SIGIR 2007 Workshop

Searching Spontaneous Conversational Speech

held in conjunction with the 30th Annual International ACM SIGIR Conference

27 July 2007, Amsterdam

Organizers

Franciska de Jong
University of Twente, The Netherlands
Douglas Oard
University of Maryland, USA
Roeland Ordelman
University of Twente, The Netherlands
Stephan Raaijmakers
TNO ICT, The Netherlands

CIP GEGEVENS KONINKLIJKE BIBLIOTHEEK, DEN HAAG

Jong de, F.M.G., Oard D.W., Ordelman R.J.F., Raaijmakers S.A.

Proceedings of the ACM SIGIR Workshop
'Searching Spontaneous Conversational Speech',
held in conjunction with the
30th Annual International ACM SIGIR Conference
27 July 2007, Amsterdam

ISBN: 978-90-365-2542-8

Centre for Telematics and Information Technology, Enschede, The Netherlands

contact: hmi_secr@ewi.utwente.nl
workshop url: http://hmi.ewi.utwente.nl/sscs

Preface

Nearly a decade ago, we learned from the TREC Spoken Document Retrieval (SDR) track that searching speech was a "solved problem." Three factors were key to this success:

- Broadcast news has a "story" structure that resembles written documents.
- The redundancy present in human language meant that search effectiveness held up well over a reasonable range of transcription accuracy.
- Sufficiently accurate Large-Vocabulary Continuous Speech Recognition (LVCSR) systems could be built for the planned speech of news announcers.

The long-term trend in speech recognition research has been toward transcription of progressively more challenging sources. Over the last few years, LVCSR for spontaneous conversational speech has improved to a degree where transcription accuracy comparable to what was previously found to be effective for broadcast news can now be achieved for a diverse range of sources. This has inspired a renaissance in research on search and browsing technology for spoken word collections in several communities, including those focused on:

- Archived cultural heritage materials (e.g., interviews and parliamentary debates).
- Discussion venues (e.g., business meetings and classroom instruction).
- Broadcast conversations (e.g., in-studio talk shows and call-in programs).

Clearly we will lack as accessible a "story" structure in at least some of these application as we had in broadcast news. Some of these applications pose new challenges for core speech technology such as acoustic channel adaptation and language modeling. And in some cases users seeking to craft highly selective queries may take us outside the LVCSR lexicon, or into aspects of language use for which we simply lack enough evidence to build good statistical models. We need to learn how to both ask and answer these new questions.

Test collections are being developed in individual projects around the world, and some comparative evaluation activity for searching spontaneous conversational speech has developed. The time now seems right to look more broadly across the interested research communities for potential synergies that can help to shape their information retrieval research agendas by sharing ideas and resources. This ACM SIGIR Workshop on Searching Spontaneous Conversational Speech aims to help to foster that discussion. The papers contained in this proceedings volume reflect some of the emerging focus areas and cross-cutting research topics, together addressing evaluation metrics, segmentation methods, workflow aspects, rich transcription, and robustness. We would like to thank all of the authors who submitted papers for the hard work that went into their submissions, and the members of the programme committee for their thorough reviews. Special thanks go to IST project AMIDA (http://www.amidaproject.org) and SRO NICE (CTIT: http://www.ctit.utwente.nl/research/sro/nice/) for their generous support.

Franciska de Jong, Douglas W. Oard
Roeland Ordelman, and Stephan Raaijmakers
July 2007

Programme Committee

Samy Bengio (Google)
Herve Bourlard (IDIAP)
Sadaoki Furui (TITECH)
Marcello Federico (FBK-IRST)
Jon Fiscus (NIST)
John Garofolo (NIST)
Sam Gustman (USC)
Thomas Hain (Sheffield)
John Hansen (UT Dallas)
Alex Hauptmann (CMU)
Julia Hirschberg (Columbia)
Diana Inkpen (Ottawa)
Gareth Jones (DCU)
David van Leeuwen (TNO)
Lori Lamel (LIMSI)
Christian Müller (ICSI)
Steve Renals (Edinburgh)
Salim Roukos (IBM Research)
Elizabeth Shriberg (SRI and ICSI)
Rainer Stiefelhagen (Karlsruhe)
Alessandro Vinciarelli (IDIAP)

Sponsors

Contents

Searching Conversational Speech

Mark Maybury
Information Technology Center
The MITRE Corporation
202 Burlington Road
Bedford, MA 01730, USA
maybury@mitre.org
Tel: (781) 271-7230
itc.mitre.org

ABSTRACT
In this presentation we begin by summarizing the challenges posed by searching spontaneous conversational speech. We then summarize two MITRE efforts to address this challenge. We first describe Audio Hot Spotting (AHS) and then Cross Language Automated Speech Recognition (CLASR). We conclude by outlining some promising opportunities for future research.

THE CHALLENGE OF CONVERSATIONAL SPEECH

Searching conversational speech is a grand challenge. Telephone conversations alone illustrate the scale of the challenge with over a billion fixed lines worldwide creating 3,785 billion minutes (63B hours) of conversations annually, equivalent to about 15 exabytes of data (ITU 2002). Add to this voluminous and highly variable mobile and wireless communication. In addition, 47,776 radio stations add 70 million hours of original radio programming per year, some of which are conversations. Add to this 31 million hours of original television programming/year from 21,264 stations (CIA World Factbook 2002). To further complicate this, we are faced with thousands (approximately 6,800) of languages and as much as ten thousand dialects globally. While the potential human need to search audio is unknown, even if only a small portion of the roughly 5 billion searches per month on the Internet (searchengine.com) could be satisfied from spoken language collections, this still remains a large requirement.

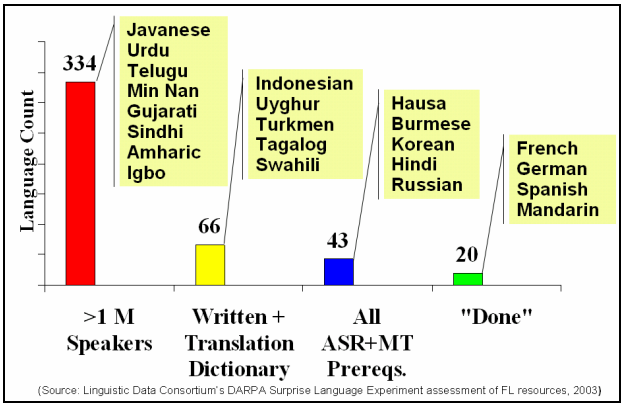

Figure 1. Foreign Language Spoken Language Needs

As Figure 1 illustrates, there are over 300 spoken languages with more than one million speakers but only 66 of these are written and for which we have a translation dictionary. Of these, we have ASR and MT for only 44, and only 20 of these are considered "done" in the sense that systems exists for automated transcription and translation.

