only way to provide means for searching the collections. For larger organisations with audiovisual archives, administrative metadata, such as rights metadata (who is the legal owner of a video item) and technical metadata (such as the format of a video item), is usually available. Often even descriptive human-generated metadata is also preserved with the collection items, such as the title, duration, a short content description and a list of names and places that are mentioned. Although the information density in this type of metadata is usually low, it can nevertheless be very helpful for locating specific documents in a collection. This type of information is therefore sometimes referred to as "bibliographic" or "tombstone information" among librarians[4]. However, locating specific parts within documents (e.g., the passage where a specific subject is addressed) remains time consuming as it requires manual scrolling through a (possibly large) multimedia document. Although multimedia documents can often be structured using available information sources (such as the occurrence of speaker changes, speech/non-speech boundaries, large silence intervals or shot boundaries) and techniques are being developed to support quick browsing of video documents (e.g., by enabling fast playback of speech), going manually through large video documents can still be cumbersome.

1.3.1 Manual versus automatic annotation

In theory, instead of using speech recognition technology, the annotation of the speech could as just well be carried out manually. In fact, in the meeting domain, manual annotation of the speech is quite common. Usually meeting minutes are generated, either stenographically or not. When the meetings are recorded on video, such annotations can very well be used as a textual representations for indexing and retrieval. Often however, such annotations do

[4] Note that besides administrative and descriptive metadata, a third type of metadata is often distinguished: information about the structure and organisation of a multi-part digital object that can be encoded MPEG-7/21

not exist and it is usually too expensive to generate them manually. In specific cases though, manual annotation can be a valid option, especially when the collection is relatively small and fixed, and/or the quality of automatic annotations is too low.

The application of automatic speech recognition (ASR) technology for indexing purposes has been made possible thanks to the large improvements in the performance of ASR systems in recent years. This is partly due to the increase in computer power but also to massive speech recognition research efforts especially in the context of benchmark evaluations (often sponsored by DARPA[5]) ranging from evaluations focusing on read speech (Wall Street Journal) in the early nineties, via broadcast news speech in the second half of the nineties and conversational speech early this decennium to meeting room speech most recently.

1.3.2 Requirements for recognition performance

IR-oriented benchmarks such as TREC (Text Retrieval Conference) demonstrated that deploying ASR techniques has become more than a theoretical option for the automatic annotation of speech for retrieval purposes. This is especially the case in the broadcast news domain which is very general and makes data collection for training a speech recognition system relatively easy. For the broadcast news domain, speech transcripts approximate the quality of manual transcripts, at least for several languages. Spoken document retrieval in the American-English broadcast news (BN) domain was even declared "solved" with the Spoken Document Retrieval track at TREC in 2000. However, in other domains than broadcast news and for other languages, a similar recognition performance is usually harder to obtain due to a number of factors including the lack of well-balanced speech databases for certain languages, the lack of domain-specific training data for certain domains, and of course due to the large variability in audio quality and speech characteristics. Some of these issues will be discussed in more detail in section 1.4, but first a number of general techniques used in spoken document retrieval will be outlined.

1.3.3 Spoken document retrieval techniques

Below, a number of techniques that can play a role in spoken document retrieval will be described in detail. Which combination of techniques is chosen for the disclosure of a specific multimedia database depends on a number of factors. For each of the techniques these factors will be listed. We will discuss

- synchronisation of available textual resources

[5] The Defence Advanced Research Projects Agency (DARPA) is the central research and development organisation for the US department of Defence

- large vocabulary speech recognition
- keyword spotting
- using sub-word as unit for representation

The first two are the most frequently used techniques. The latter two are sometimes used as the primary annotation strategy, but often in combination with the first two techniques.

Synchronisation of collateral data sources

To allow the conceptual querying of video content without having to set-up a speech recognition system to generate full-text transcriptions, collateral textual resources that are closely related with the collection items can be exploited. A well known example of such a textual resource is subtitling information for the hearing-impaired (e.g., *CEEFAX* pages 888 in the UK) that is available for the majority of contemporary broadcast items, in any case for news programs. Subtitles contain a nearly complete transcription of the words spoken in the video items and provide an excellent information source for indexing. Usually, they can easily be linked to the video by using the time-codes that come with the subtitles. The Dutch news subtitles even provide topic boundaries that can be used for segmenting the news show into sub-documents. Textual sources that can play a similar role are teleprompter files (the texts read from screen by an anchor person, also referred to as auto-cues) and scenarios. Teletext subtitles can relatively easy be obtained using the teletext capturing functionality in most modern television boards. Teleprompter files and scenarios of course have to be provided by the producers of the videos.

The time labels in these sources are crucial for the creation of a textual index into video. In the collateral text sources mentioned above, the available time-labels are not always fully reliable and can even be absent. In that case the text files will have to be synchronised. Examples of such text sources are minutes of meetings or written versions of lectures and speeches. Below, two approaches for the automatic generation of time-stamps for minutes will be described using two scenarios: the synchronisation of (i) the so-called *Handelingen*, i.e., the meetings of the Dutch Parliament, and (ii) the minutes of Dutch city council meetings to the video recordings of the meetings. Due to the difference in accuracy of the minutes, two different approaches are needed.