In addition to the challenge of lack of written materials, which we will return to subsequently, there are many challenges beyond scale. These include challenges with language in general, such as polysemy, ambiguity, imprecision, malformedness, intention, and emotion. And in addition to the traditional set of challenges with automated speech recognition such as noise, microphone variability, and speaker disfluencies, the kind of conversational speech that occurs in telephone calls, meetings, interviews has additional challenges including:

- *Multiparty* – multiple, interacting speakers
- *Human and Machine* – can include human-human conversations and human-machine conversation
- *Talkover* – multiple simultaneous speakers talk over speaker turns
- *Spontaneity* – unpredictable shifts in speakers, topics, and acoustic environments.
- *Diverse genre* – Conversation is found in many venues including meetings, radio/TV talk shows, interviews, town halls, debates, presentations which vary in degree of structure, roles of participants, lengths, degree of formality, and so on.
- *Multiple media* – conversational speech is found in audio, video
- *Real time and retrospective* – access during the speech event or after
- *Tasks*: document routing, (doc/passage/fact) retrieval or question/answering, browsing, tracking entities and events , summarization (e.g., speakers, topics)
- *Multilingual* – multiple languages, sometimes from the same speaker
- *Acoustic challenges* – spoken conversations often occur over cell phones or hand held radios which come in and out of range and have highly variable signal to noise ratios.

Keynote for Workshop on Searching Spontaneous Conversational Speech
2007 International Conference on Information Retrieval (SIGIR-07).
27 July 2007 Amsterdam, Netherlands.

- *Non-acoustic conversational elements* – speakers use clapping, laughing, booing, whistling, and other sounds and gestures to express agreement, disagreement, enjoyment, and other emotions.
- *References*: Since conversations are often performed in a physical context, the language often contains references to items therein (exophora)

Compounding these challenges, expert translators, particularly for low density languages are expensive and scarce.

In addition to the challenges with speech, there are many retrieval challenges such as query formulation, query expansion, query by example, results display, browsing, and so on.

AUTOMATED PROCESSING OF CONVERSATIONS

Figure 2 illustrates the significant progress made over the years in spoken language processing. The Figure shows best systems each year in competitions administered by NIST to objectively benchmark performance of speech recognition systems over time. The graph reports reduction of word error rate (WER) over time. The systems were assessed on a wide range of increasingly complex and challenging tasks moving from read speech, to broadcast (e.g., TV, radio) speech, to conversational speech, to spontaneous speech, to foreign language speech (e.g., Chinese Mandarin, Arabic). Over time tasks have ranged from understanding read Wall Street Journal text, to understanding foreign television broadcasts, to so called "switchboard" (fixed telephone and cellphone) conversations. Future plans include meeting room speech recognition (NIST; Zechner and Waibel 2000).

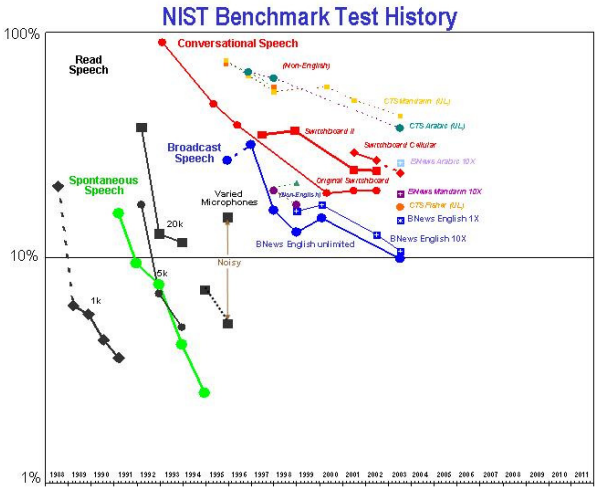

Figure 2. NIST Benchmarks Over Time
(http://www.nist.gov/speech/history)

As Figure 2 illustrates, while recognition rates for English (clean, well-formed, single speaker, speaking clearly to computer) are well below 10% word error rates. For example, computers can achieve understand someone reading the Wall Street Journal with a 5% Word Error Rate (WER) (1 word in 20 wrong). Conversations are harder, with broadcast news often achieving only a 15-20% WER and the CALLHOME Data Collection (Phone calls) achieving 30-40% WER.

HUMAN LANGUAGE TECHNOLOGY AT MITRE

MITRE has engaged in a highly diverse HLT program over multiple decades. This has resulted in operational systems of integrated capabilities such as the DARPA MITRE Text and Audio Processing System (MiTAP) and the Translingual Instant Messaging (TrIM). MITRE has made a number of its contributions available via open source. These include:

- DARPA Galaxy Communicator architecture (~800 downloads at communicator.sourceforge.net)
- Midiki MITRE dialog manager toolkit (200+ downloads at midiki.sourceforge.net)
- Callisto annotation tool framework (~900 downloads at callisto.mitre.org)

We have also been active in facilitating the community to advance a number of key standards such as TIMEX2 (Ferro et al 2005), TimeML (timeml.org, Pustejovsky, et al. 2005), and more recently an effort to create SpatialML (2007). We have been awarded a patent for our effort in Broadcast News Navigation (US Patent 6,961,954 ; Maybury et al 1997) and have patents submitted for Personalcasting and Audio Hot Spotting.

AUDIO HOT SPOTTING

The Audio Hot Spotting project (Hu et al. 2003, 2004ab) aims to support natural querying of audio and video, including meetings, news broadcasts, telephone conversations, and tactical communications/surveillance. As Figure 2 illustrates, the architecture of AHS integrates a variety of technologies including speaker ID, language ID, non speech audio detection, keyword spotting, transcription, prosodic feature and speech rate detection (e.g., for speaker emotional detection), and cross language search.

An important innovation of AHS is the combination of word-based speech recognition with phoneme-based audio retrieval for mutual compensation for keyword queries. Phoneme-based audio retrieval is fast, more robust to spelling variations and audio quality, and may have more false positives for short-word queries. In addition, phoneme-based engines can retrieve proper names or words not in the dictionary (e.g., "Shengzhen") but, unfortunately, produces

no transcripts for downstream processes. In contrast, word-based retrieval is more precise for single-word queries in good quality audio and provides transcripts for automatic downstream processes. Of course it has its limitations too. For example, it may miss hits for phrasal queries, out-of-vocabulary words, and in noisy audio and is slower in pre-processing.

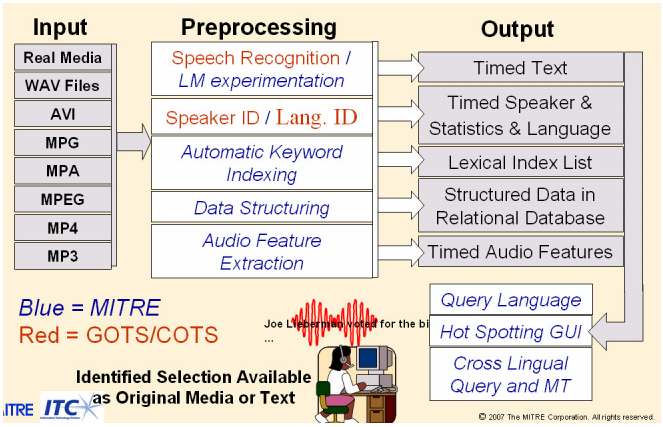

Figure 2. AHS Architecture

Figure 3 illustrates the user interface for speech search, and includes a speaker and keyword search facility against both video and audio collections. The user can also search by non speech audio (e.g., clapping, laughter).