The minutes of the meetings of the Dutch Parliament are stenographic minutes that closely follow the discourse of the meeting, only correcting slips of the tongue and ungrammatical sentences. Given the close match with the actual speech, a relatively straightforward so-called forced alignment procedure could be used. Forced alignment is a technique commonly used in acoustic model training in automatic speech recognition (ASR). In order to be able to train phone models, words and phones in pre-segmented sentences are aligned

to their exact location in the speech segments using an acoustic model[6]. Given a set of words from a sentence the acoustic model tries to find the most optimal distributions of these words given the audio signal on the basis of the sounds the words are composed of. When using alignment for indexing, presegmented sentences are evidently not available but as long as the text follows the speech well enough, the word alignment can be found by using relatively large windows of text.

The alignment procedure works well even if some words in the minutes are actually not in the speech signal. However, if the text to be aligned does not match the speech too well, as was the case with city council meetings, and if the text segments are too large, the alignment procedure will fail to find a proper alignment. In order to produce suitable segments, a two-pass strategy can be used, incorporating the following steps as visualised in Figure 1.3:

1. a baseline large vocabulary speech recognition system[7] is used to generate a relatively inaccurate transcript of the speech with word-timing labels, referred to as hypothesis;
2. the hypothesis is aligned on the word level to the minutes using a dynamic programming algorithm;
3. at the positions where the hypothesis and the minutes match (a match may be defined as three words in a row are correctly aligned), so called 'anchors' are placed.
4. using the word-timing labels provided by the speech recognition system, the anchors are used to generate suitable segments;
5. individual segments of audio and text are accurately synchronised using forced alignment;

The described methods allow for the synchronisation of audiovisual data to available linguistic content that approximates to a certain extent the speech in the source data and they enable the processing of conceptual queries of the audiovisual content. In the second alignment procedure, an initial hypothesis is generated by a large vocabulary speech recognition system. As this hypothesis is only needed for finding useful segments, the performance of the system is not crucial as long as it is able to provide 'anchors'. However, the relevance of speech recognition performance increases when textual resources suitable for alignment with audiovisual data are *not* available. In the next section, the application of speech recognition technology as the *primary* source for generating a textual representation of audiovisual documents that can be linked to other linguistic content, is described.

[6] In the first iteration usually an 'averaged' bootstrap model is used. The alignment and the model should improve iteratively

[7] Optionally the speech recognition is somewhat adapted to the task for example by providing it with a vocabulary extracted from the minutes

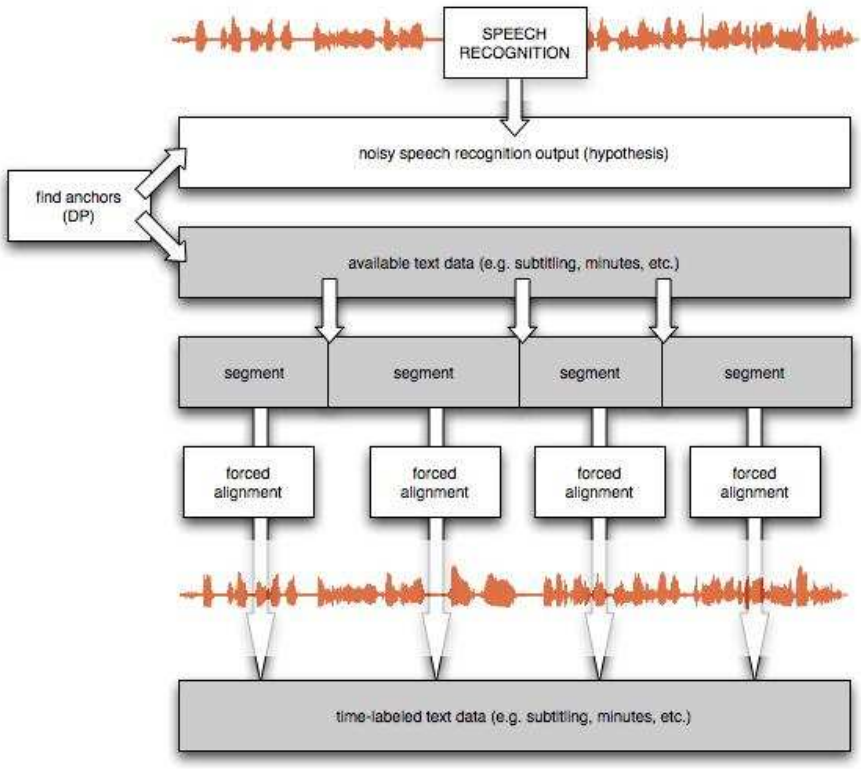

Fig. 1.3. Alignment procedure: synchronisation of text and audio in a number of steps.

Large Vocabulary Speech Recognition

Using speech recognition technology to convert spoken audio into text for retrieval purposes, may seem a rather obvious solution. However, in order to obtain reasonable retrieval results, a speech recognition system has to produce reasonably accurate transcription of what was actually spoken. When a system produces lots of errors, successful retrieval will be doubtful. When it produces perfect transcripts, retrieval will resemble the performance of retrieving text documents. How accurate exactly speech recognition should be for acceptable retrieval performance was uncertain at the outset of SDR research, although some experience was gained with the retrieval of corrupted documents (e.g., from OCR) at TREC-5 in 1996. At the first SDR evaluation at TREC-6 on broadcast news data, word error rates fell between 35 % and 40 % which appeared to be good enough for acceptable retrieval results in a known-item retrieval task, simulating a user seeking one particular document.