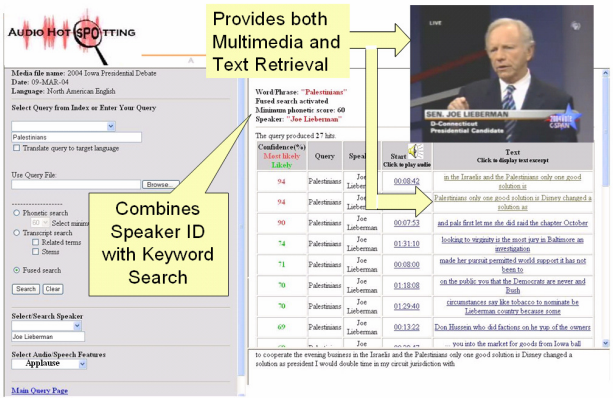

Figure 3. AHS Search Interface

A recent extension enables a user to query in English, have this query translated to a foreign language (e.g., Spanish, Arabic), use this query to retrieve hot spots in a transcription of the target media, which is then retrieved and translated into the query language. Figure 4 illustrates this in action. The user has typed in the word "crisis" which is subsequently translated into Arabic query term, أزمة, which is then used to search the target media which is subsequently translated as shown.

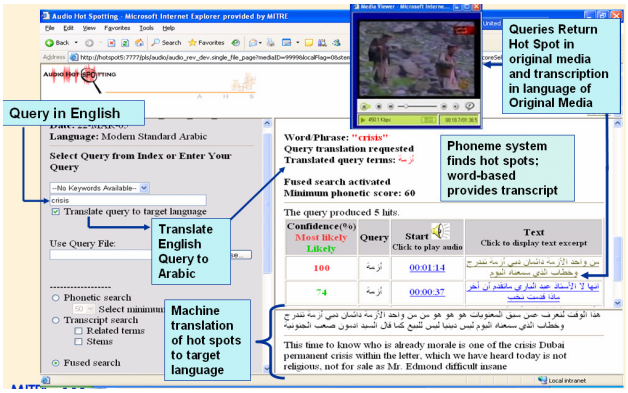

Figure 4. AHS Cross-Lingual Audio Hot Spotting

CROSS LANGUAGE ASR (CLASR)

Access to foreign language spoken discourse is challenging. Building systems to do so is even more difficult when there exist no written resources for that language. The Cross Language Automated Speech Recognition (CLASR) effort is investigating a new approach for spoken language translation of languages that lack significant written resources. This effort is exploring the hypothesis that recent advances in both speech recognition and machine translation enable a fresh approach.

In particular, CLASR aims to build a process that goes from audio in a foreign language to text in English, addressing languages that do not have the right quantity and type of language resources for the current approaches. Current approaches to this challenge go from source language acoustics to source language written form, then from the source language written form to the English written form. Typically they use 1-best ASR output although some use n-best, but in all cases they output written form.[1] CLASR simplifies this process and folds the translation model and acoustic model into one cross-language acoustic model.

While CLASR aims to address low resource languages, experiments are being performed on well-known languages (Spanish and Mandarin) to compare the new single stage approach to the traditional two stage pipe-line system, i.e., ASR+MT. In particular, CLASR uses an open source toolkit for ASR (HTK from Cambridge University) and a development kit for MT (GIZA++ and PHARAOH (JHU, MIT)). Our Spanish experiments are based on 30 hours of

[1] Siegler (1999) evaluated the IR performance of indeces based on different types of ASR output and found that the use of n-base lists was superior to individual word probabilities from lattices.

broadcast news audio using audio from Central America and transcripts in Spanish which have been translated into English. Initial results with Spanish with no additional language model have been promising as assessed by BLEU (BiLingual Evaluation Understudy), i.e., the portion of 4-word sequences in MT output that are found in reference translations with a range from 0 (poor) to 100 (good). The very first single-stage score, an initial foothold as we begin hill climbing, was a BLEU score of 8. By contrast, the 2-stage ASR+MT scores achieved a word error rate of 45 and & a BLEU score of 13. Our recent system Spanish-English MT system has a BLEU score of 21, outperforming the two stage baseline.

In summary, this approach is analogous to the results reported in this workshop by Olsson (2007) in which a single, integrated model outperforms a sequence of transcription and retrieval. Notably, CLASR's combined approach shows promise both performance-wise as well as in terms of its limited requirement for language resources.

RESEARCH OPPORTUNITIES

Spoken dialogue retrieval is an exciting research area precisely because it contains all the traditional challenges of spoken language processing together with the challenges imposed by the retrieval task. Some important spoken conversation processing challenges include:

- dealing with multiple speakers
- dealing with foreign language and associated accents
- incorporating non-speech audio dialogue acts (e.g., clapping, laughter)
- conversational segmentation and summarization
- discourse analysis, such as analyzing speaking rates, turn taking (frequency, durations), concurrence/disagreement which often provides insights into speaker emotional state, attitudes toward topics and other speakers, and roles/relationships.

Fiscus et al. (2007) in the workshop proceedings report spoken term detection evaluation results and note that scalability and domain independence remain areas for future evaluation.

Some important speech retrieval challenges include:

- How can we provide a query by example for a speech or audio signal, e.g., find speech that sounds (acoustically, perceptually) like this? (See Sound Fisher in Maybury 1997)
- How can we provide (acoustic) relevancy feedback to enhance subsequent searchers?
- How do we manage whole story/long passage retrieval which exposes users to too much errorful ASR output or too much audio to scan?
- Because text-based keyword search alone is insufficient for audio data, how do we retain and expose valuable information embedded in the audio signal?
- Are non-linguistic audio cues detectable and useful?
- Can we utilize speech and conversational gists (of sources or segments) to provide more efficient querying and browsing ?

Some interesting application challenges are raised such as dialogue visualization, dialogue comparison (e.g., call centers), or dialogue summarization. And, of course, the challenge of addressing speech and dialogue simultaneously.

ACKNOWLEDGEMENTS

This paper is dedicated to the memory of my doctoral supervisor Karen Spärck Jones for her generosity and indefatigable support of her students. This paper includes a summary of the contributions of Qian Hu's AHS project and John Henderson's CLASR project and their associated technical teams, all of whom deserve the recognition for these technical achievements. I thank Qian and John as well as Inderjeet Mani and Lisa Ferro for their comments on drafts of this paper.

REFERENCES

1. Ferro, L., Gerber, L., Mani, I., Sundheim, B. and Wilson G. (2005) "TIDES 2005 Standard for the Annotation of Temporal Expressions" April 2005, Updated September 2005. http://timex2.mitre.org/annotation_guidelines/2005_timex2_standard_v1.1.pdf
2. Fiscus, J., Ajot, J., Garofolo, J. and Doddington, G. Results of the 2006 Spoken Term Detection Evaluation. 2007 SIGIR Workshop on Searching Spontaneous Conversational Speech, Amsterdam, 27 July 2007. p. 45-51.
3. Hu, Q., Goodman, F., Boykin, S., Fish, R., and Greiff , W. 2004. "Audio Hot Spotting and Retrieval Using Multiple Audio Features and Multiple ASR Engines". Rich Transcription 2004 Spring Meeting Recognition Workshop at ICASSP 2004, Montreal, Canada. http://www.nist.gov/speech/test_beds/mr_proj/icassp_program.html
4. Hu, Q., Goodman, F., Boykin, S., Fish, R., and Greiff ,

W. 2004. "Audio Hot Spotting and Retrieval Using Multiple Features". Proceedings of the HLT-NAACL 2004 Workshop on Interdisciplinary Approaches to Speech Indexing and Retrieval. Boston, USA, pp. 13 - 17.

5. Hu, Q., Goodman, F., Boykin, S., Fish, R., and Greiff, W. 2003. "Information Discovery by Automatic Detection, Indexing, and Retrieval of Multiple Attributes from Multimedia Data", *The 3rd International Workshop on Multimedia Data and Document Engineering*, September 2003, Berlin, Germany, pp. 65-70.
6. Maybury, M., Merlino, A., and Morey, D. 1997. Broadcast News Navigation using Story Segments, ACM International Multimedia Conference, Seattle, WA, November 8-14, 381-391.
7. Maybury, M. (ed.) 1997. *Intelligent Multimedia Information Retrieval*. Menlo Park: AAAI/MIT Press. (http://www.aaai.org:80/Press/Books/Maybury-2/)
8. Maybury, M. editor. 2004. *New Directions in Question Answering*. AAAI/MIT Pres.
9. NIST Meeting Room Project: Pilot Corpus. http://www.nist.gov/speech/test_beds/mr_proj/meeting_corpus_1/
10. Olsson, S. Improved Measures for Predicting the Usefulness of Recognition Lattices in Ranked Utterance Retrieval. 2007 SIGIR Workshop on Searching Spontaneous Conversational Speech, Amsterdam, 27 July 2007. p. 1-5.
11. Pustejovsky, J., Ingria, B. Sauri, R., Castano, J., Littman, J., Gaizauskas, R., Setzer, A., Katz, G. and Mani, I. 2005. The Specification Language TimeML. In Mani, I., Pustejovsky, J. and Gaizauskas, R. (eds.), The Language of Time: A Reader, 545-557. Oxford University Press. http://timeml.org
12. Siegler, M. 1999. Integration of Continuous Speech Recognition and Information Retrieval for Mutually Optimal Performance. PhD Thesis. Carnegie Mellon University.
13. SpatialML: Annotation Scheme for Marking Spatial Expressions in Natural Language, March 30, 2007. MITRE Technical Report. http://sourceforge.net/projects/spatialml/
14. Zechner, K. and Waibel, A. 2000. DiaSumm: flexible summarization of spontaneous dialogues in unrestricted domains. Proceedings of the 18th Conference on Computational Linguistics, 968 - 974.

Improved Measures for Predicting the Usefulness of Recognition Lattices in Ranked Utterance Retrieval

J. Scott Olsson
Appl. Math. and Sci. Comp.
University of Maryland
College Park, MD 20742
olsson@math.umd.edu

ABSTRACT

We consider the problem of evaluating automatic speech recognition lattices to predict their usefulness in speech retrieval applications. In particular, we focus on ranking utterances by our confidence that they contain a query term. Our purpose is to close the gap between recognition efforts, which have traditionally focused on producing one-best transcripts, and recent retrieval systems, which may utilize multiple transcript hypotheses in indexing and search. We present a simple framework for comparing the ability of two measures to predict how well a system can retrieve a matching lattice. In a comparison with the traditional measure, simple accuracy (or word error rate), we show with statistical significance that two new measures are superior at predicting a vocabulary independent utterance retrieval system's rank ordering of speech utterances.

Categories and Subject Descriptors: H.3 [Information Storage and Retrieval]: Miscellaneous

General Terms: Design, Measurement, Experimentation

Keywords: speech utterance retrieval, keyword spotting, speech recognition lattices

1. INTRODUCTION

Early work in spoken document retrieval (SDR) focused largely on applying an available information retrieval (IR) system to the one-best transcript output of an available automatic speech recognition (ASR) system [1]. The complex components in this cascade approach were largely black boxes and often could only be superficially integrated. Advancements centered not on improving the integration of these technologies, but largely on coping for their deficiencies (e.g., document expansion techniques to recover from mediocre recognition transcripts).

This modular approach was a reasonable thing to do because speech people already knew something about producing these transcripts. Transcripts are, after all, the natural output for *human* users. Speech people also knew how to measure their performance at the task, using *word error rate* (WER) or *accuracy*. For the transcription problem, WER makes sense: given a transcript, WER gives something like the proportion of recognized words which need to be fixed. It's also simple to explain: given a reference and hypothesized transcript, with an alignment allowing for some insertions, substitutions, and deletions, we define $\text{WER} \equiv \frac{Ins+Sub+Del}{N}$, where N is the length of the reference transcript. Accuracy is $1 - \text{WER}$.

This resulted in IR people asking the question: how good does my WER need to be to do reasonable IR? Or if an IR person was speaking to a speech person, he might say: just make my WER smaller. The assumption, which we will soon re-examine, has often been that a smaller WER meant better retrieval.

In this paper, we focus on the problem of ranking short speech utterances by our confidence that they contain a query term. We refer to this problem as *ranked utterance retrieval*. We emphasize here that SDR cannot be reduced to simply the problem of keyword spotting or utterance retrieval. However, finding the occurrence of terms is at least a necessary component of every SDR system. More importantly, this limited focus affords us greater clarity when considering the evaluation of speech recognition output. In particular, we can avoid the mixing effects of full SDR (e.g., a misrecognized word may not hurt ranking if other query terms hit). Alternatively, we can think of utterance retrieval roughly as SDR with very short queries (an important difference being that we have only a trivial notion of "relevance").

2. A PROBLEM

Unfortunately, WER, a measures on transcripts, is poorly matched to the IR problem. Consider the four hypotheses for the sentence "The man is tall" below:

Sentence	mistake	WER
The man is the	1 Sub	.25
The man tall	1 Del	.25
The man is is tall	1 Ins	.25
The man is tall tall	1 Ins	.25

With respect to WER, these sentences are all the same. Clearly, for IR systems which count words—and care most about the informative ones, WER isn't adequately charac-

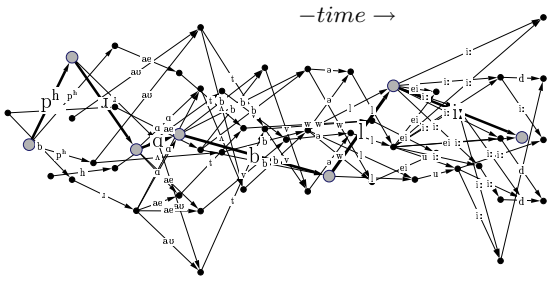

Figure 1: A phone lattice snippet containing the word "probably". Arc thickness is proportional to the transition probability. Large filled circles mark the most probable path.

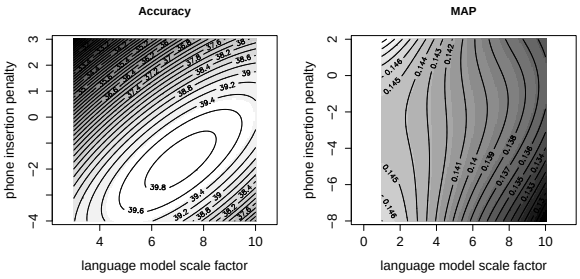

Figure 2: Second degree polynomial trend surfaces fit by least squares for (1) phone accuracy and (2) MAP. Notice that the greatest accuracy is near the middle (white space), while MAP is best (and getting better) in the top left.

terizing the utility of the different hypotheses.[1] While artificial, this example illustrates the difficulty: there is a mismatch between WER and the IR problem.