Already at TREC-7, where the known-item retrieval task was replaced by the ad-hoc retrieval task of searching multiple relevant documents from single topics, speech recognition performance was improved substantially—the University of Cambridge HTK recognition system produced error rates below the 25 %—and almost all retrieval systems performed reasonably well. Also at TREC-7, evidence could be provided for the assumption that better speech recognition performance will also result in better retrieval performance. A speech recognition performance of 50 % WER is on the other hand regarded as a minimum for obtaining useful retrieval performance.

In the TREC SDR tracks, retrieval systems that used automatic annotations from a speaker-independent large vocabulary speech recognition systems outperformed other approaches (e.g., phone based approach, see below). Most multimedia retrieval systems that use speech transcripts nowadays use such large vocabulary systems that exists in many flavours and configurations. All systems are speaker-independent and have large vocabularies. The speaker-independency is required in the context of indexing multimedia collections as it is usually not known which speakers appear in the collection. Although speaker adaptation strategies are often applied (for example by clustering audio of single speakers and creating adapted, speaker specific models), speaker-independent models are the basis for LVCSR in the context of multimedia indexing. Speaker-independent models are typically trained using a large amount of example audio data from the task domain to make sure that most of the inter-speaker variabilities are captured. The large vocabulary is also a prerequisite, given the variety of words encountered in fluent speech and the fact that it is often very hard to predict which words are going to be used by speakers in a task domain. Typically a vocabulary of 65 thousand words (65 K) is used in LVCSR. Large corpora of text data are needed to train language model probabilities for the words in the vocabulary.

Because of the the requirements for training a LVCSR system, setting up a large vocabulary speech recognition system for a certain language in a specific domain is complex and time-consuming. It is crucial that sufficient amounts of in-domain training data are available to enable the capturing of the acoustic and linguistic variability in the task domain and to train robust acoustic models and language models. To give an impression of the amounts of data that are used for typical systems in the English broadcast news benchmark tests, the LIMSI/BBN[8] 2004 English Broadcast News speech recognition system uses for acoustic model training some 140 hours of carefully transcribed broadcast news audio data and for language model training the manual transcriptions of the acoustic BN data (1.8 M words), the American English GigaWord News

[8] The Computer Sciences Laboratory for Mechanics and Engineering Sciences (LIMSI) is a research laboratory associated to Paris-6 and Paris-11 Universities and is one of the major players in large vocabulary speech recognition research (`www.limsi.fr`). BBN technologies is a US company that has been performing pioneering research in automatic speech recognition since the early 1970s (`www.bbn.com`).

corpus for a total amount of approximately 1 billion words of texts, and a few hundreds of million words of other text data. A large part of this data became available with the broadcast news benchmark evaluations (Hub4). For other languages than English and for other domains than news such amounts of annotated data can often not be laid hands on easily. In such cases, one has to come up with strategies to deal with this lack of training data.

Especially in the context of retrieval, the out-of-vocabulary (OOV) problem in speech recognition deserves special attention. When a word is not in the speech recognition vocabulary, it can not be recognised and hence, will not turn up in an annotation of a video document. In section 1.4 strategies that try to deal with this OOV problem are described in detail. One of these strategies is making use of keyword spotting that is addressed next.

Keyword Spotting

Because of its relative simplicity, earliest attempts to deploy speech recognition technology in SDR made use of word spotting techniques to search for relevant documents in audio material. A keyword spotter searches the audio material for single keywords or multi-word expressions (such as "New York" or "football game"). An acoustic model is used to recognise phones and a small vocabulary of keywords with phonetic transcriptions provide the link to the keywords. Keyword searches are often weighted using a simple grammar (such as a Finite State Grammar). Weighting can be uniform for all keywords or be based on the probability distribution of the keywords in the database. Normally the spotter has a facility to reduce incorrect keyword hypotheses (false alarms). This may be one single "garbage" model matching all non-keywords or even a vocabulary of non-keywords.

A speech recogniser in keyword-spotter mode has the advantage of being relatively light-weight as it does not use a computationally costly language model. Therefore, keyword spotting was a feasible approach at times when computer power was still limited. In early systems, keywords were usually carefully fixed in advance. After the keyword spotting process was performed, the spoken documents in the collection could be represented in terms of the keywords found in the documents. Although, this method worked well within a very restricted domain (such as the detection of weather reports) or topic identification in speech messages, the fixed set of keywords often appeared to be too limited for realistic tasks.

As computer power increased, keyword spotting could also be deployed at retrieval time, enabling the search for any keyword given by the user. However, keyword spotting at retrieval time may result in unacceptable delays in response time, especially when the document collection is large. To avoid this, an alternative word spotting technique called phone lattice scanning (PLS) can be deployed. In PLS word spotting, phone lattices are created and searched for the sequence of phones corresponding to a particular search term. In this way keywords do not need to be chosen a priori so that any set of words

can be searched, and as the phone lattices are created before retrieval time, delays in response time can be minimised.