More recently, speech retrieval research has moved away from using only the one-best transcript of the audio. Just as cross-language IR can be improved by using probabilistic term frequency translation instead of a one-best hypothesis (i.e., by more fully capturing the uncertainty in translation), spoken-document IR can be improved by searching the *lattice* of utterance hypotheses rather than only the one deemed most probable. A lattice is just a simple way of representing the ASR system's uncertainty in prediction. Consider Figure 1. Moving from left to right, the lattice arcs show which phone was spoken and with what probability (according, in this case, to an acoustic and language model).

Recognition lattices might contain words or phones (or other things). This work focuses on phones, in part because phone recognition is hard. With phone recognition, there is no shortage of uncertainty to carefully represent in the lattice (so that we are particularly concerned with evaluating the utility of phone lattices). Phone (or subword) lattices are also especially promising for IR applications. Because they facilitate *vocabulary independent* term detection, we expect them to be more useful at finding the rarest (i.e., most informative) words. These words, because they are rare, are likely to fall outside of the fixed vocabulary of a conventional word-based ASR system (and the lattices it may produce).

It may be tempting to think of words and phones in lattices interchangeably (as though phones were just really short words), but this is a misleading simplification. It isn't enough to just estimate the expected number of occurrences of each phone in the lattice, as you might to produce a "bag of words" for indexing. Certainly, representing a spoken document as a "bag of phones" doesn't seem very useful. Another difference is that, with words, you might not worry too much about misrecognizing a few and recovering the downstream retrieval performance with tricks like document expansion. That technique won't directly translate to the phone case: the phones we recognize are only useful if they are recognized in the right context (that is, after all, where phones can convey meaning). This connectedness requirement for phone sequence hypotheses again motivates the evaluation of lattices.

Moreover, using lattices rather than only a one-best transcript dramatically changes how we must evaluate speech systems for IR. Keep in mind that WER is really a measure of the utility of a transcription. For the speech transcription problem, there may be a difference between "recognize speech" and "wreck a nice beach," but if both these hypotheses are captured in the lattice with some probability, an IR system which uses the lattices still has a fighting chance of finding your document. This is to say, the evaluation of recognition output should also consider (with an appropriate weighting) less probable utterance hypotheses.

The mismatch, we see now, is that one-best WER is a measure designed for humans (i.e., transcripts), not state of the art speech retrieval systems. And this mismatch occurs in the real world[2]. Suppose your speech guy is tuning a phone recognizer to maximize the phone accuracy for your phonetic-lattice based spoken utterance retrieval system. To measure the downstream IR performance, you sort the utterances by your confidence that they contain a query word and compute the average precision (AP) over the list. Averaging over many query words, this is mean average precision (MAP)[3]. The speech guy works on improving the phone accuracy by tuning two common parameters on the output lattices: the phone insertion penalty and the language model scale factor[4]. The results from such a scenario are plotted as trend surfaces in Figure 2. In the figures, the phone insertion penalty and language model scale factor are being varied as phone accuracy and mean average precision are

[1] Recently, [3] proposed some measures which indicate the "proportion of information communicated" by the hypothesis. Again, however, their focus was on a human reader, not producing input for an IR system (which may consider multiple hypotheses).

[2] The data for this example was produced using the utterance retrieval system described in [4] on a set of CallHome English telephone speech.

[3] This might seem a bit funny to IR people, who are used to averaging over *topics*. Just think of MAP as a way of evaluating something with good and bad items in a ranked order (in this case, short utterances).

[4] See [6] for definitions.

observed. Unfortunately, as we see here, you can spend a lot of time searching for the best accuracy (or WER), only to move away from the best MAP. Optima for the scores are at different settings of the parameters!

Unfortunately, due to the complexities of IR systems, it seems hopeless to search for a new speech evaluation measure whose optimum will exactly correspond to the best IR performance. At this point, we might throw up our hands, simply forget about evaluating the recognition component, and instead evaluate end to end: tune recognition parameters and observe MAP, tune parameters, observe MAP, and so on. Unfortunately, this is rather costly. A better choice is a new measure for the recognition component which is a better predictor for search performance.

3. NEW MEASURES

Retaining only the one-best hypothesis from recognition is fundamentally a precision-oriented strategy. But to maximize retrieval utility (as modeled by MAP), we also need to concern ourselves with recall. That is to say, we have to consider less probable paths through the lattices. The difficulty then, is determining how much weight to assign to these less-probable lattice traversals. In this study, we are concerned with choosing a measure on lattices that respects the need to consider non-best hypotheses.

3.1 Expected Accuracy

For IR systems which consider the entire lattice of speech hypotheses, a natural alternative to simple one-best accuracy is the *expected accuracy* over the lattice. Given a lattice L containing many paths ℓ (i.e., $\ell \in L$), the expected (phone) accuracy over all paths is

$$E_{P_L}[\text{Acc}] = \sum_{\ell \in L} P_L(\ell)\ \text{Acc}(\ell).$$

Here, $\text{Acc}(\ell)$ denotes the phone accuracy along one lattice path ℓ. The posterior distribution $P_L(\ell)$ is defined as

$$P_L(\ell) = \frac{\exp\{\sum_{\alpha \in \ell} S(\alpha)\}}{\sum_{\nu \in L} \exp\{\sum_{\beta \in \nu} S(\beta)\}},$$

where $\exp\{\cdot\}$ denotes exponentiation; we assume the score $S(\alpha)$ for an arc α on the path is a log probability (e.g., the sum of the acoustic and language model log probabilities). This distribution may be efficiently computed using a variant of the forward-backward algorithm. This functionality is currently supported by the SRI language modeling toolkit [5].

In practice, we estimate the expected accuracy by randomly generating M paths through the lattice (in this study, $M = 500$) and then averaging the accuracies computed for each traversal:

$$E_{P_L}[\text{Acc}] \approx \frac{1}{M} \sum_{i=1}^{M}\ \text{Acc}(\ell_i).$$

Here, ℓ_i is just the ith random traversal ($\ell_i \in L$). Figure 3 shows the mean sample accuracy quickly converges to the true expected lattice accuracy as M increases.

3.2 Min. of Expected and Simple Accuracy

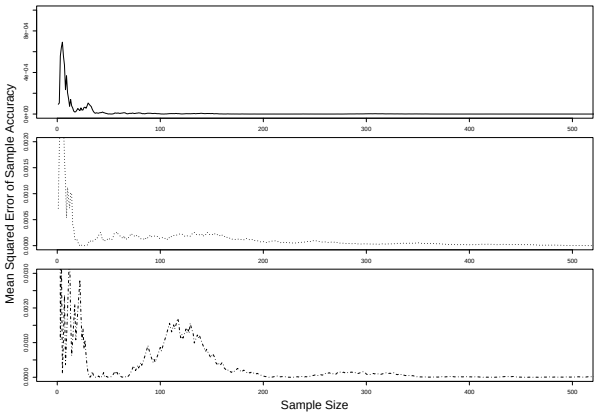

Figure 3: Mean squared error in mean sample accuracy *vs.* sample size for the three lattices from Figure 4. Lines are drawn in the same way in both plots. We see the misbehaving lattice from Figure 4 takes the most samples for convergence.