But using keyword spotting for retrieval purposes has disadvantages. Retrieval will suffer from false alarms and missed keywords and especially short words are hard to spot as keyword spotting relies solely on acoustic information. This attracted SDR researchers to use large vocabulary speech recognition systems (LVCSR, discussed below) that can benefit from the restrictive power of language models or to combine other speech recognition techniques with word spotting. Especially when the mismatch between speech recognition vocabulary and domain vocabulary is hard to model and tends to produce many out-of-vocabulary words, having word spotting functionality available as an ad-hoc tool for searching either the audio directly or a phone or phone lattice representation of the document be profitable. A typical example of a deploying word spotting approach in combination with a full text transcription approach would be the following strategy to recover names that were misrecognised: (i) the initial speech recognition transcript is used to find related collateral text data; (ii) named entity detection in the collateral data source provides relevant named entities given the document topic; (iii) the occurrence (and timings) of these named entities in the source data are recovered using a word spotting approach.

In spite of its disadvantages, keyword spotting can be regarded as a useful technique for the retrieval of spoken documents. The focus of the SDR community however shifted toward large vocabulary speech recognition in the late nineties due to massive research efforts resulting in substantial improvements in speech recognition performance in SDR. But utilising word spotting techniques, either alone or in combination with other speech recognition techniques, remains a good choice for a variety of applications. Especially when heavy-weight speech recognition is not feasible or useful.

SDR using sub-word unit representations

While keyword spotting and LVCSR approaches largely focus on words as representation units of the decoded speech in the document, an alternative category of SDR approaches use sub-word unit representations such as phones, phone n-grams, syllables or broad phonetic classes to deal with the retrieval of spoken documents. Sub-words are generated by either taking the output of a phone recogniser directly (phones) or by post-processing this output to acquire phone N-grams or other representations. A significant characteristic of sub-word based approaches is that the document is represented in terms of these sub-word units. At retrieval time, query words are translated into a sequence of sub-word units which are matched with sub-word document representations.

Note that keyword spotting using a phone lattice as described earlier, resembles this type of approaches in the way that the query is translated into a sequence of sub-word units, namely phones, that are matched with the phone

representation of the documents. However, keyword spotting aims at matching particular sequences of phones in the document representations themselves in order to map them to words, whereas in sub-word based approaches, the matching is done using sub-word indexing terms.

As a phone recogniser requires only an acoustic model and a small phone grammar to generate sequences of phones, the recognition process can do with a relatively simple decoding algorithm. Compared to computationally expensive large vocabulary speech recognition approach, the decoding step of a sub-word based approach is much faster. Also, by deploying a phone recogniser, collecting large amounts of domain specific text data (that may be unavailable) for language model training can be circumvented, which reduces training requirements to the acoustic model training. Finally, as the phone recogniser does not need a vocabulary of words, a sub-word based approach is less sensitive to out-of-vocabulary words, provided that the query words can be converted to the sub-word representations using grapheme-to-phoneme conversion tools.

However, depending solely on acoustic information, phone recognition systems tend to produce higher error rates, resulting in less accurate document representations. To compensate for the decrease in precision, hybrid approaches have been proposed where for example the sub-word unit approach is used as a pre-selection step for a word spotting approach.

1.4 Robust speech recognition and retrieval

If speech recognition technology is deployed to support retrieval tasks, the recognition accuracy must be analysed in this context. Whereas in dictation systems for example, it is of utmost importance to have a high speech recognition accuracy level for all words, for retrieval purposes it is important to have at least the content words right, e.g., nouns, proper names, adjectives and verbs. During the indexing process, function words (articles, auxiliary verbs, etc.) are filtered out anyway. The usually smaller function words bear less acoustic information and therefore have a high change to be misrecognised by speech recognition systems. Therefore, analysing the global word error rate of an ASR system for evaluating its feasibility for retrieval may not always be adequate. The word error rate is based upon a comparison of a reference transcription of the test material with the output of the recogniser referred to as the hypothesis transcription. The scoring algorithm searches for the minimum edit distance in words between the hypothesis and reference and produces the number of substitutions, insertions and deletions that are needed to align 'the reference with the hypothesis. The word error rate is then defined as:

$$\mathrm{WER} = \frac{\mathrm{Insertions} + \mathrm{Deletions} + \mathrm{Substitutions}}{\mathrm{Totalwordsinreference}} \tag{1.8}$$

Disregarding words that are in a list of stop-words during evaluation is one of the strategies that can be used for scaling the word error rate. Another

strategy is to compute the term error rate, that is defined as:

$$\text{TER} = \frac{\sum_{t \in T} |R(t) - H(t)|}{|T|} \tag{1.9}$$

where $R(t)$ and $H(t)$ represent the number of occurrences of query term t in the reference and the hypothesis respectively, and $+|T| = \sum_{t \in T} R(t)$. The TER gives a more accurate measure of speech recognition performance conditioned on a retrieval system as it takes only the mis-recognised query terms into account.

In practice, the word error rate is nevertheless frequently used as an indication of quality. In general, a word error rate of 50% is regarded as an adequate baseline for retrieval. However, for certain collections, it can be hard enough to achieve this goal of having at least half of the words right. In the following section, we describe a strategy that aims at improving retrieval based on noisy speech recognition transcripts itself by making use of parallel text corpora (section 1.4.1). Next, three strategies to reach optimal speech recognition are described: optimisation of the speech recognition vocabulary in section 1.4.2, generation of topic specific language models in section 1.4.3, and acoustic model adaptation in section 1.4.4.

1.4.1 Query and document expansion

A technique applied in information retrieval that is specifically worth mentioning in the context of spoken document retrieval, is query expansion. As the term already suggests, this technique simply adds words to the query in order to improve retrieval performance. In query expansion the document search is basically performed twice. After an initial run, a selection of the top N most relevant documents generates a list of terms ranked by their weight (e.g., a tf.idf weight). The top T terms of this list are then added to the query and the search is repeated using the enriched query. Query expansion can be performed using retrieved documents from the same collection, or using retrieved documents from another (parallel) corpus. In the former case, query expansion is referred to as blind relevance feedback, in the latter it is called parallel blind relevance feedback. As the speech recognition system in a spoken document retrieval task may have produced errors or may have missed important words, it can be useful to apply parallel blind relevance feedback using a corpus without errors—such as a manually transcribed corpus—in order to reduce retrieval misses due to speech recognition errors. In other approaches to query expansion, compound words are split, geographic names are expanded (e.g., "The Netherlands" to "Amsterdam, ..., Zaandam") and hyponyms of unambiguous nouns are added (e.g., "flu, malaria, etc." are added given "disease") using thesauri and dictionaries. Also the opposite approach, document expansion is applied to alleviate the effect of speech recognition errors on retrieval performance. However, this approach does not work that well when story segmentation is unknown.

1.4.2 Vocabulary optimisation

For successful retrieval, the minimisation of out-of-vocabulary (OOV) words in the speech-to-text conversion step is important. OOV words may result in OOV query words (QOV): words that appear in a user's query and also occurred in the audio document but – as they were OOV – could not be recognised correctly. OOV's damage retrieval performance in two ways: firstly, given a query with a QOV word, the QOV word leads to a word miss in searching. Secondly, its replacement potentially induces a false alarm for other queries. Document expansion and query expansion techniques are often deployed to compensate for QOV's in information retrieval. However, especially when OOV words are named entities, attempts to minimise the number of OOV words beforehand is beneficial. To enable the selection of an appropriate set of vocabulary words, typically text data that closely resemble the task domain are deployed to obtain an indication of the word usage in the task domain. The broadcast news (BN) domain is a relatively "open" with respect to word usage. Predicting exactly which words are to be used in news items is virtually impossible and as a result of this, the usual approach is to include as many words as possible in the vocabularies so that at least the majority of the words occurring are covered. The maximum number of words that can be included in the vocabulary is restricted by the number of words a speech recognition system can deal with, typically 65 K words[9] But as news topics are constantly changing, it is necessary to revise the selection of vocabulary words with regular intervals. By doing so, words that have shown an increased news value due to recent events, but were not in a vocabulary created earlier, can be recognised as well. On the other hand, words that become outdated, such as for example the name of a former minister that has a very low chance to appear in the news again, should be removed in order to avoid that these obsolete words are confused acoustically with other words in the vocabulary.

In the figures below the changing news value of words are visualised. Relative frequency statistics of word occurrences in a newspaper database are plotted in time (1999-2003). Figure 1.4 shows the decrease in news value of the word "Clinton" in favour of "Bush" whereas Figure 1.5 shows the word "poederbrief" (letter containing possibly poisonous powder) that suddenly appears after the 9/11 terrorist attacks. In the first example, one could think of removing "Clinton" from the vocabulary after a certain period of time. The appearance of word in the second example however cannot be "foreseen" (it suddenly appears) and can only be incorporated in the speech recognition vocabulary after it is first seen. Another category of words appear only (or mostly) in certain times of the year such as "Santa-Claus", "Christmas" and

[9] The reason for limiting the vocabulary to 65 K words in large vocabulary speech recognition is often for efficiency reasons. Words in the language model are represented by an integer index, which fit in 16-bit integers when the vocabulary size is limited to $2^{16} = 65536$.

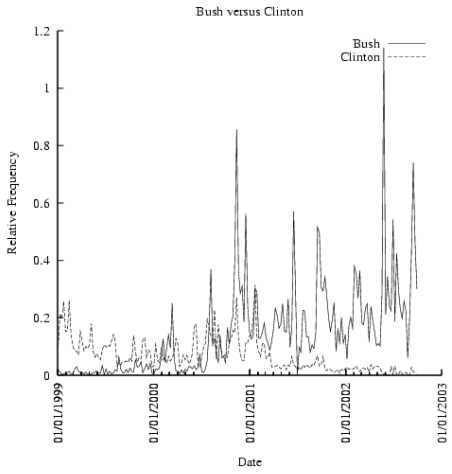

Fig. 1.4. Relative frequency statistics plotted in time of the words "Clinton" and "Bush". "Clinton" is showing a decreasing news value, whereas "Bush" is increasing.

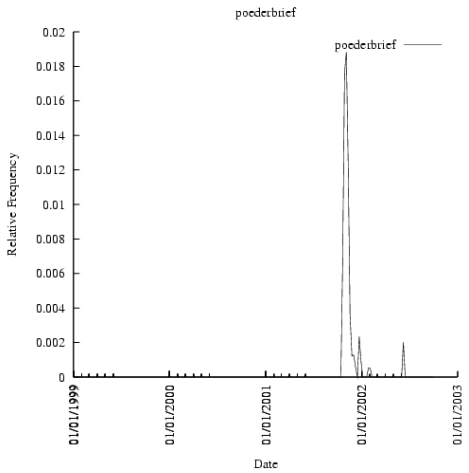

Fig. 1.5. Relative frequency statistics plotted in time of "Poederbrief" showing that word sudenly appears in the news.