Figure 4 shows the distribution of accuracies for three example lattices, as well as their mean (expected) accuracy and the accuracy from their most-probable path. Not surprisingly, for lattices which exhibit small variance in accuracy, the one-best and expected accuracies are close together. We might expect that, as the variance (and thus the risk of traversing low-accuracy paths) increases, the expected accuracy would tend to be increasingly conservative in estimating the utility of the lattice. For example, the multimodal (misbehaving) distribution seen in Figure 4 has a much lower mean than one-best accuracy. Another way of saying this is that we expect that the accuracy on the "best" path ought to be better than the average accuracy for a randomly sampled path.

But this is not always the case. Consider Figure 5, which plots the expected accuracy for each lattice against its one-best accuracy, conditioned on the standard deviation in accuracy. For the riskiest lattices, we see that expected accuracy is often greater than the one-best accuracy. Since it seems reasonable to suppose that a lattice having higher expected than simple accuracy is poorly behaved, we take as our second measure the minimum of both measures,

$$\text{Acc}_{\min} = min(\text{Acc}_{\text{exp.}}, \text{Acc}_{\text{one-best}}).$$

4. EVALUATION

For our lattice retrieval system, we use the phonetic-lattice indexing approach described in [4]. To evaluate whether our new measures are better predictors of retrieval performance, we consider again the utterance retrieval task.

Suppose the recognizer does a good job producing the phone lattice. In this case, we'd expect utterances to be ranked highly by the retrieval system for terms they contain. Likewise, we'd expect utterances to have a lower rank when they are poorly recognized. In other words, if a recognition measure is doing a good job at predicting how easy it is to find the lattice, then the measure ought to be strongly correlated

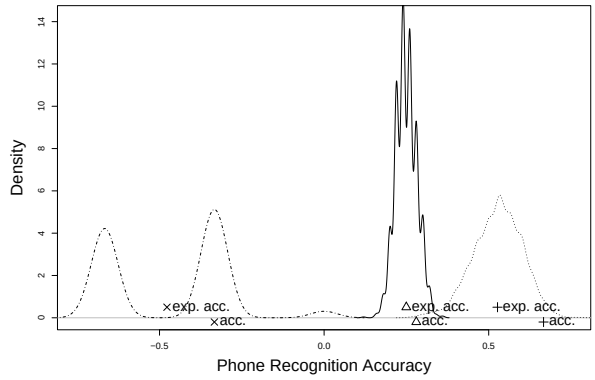

Figure 4: The distribution of accuracies for three lattices. Note, accuracy is quantized (not continuously distributed), so the density estimates are just to aid in visualization.

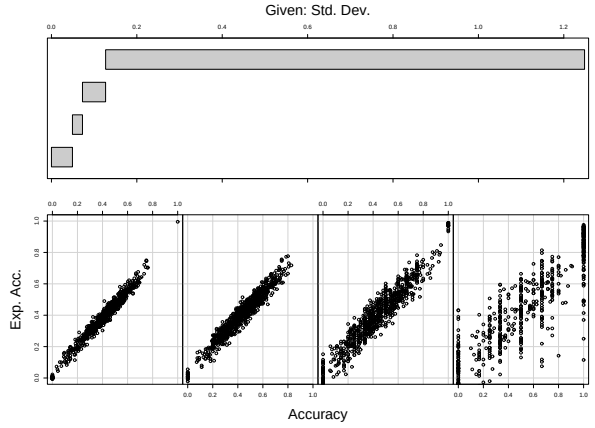

Figure 5: Expected accuracy *vs.* one-best accuracy. Each of the four subplots contains one fourth of all the lattices, where we partition by their standard deviation in accuracy.

with the rank order of the lattice in the search results.

Figure 6 shows expected accuracy and one-best accuracy plotted against the rank of the correct lattices for one query term, "about". Using Kendall's τ to test correlation for each measure and the lattice's rank order, we see in this case that the expected accuracy produces a stronger correlation ($\tau = -0.29$) than one-best accuracy ($\tau = -0.21$). Note that the correlation is negative because the rank position of a lattice gets bigger (worse) as the accuracy decreases.

To determine if this improvement is significant, we compare Kendall's τ on 868 query words having more than one correct lattice. Of these, 489 queries have the same value of τ (this is so because many queries have very few correct lattices). A remaining 210 have higher τ value using expected accuracy and 169 with one-best accuracy. Accordingly, we may apply the Fisher sign test, where our null hypothesis is that neither evaluation measure is superior at predicting the lattices rank order. We find that expected accuracy is

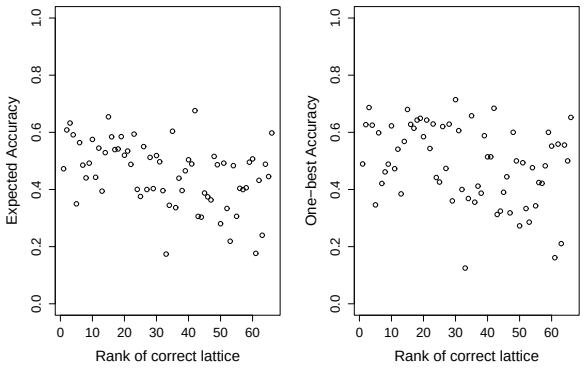

Figure 6: Expected accuracy and one-best accuracy plotted against the rank order of retrieved utterances for the query term "about".

significantly better, with p-value 0.0199. Along the same lines, we compare $\text{Acc}_{\min}$ and one-best accuracy, and find that $\text{Acc}_{\min}$ is significantly better, now with p-value 0.0017.

One reason why expected accuracy may outperform simple accuracy at this task is because simple accuracy tends to be heavily quantized for short utterances. This is essentially the same problem we outlined in Section 2's toy sentence example. By taking a weighted average over many traversals' accuracies, expected accuracy reduces this quantization effect, as we see in Figure 5: expected accuracy distinguishes between lattices that simple accuracy does not (note the many lattices spread along $\text{Acc} = 0$).

A second reason why these new measures may outperform simple accuracy relates to the distribution of path probabilities in the lattice. Suppose the most probable path is also the best (lowest WER) path with probability p_1. For computing one-best accuracy, we don't care whether the second best path has probability $p_1/2$ or $p_1/1000$. The result is that we only care about the relative (not absolute) scaling of language model vs. acoustic model scores. If the relative scaling of language to acoustic model scores is fixed, the rank ordering of the paths will remain fixed. If however, we intend to compute posteriors on sequences in the lattice—as we commonly do for indexing, then we care not only about the relative, but also the absolute value of the scale factors. This is so because the scaled model log-probabilities are exponentiated in the computation of posteriors. Consequently, both the relative and absolute values of the scale parameters effect the index, but only the relative value effects one-best accuracy. Expected accuracy, which weights paths by their posterior probability, is thus able to better predict the utility for indexing.

5. CONCLUSION AND FUTURE WORK

We've introduced two simple new evaluation measures for speech recognition lattices: expected accuracy and the minimum of expected and simple accuracy. We've shown that the traditional measure for recognition evaluation, WER, may be a worse predictor for downstream retrieval performance,

particularly amongst the more recent generation of lattice-based systems which consider multiple speech hypotheses in indexing and search. In a comparison with simple one-best accuracy, we experimentally validated that these new measures can be superior at predicting a system's ability to utilize (i.e., highly rank) the recognition output. We hope such measures may reduce the end to end development cost of these systems.

An important limitation of this work is our focus on ranked utterance retrieval as opposed to full SDR (i.e., real information needs, relevance assessments, and human formulated topics). Our goal here was to avoid confounding effects associated with long queries and to focus on the foundational problem of detecting a term's presence in the speech. It remains to be seen if these evaluation measures will retain their advantage in the case of full SDR.

We also would like to investigate how these measures suggest new designs for speech recognizers. While speech recognition systems have already been developed which explicitly attempt to minimize WER in expectation (e.g., minimum Bayes-risk decoders [2]), it seems plausible that loss functions more closely tied to the IR problem could provide additional gains, e.g., by incorporating a notion of term discriminativeness.

Acknowledgements

The author is grateful to Douglas W. Oard for his many helpful suggestions. This work was funded in part by NSF IIS award 0122466 (MALACH).

6. REFERENCES

[1] J. Garofolo, G. Auzanne, and E. Voorhees. The TREC Spoken Document Retrieval Track: A Success Story. In *RIAO '00*.

[2] V. Goel and W. Byrne. *Pattern Recognition in Speech and Language Processing*, chapter Minimum Bayes-risk automatic speech recognition. CRC Press, 2003.

[3] A. C. Morris, V. Maier, and P. Green. From WER and RIL to MER and WIL: improved evaluation measures for connected speech recognition. In *INTERSPEECH'04*.

[4] J. S. Olsson, J. Wintrode, and M. Lee. Fast Unconstrained Audio Search in Numerous Human Languages. In *ICASSP'07*.

[5] A. Stolcke. SRILM – an extensible language modeling toolkit. In *ICSLP'02*.

[6] S. Young, G. Evermann, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and Woodland. The HTK Book (Version 3.2.1). 2004.

Evaluating ASR Output for Information Retrieval

Laurens van der Werff
University of Twente
PO Box 217, NL-7500AE
Enschede, The Netherlands
laurensw@ewi.utwente.nl

Willemijn Heeren
University of Twente
PO Box 217, NL-7500AE
Enschede, The Netherlands
w.f.l.heeren@ewi.utwente.nl

ABSTRACT

Within the context of international benchmarks and collection specific projects, much work on spoken document retrieval has been done in recent years. In 2000 the issue of automatic speech recognition for spoken document retrieval was declared 'solved' for the broadcast news domain. Many collections however, are not in this domain and automatic speech recognition for these collections may contain specific new challenges. This requires a method to evaluate automatic speech recognition optimization schemes for these application areas. Traditional measures such as word error rate and story word error rate are not ideal for this. In this paper, three new evaluation metrics are proposed. Their behaviour is investigated on a cultural heritage collection and performance is compared to traditional measurements on TREC broadcast news data.

General Terms

Automatic Speech Recognition, Spoken Document Retrieval, Lattices, Evaluation

1. INTRODUCTION

Several developments in recent years have led to an increased interest in improving access to spoken word collections. The reduced cost and increased capacity of random access media (e.g., harddrives), combined with the increased speed of Internet connections, means that it is now quite feasible to access such collections online. In contrast to these technological opportunities stands the reality of current practice: many existing collections have not been properly digitized yet since this requires a lot of manual effort. Those that have been digitized are often not searchable for a variety of reasons, ranging from intellectual property issues to technical and implementation issues.

Searching in spoken content implies the application of information retrieval (IR) techniques to speech. Since searching in speech directly is unfeasible, a more computer-processable representation has to be used. From an (automatic) indexing perspective, spoken word collections can be approached in several ways based on the amount of available collateral data. Collections that are up to a few hundred hours in size can usually be made accessible through some human effort: either by labelling segments of speech with keywords and named entities or by manually creating a full transcription. This can then be automatically aligned to the audio using standard Viterbi techniques [20] and indexed as any other textual document. When an audio collection is too large to be disclosed manually, it must be done using a more or less automated process. In such cases it is expected that an automatic speech recognition (ASR) system can be used to provide a full, though imperfect, transcription of the audio.

Of great importance for the accessibility of a spoken document collection is the quality of the index. The quality of an index based on ASR output will be highly dependent on the characteristics of the speech. Since ASR is probabilistic and based on models that are estimated from statistics, performance of ASR is determined largely by the match between those models and the speech that is processed. Spontaneity and noise typically cause problems for ASR systems due to the fact that they make the speech signal less predictable and so by definition reduce the match. ASR therefore tends to perform best on material that is generated under highly controlled circumstances, for example broadcast news (BN) or dictation. Many of the collections that are considered interesting are not of this type, such as historical audio or oral history collections. These may contain noisy spontaneous speech or highly accented speech by non-professional speakers, often recorded under suboptimal conditions using old-fashioned equipment. These circumstances typically cause a doubling of the number of ASR errors and thus reduce the reliability of the automatically generated transcription.

Many optimization methods for ASR on noisy and/or spontaneous speech have been extensively studied in the past [5]. Most of these studies have employed ASR as a 'dictation machine', meaning that the primary task of the system was to generate a literal transcription of every word that was uttered. Traditionally, the performance of such ASR systems is measured using the word error rate (WER). In the context of spoken document retrieval (SDR), ASR is not so much a dictation machine as it is a means to generate some representation that is suitable for building an index. The literal transcription is just a (potential) by-product of this process. WER is a flawed optimization criterion for ASR

in this context because (i) it is only defined as such on a (literal) transcription and can therefore not be calculated on ASR output such as n-best lists or lattices, and (ii) IR performance depends not only on the *amount* of errors but also on the *type* of errors.

Performance of IR systems is typically measured using the mean average precision (MAP), a score that is calculated based on the amount of relevant documents found for some set of queries, the amount of non-relevant documents that is produced and their ranking. Calculating such a score can only be done using an evaluation platform that contains ground-truth (i.e. human) relevance judgments for a set of queries and documents. When applying ASR to a collection for which such a platform exists, the MAP should be used as an optimization criterion.

In practice, IR evaluation platforms are only readily available for a limited amount of collections. When optimizing the ASR component of an SDR system for a collection for which no matching evaluation platform can be found, developing a new evaluation set requires a prohibitive amount of work. Instead, some ASR for IR optimization criterion is needed that can be used to predict the MAP, or at least the relative improvement in MAP, for collections where this score cannot be calculated. In this work, three new performance measures for ASR in an SDR context are introduced.

This paper is organised as follows: Section 2 touches on some previous efforts to find the relationship between ASR and IR performance, strengthening the argument that WER is not a good criterion for optimizing ASR in an IR environment. Section 3 first explains the workings of an SDR system and why current evaluation metrics for ASR can be problematic in this context. Then three new performance measures will be proposed that are more appropriate versions of the traditional measures WER, Story WER (SWER) and Out-Of-Vocabulary (OOV) rate. It is argued that the ASR output can only be assessed properly when some particular characteristics of the IR system are incorporated into the evaluation. The behaviour of the measures in combination with standard IR techniques is investigated in Section 4 and in Section 5 a comparison is made between the traditional ASR measures and the new ones on a TREC BN collection. Finally, Section 6 contains some conclusions and gives suggestions for future work.

2. RELATED WORK

The performance of ASR in the context of IR has been studied for many years, mainly in the context of TREC since 1997 [21]. In [4] it was noted that there is a high correlation between WER (actually SWER) and retrieval performance as measured with MAP. This correlation was even higher when instead of SWER a Named Entity SWER (NE-SWER) was used, measuring exclusively the named entity performance of the ASR system. In [9] some experiments were done with Term Error Rate (TER) as a performance measure. A high correlation was found between the TER and the MAP score of the systems, however no such clear relationship was found with the R-precision score. Since SDR performance on ASR based transcriptions was only marginally worse than on human transcriptions, ASR-based indexing was considered 'solved' for the BN domain [3].