"Prinsjesdag"(See Figure 1.6). One could decide to include such words in the vocabulary only during the relevant periods with a certain overlap.

In order to use a dynamic vocabulary that is updated on the basis of the news value of words, a parallel text corpus is needed for generating word occurrence statistics. An obvious approach is to extract text data from internet news sites on a daily basis. Having such a corpus available a vocabulary selection strategy has to be chosen. Typically new words that exceed a certain

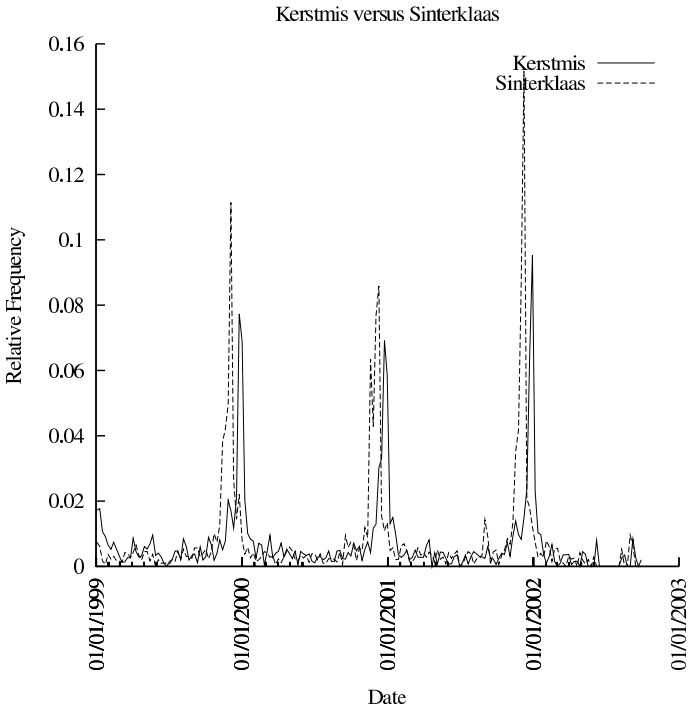

Fig. 1.6. Relative frequency statistics plotted in time of "Christmas" and "Santa-Claus".

frequency threshold are added to the vocabulary but other, more fine-grained strategies are conceivable, depending on available parallel text corpora. Note however that for commercial applications, intellectual property rights (IPR) can make the exploitation of text corpora difficult.

1.4.3 Topic-based language models

The use of topic-based language models is a somewhat different approach towards domain adaptation on the word level then the vocabulary selection approach discussed in the previous section. Here, instead of selecting words that are expected to appear in a task domain globally, words are selected with a focus on a specific segment of an audio document. Moreover, topic-based language model also try to incorporate those n-gram that are specific for a certain topic. A very simple example would be n-grams concerning the word 'bank': in a financial topic the four-gram 'go to the bank' is more likely then 'sit on the bank'. A financial-specific language model should reflect these topic specific statistics. More generally, one could interpret topic-specific language models as attempts to model topic-specific 'matters of speaking', more accurately.

Building topic-based language models in principle requires five steps:

- segmentation of the audio file
- initial speech recognition on the audio segments
- defining the 'topic' on the basis of the speech transcripts
- creating a topic specific language model
- final speech recognition run using topic-based language model

Ideally a segmentation of the audio file in order to apply topic-based language models results in a number of subsequent segments that can be interpreted to be on one single topic. In practice, real topic segmentations are usually not known a-priori so that often readily available segmentations, such as on the change of speaker, on longer silence intervals or even fixed time-windows (see also section 1.5 below), are chosen to divide the audio document in smaller parts. These parts are then further regarded as representing single topics. For each segment, a baseline speech recognition system provides an initial textual transcription.

On the basis of the speech transcript a topic must be assigned to the segment, either implicitly or explicitly. An explicit topic assignment refers to using specific topic labels, for example generated on the basis of a topic-classification system that assigns thesaurus terms. From a collateral text corpus (e.g., a newspaper corpus) that is labelled with the same thesaurus terms documents that are similar to the topic in the segment can be harvested for creating a topic-specific language model. For implicit topic assignment, an Information Retrieval system is used for the selection of documents from an unstructured collateral text source that have a similar topic: on the basis of the speech transcript (stop-words removed) that serves as a query, a ranked list of similar documents is generated; the top N documents of the list in turn serve as input for language modelling. Having created a topic specific language model a second speech recognition run is performed on the same segment with the new language model to generate the final transcript. The procedure is visually depicted in Figure 1.7.

A drawback of the procedure is that the two recognition runs, the search for related text documents and building the language models takes quite some time. When the topics are broadly defined (e.g., economics, sports, etc.) the language models could best be created a priori therefore. Note also that in an alternative set-up, the segmentation can in theory be done on the basis of topic segmentation on the speech transcript of the complete audio document. However, as the initial speech transcripts has errors, the accuracy of the text-based topic segmentation may be low.

1.4.4 Acoustic adaptation

A robust speech recognition system can be defined as a system that is capable of maintaining good recognition performance even when the quality of the speech input is low (environment, background noises, cross-talk, low audio

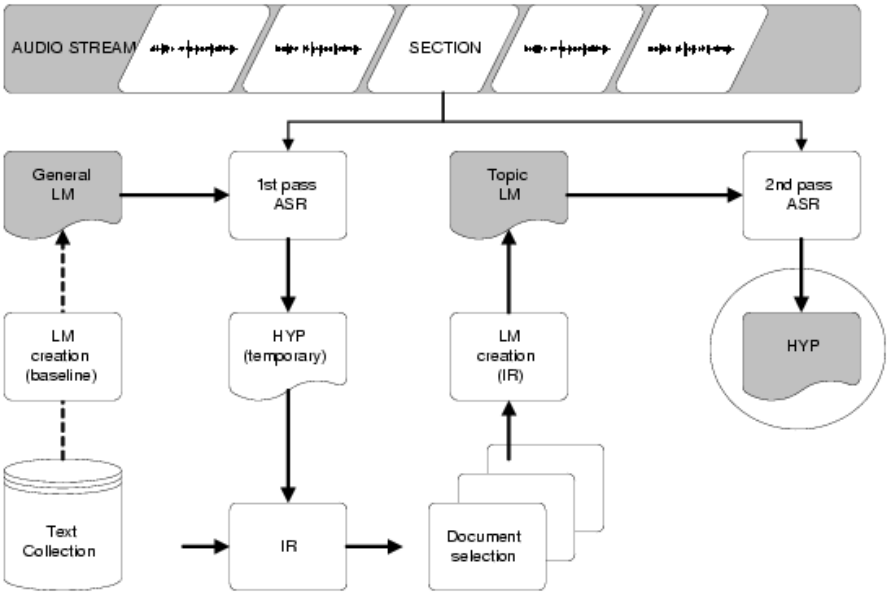

Fig. 1.7. Creation of topic based LM

quality) or when the acoustical, articulatory, or phonetic characteristics of the speech encountered in the training data differ from the speech in the task data. Speaker-to-speaker characteristics may vary enormously due to a number of factors, including:

- vocal tract length difference (gender)
- age
- speaking style (pronunciation, speed)
- regional accents
- emotion
- non-nativeness

Even systems that are designed to be speaker independent cannot cover all the speaker variations that may occur in the task domain. A possible strategy to deal with the variation encountered in the task domain would be to collect domain specific data for additional training in order to capture the environment characteristics or, when there are only a few speakers, train speaker dependent models. This strategy however has some drawbacks. Firstly, setting up a domain specific training collection is costly. For acoustic training purposes the speech data needs to be annotated on the word level, which takes an experienced annotator approximately 10 hours for every hour. Secondly, the variability in the task domain can simply be too large for additional train-

ing to be successful (multiple acoustic conditions, large number of speakers of varying signature, etc).

The alternative of doing additional training is to apply normalisation procedures (speaker normalisation, noise suppression) and dynamic adaptation procedures. The idea is to start with a general, relatively stable baseline system and tune this system to the specific conditions in the task domain automatically. Below, three frequently used techniques are discussed in brief.

Vocal tract length normalisation

The vocal tract is the area in between the lips and the glottis. It is often imagined as a tube which length and shape have a determining effect on the resonance characteristics and hence the characteristics of the speech. The length of the vocal tract of speakers differs. The average length for white American adult males is said to be 17 cm, but this varies strongly with the physical dimension of the person. A short vocal tract tends to result in formants at a higher frequency and long vocal track lengths with lower formant frequencies.

Over the past 10 years VTLN has become a standard normalisation technique in speaker-independent speech recognition. With vocal tract length normalisation (VTLN) the aim is to compensate for the acoustic difference due to vocal tract length by normalising the spectra of speakers or clusters of speakers to that of a "generic" speaker during training and testing. The normalisation is done by warping the frequency axis of the spectra by an appropriate warp factor prior to the feature extraction procedure. Different warping techniques have been reported: frequency warping both linear and exponential non-linear and Bark/Mel scale warping.

MAP and MLLR adaptation

Whereas with VTLN the adaptation is done by normalising spectral information (feature space normalisation), other adaptation methods aim at adjusting the model parameters (model-space transformation). The advantage of the model-space transformation is that the normalisation has to be performed only once instead of every time new speech input has to be decoded. A disadvantage however is that one may end up with a variety of adapted models.

Model adaptation can be done off-line or at preparation time and online, at runtime. Off-line (or batch) adaptation refers to situations in which is known that the acoustic model has to be adapted for one single speaker or acoustic condition (typically in dictation tasks). The approach here is to collect as little adaptation data as possible to achieve an acceptable performance as collecting the adaptation data can be expensive (as discussed above). In online adaptation, the adaptation is during at recognition time. As a consequence only very little data is available and the adaptation algorithms should not be

too complex in order to avoid huge delays. Often, online-adaptation requires multiple decoding passes.

As in the context of spoken document retrieval on-line adaptation is most needed, we restrict ourselves to this adaptation mode. A very effective and popular model adaptation technique is Maximum Likelihood Linear Regression (MLLR). With MLLR estimation the aim is to capture the general relationship between the speaker independent modal set and the current speaker by transforming the model means to fit the adaptation data. This is done by estimating a global linear model transformation matrix in order to maximise the likelihood of generating the adaptation data.

In MLLR adaptation clusters of model parameters are transformed simultaneously using a shared function that is estimated from available adaptation data. Because of this sharing, transformation-based adaptation techniques are especially attractive in situations where the amount of adaptation data is limited. MLLR adaptation is an indirect model adaptation approach. Direct model adaptation techniques do not assume any underlying functional transformation. Here, acoustic units are re-estimated for which adaptation data is available. As acoustic units that are not observed in the adaptation are not modified, this type of adaptation leads to local adaptation. In direct adaptation Bayesian learning, often implemented via maximum a posteriori (MAP) estimation, is a commonly used approach. MAP adaptation combines the information provided by the adaptation data with some prior knowledge about the model parameters described by a prior distribution. When the amount of adaptation data increases, MAP converges slowly to maximum likelihood estimation. A large amount of adaptation data is needed however to observe a significant performance improvement.

1.5 Audio Segmentation

Although time-labelled speech transcripts can directly be used to identify relevant items in within an audio collection, segmenting of longer audio documents is a helpful intermediate step. Segmentation can be done according to a particular condition such as speaker, speech/non-speech, silence, or even topic, into homogeneous sub-documents that can be accessed individually. This is convenient, as scrolling through a large unstructured audio or video fragment to identify interesting parts can be cumbersome. Audio segmentation can be advantageous from a speech recognition point of view as well, as it allows for segment based adaptation of the recognition models as will be discussed below. A frequently applied adaptation scheme is based on speaker identity.

Using a fixed overlapping time window, or fixed number of words to segment an audio stream is a simple but in cases very effective segmentation approach that does not rely on special segmentation tools. When the window and overlap ranges are chosen well, it can provide a document structure that can already usefully be deployed for certain retrieval tasks, such

as word-spotting. But a segmentation based on audio features is much more informative and helpful both from a retrieval and speech recognition point of view. With a segmentation according to speaker for example, a retrieval results can be structured and presented according to speaker identity (using an ID or even a name when combined with speaker identification). In addition speaker dependent modelling schemes can be applied in order to improve speech recognition performance. Useful segmentation cues are in general provided by techniques that aim at the labelling of the source of audio data (e.g., acoustic environment, bandwidth, speaker, gender), often referred to as 'diarisation' or 'non-lexical information generation'.

1.6 Cross-media mining

In section 1.3.3 the alignment of collateral data such as subtitling information to the audiovisual source was discussed as a convenient metadata generation approach. Ideally one would not only synchronise audiovisual material with content that approximates the speech in the data such as with subtitles or minutes, but take even one step further and exploit any collateral textual resource, or even better: any kind of textual resource that is accessible, including open source titles and proprietary data (e.g., trusted web-pages and newspaper articles). Another way of putting it is to shift the focus from indexing individual multimedia documents to video-mining in truly multimedia distributed databases. In the context of meetings for example, usually an agenda, documents on agenda topics and CVs of meeting participants can be obtained and added to the repository. Mining these resources can support information search because it yields annotations that offers the user not just access to a specific media type, but also different perspectives on the available data. An agenda could help to add structure that can for example be presented in a network representation, whereas CVs can be linked to annotations resulting from automatic speaker segmentation. In addition, both documents and CVs would allow for multi-source information extraction.

A typical example of what the cross-media perspective can yield in the broadcast news domain is the linking of newspaper articles with broadcast items and vice versa. Links can be established between two news objects which count is similar on the basis of the language models assigned to them via statistical analysis. Typically such language models are determined by the frequency of the linguistic units such as written or spoken words and their co-occurrences. The similarity between two documents can be decided for each pair of documents, but a more common approach is to pre-structure a document collection into clusters of documents with similar language models.

Similarity of language models predicts similarity of topic, and therefore this technique is known as topic clustering.[10]

In addition to linking documents with a similar topic profile, which can be supportive in a browser environment, also the available semantic annotation for documents with similar profiles can be exchanged and exploited for conceptual search. If a newspaper article has been manually classified as belonging to e.g., economy or foreign politics, a broadcast item with a similar language model can be classified with these conceptual labels as well.

1.7 Summary

In this chapter the focus was on the automatic extraction of information from the speech in multimedia documents. The larger part of this chapter was concerned with the use of speech recognition technology for automatic metadata extraction. After a brief introduction to speech recognition a number of spoken document retrieval techniques have been discussed: alignment of available textual data sources, keyword spotting, sub-word unit based approach, and finally large vocabulary speech recognition. It was shown that for certain domains acquiring a speech recognition performance that is suitable for retrieval purposes can be hard and that approaches aiming at robust speech recognition on the one hand, and search error minimisation on the other hand can be deployed in such difficult domains. The chapter was finalised by discussing some properties of segmentation, an important topic from both a information retrieval and speech recognition development point of view, and by introducing the concept of cross-media mining where the focus is less on searching (an index of) a multimedia database, but on enabling new views on the data by cross-linking all the available multifaceted information sources in a multimedia database.

1.8 Further Reading

- "Speech and Language Processing, An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition", Daniel Jurafsky and James H. Martin
- "Statistical Methods for Speech Recognition", Frederick Jelinek

[10] The functionality commonly known as *topic detection and tracking* (TDT) for dynamic news streams has been built upon it and plays a central role in the evaluation series for TDT organised by DARPA.