In [19] the IR performance of indexes based on different types of ASR output was evaluated. The incorporation of the sentence structure through the use of n-best lists was found to be superior to using individual word probabilities from lattices. Using 1-best output was found to be inferior to using either n-best or lattice representations. The IR weights were calculated by combining relevance and ASR confidence into a single probabilistic measure. The effect of the choice of ASR output type on overall retrieval performance was measured by running and evaluating a predefined set of queries on the resulting index and comparing MAP score.

More recently, research has been done on optimized indexing from ASR lattices for improved IR performance, for example through multi-word queries [1] or through combination of multiple lattice hypotheses [14]. Both techniques gave rise to some improvement.

This previous work suggests (i) that IR performance is dependent on ASR performance, (ii) that indexing from lattices or n-best lists can improve IR performance and (iii) that the way that these enriched outputs are exploited needs to be optimized.

3. EVALUATING SDR

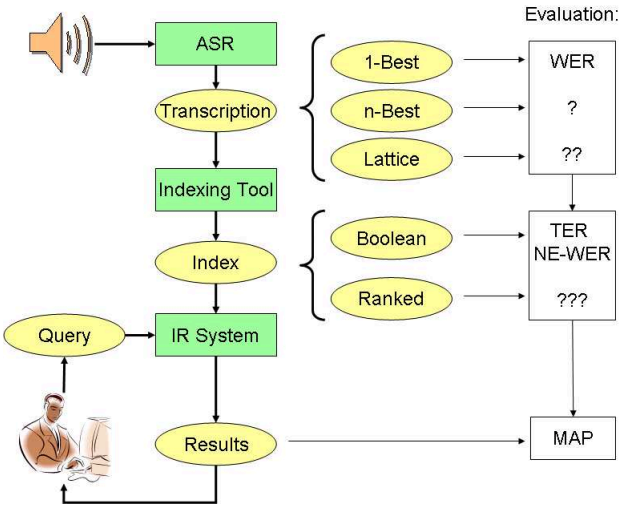

Figure 1: Anatomy of an SDR system

3.1 Anatomy of an SDR system.

A typical SDR system will contain at least the following three main components: an ASR engine, an indexing tool and an IR system. Figure 1 gives an overview of such an SDR system. The ASR engine takes as its input the audio containing the speech and produces a transcription. This can then be processed into some index representation by the indexing tool. Finally, the user enters queries into the IR system which will produce the relevant audio fragments based on the index.

These three components should work together in such a way that the final retrieval results are optimal given the user's query. MAP is the standard method for evaluating this, so optimizing the individual components for the best MAP

score is the most efficient way of improving system performance. Calculating the MAP score however is not always possible, it requires some evaluation platform which, as was mentioned in Section 1, is very time consuming to produce.

Although the inability to calculate a MAP score for most collections might limit optimization and customization of an IR component, it may still be possible to optimize the ASR system and the indexing tool. As mentioned in Section 2, there is a fairly strong correlation between ASR performance as measured with WER and the MAP score. However, as will become clear in Section 5, optimisation of the ASR component for minimal WER does not automatically lead to an SDR system with a higher MAP. Proper evaluation of the ASR output and/or the results of the indexing tool is therefore crucial for optimization of SDR systems for collections that do not allow for the calculation of MAP.

3.1.1 Transcription Types

The output of the ASR engine can take several forms. Traditionally, in dictation type applications, the 1-best output is used. It represents the sequence of words that, based on the acoustic and language models used as well as pruning parameters, gives the highest likelihood for a fragment of speech. A 1-best output normally does not contain any scoring information for individual words, meaning that confidence in its correctness is equal for each word in the transcription. Evaluation of the 1-best output is done using WER, calculated using the following equation:

$$WER = \frac{S + I + D}{N}$$

Where S, I and D represent the number of substitutions, insertions and deletions as determined through a dynamic programming, minimum Levenshtein distance function (weights: 4, 3 and 3)[13] alignment of reference and hypothesis transcription. N is the total number of words in the reference.

Alternatively, an ASR engine can produce an n-best list or a lattice structure as its output. Both of these types of output contain multiple transcriptions for the audio and may also contain some form of confidence scoring. The main difference between them is that n-best lists contain only full transcription alternatives, i.e. full sentences, while lattices contain alternatives on a word-by-word basis. A lattice structure is a relatively compact representation of the search space of the ASR engine and can be expanded into an n-best list. Lattice output is typically used as an intermediate representation that is then postprocessed/rescored into a 1-best output which in turn can be evaluated using WER. When lattice or n-best output has to be evaluated directly, no useful metrics are available.

3.1.2 Indexing

The index of an IR system links words and/or concepts to specific documents (or speech fragments in the case of SDR). In IR that is based on textual documents, the underlying data on which the index was made is, in principle, reliable. When the index is based on ASR output, the reliability of the index may suffer as a result of transcription errors. Since final retrieval performance is directly dependent on the index and only indirectly on ASR performance, evaluation of ASR output by measuring the impact of the errors on the index should, at least in theory, be more indicative of IR performance than evaluation of the ASR output by itself. Evaluation of an ASR-based index can be done by building an index both on a reference transcription and on the ASR output and comparing the two.

For a Boolean retrieval system, each index term represents an unambiguous set of documents: those that contain it. Measuring the impact of ASR errors on the index is therefore a matter of counting these errors, for example using the term error rate (TER) as proposed in [8]:

$$TER = \frac{\sum_w |A(w) - B(w)|}{W}$$

Where W is the total number of words in the reference and $A(w)$ and $B(w)$ represent the number of times word w occurs in the reference A and the transcription B, thereby modeling a traditional substitution as two errors. Since the number of occurrences of a word is of no importance in a Boolean system – a document is either a member of a set or it is not – a unique term error rate (UTER) value may be more appropriate. This can be calculated by using $A(w)$ and $B(w)$ only to indicate the presence (value=1) or absence (value=0) of word w in the document.

The family of ranked retrieval models is characterized by the inclusion of a – usually statistically motivated – weighting scheme on the index terms. Such a scheme is typically based on some form of term frequency (tf) and document frequency (df) combination. Several approaches exist for exploiting and calculating these measures, for example the Vector Space Model (VSM) [17] and Okapi [11].

Measuring the impact of ASR errors in a ranked retrieval environment is not simply a matter of counting, since errors now impact weights in a complex manner. A deletion of a term will decrease its tf for that document, but will also decrease the df that is calculated over the whole set, thereby increasing the weight for this term in all other documents.

The TER can be adapted as proposed in [7], so that the error count of each term is multiplied by an individual weight. This can be used to simulate the non-uniform impact of ASR errors, but finding a suitable weighting function may be quite difficult and the total error is still determined by simply counting the number of insertions and deletions.

3.2 SDR Evaluation Metrics

In the systems that were enrolled in the TREC benchmarks, ASR performance was measured using the WER (all systems used 1-best ASR output only)[4]. By comparing the ranked retrieval IR performance of the systems on each of the various ASR outputs, a correlation between ASR performance and retrieval performance could be established. As it turned out, the correlation coefficient in the TREC-7 systems between WER and MAP was 0.87, meaning a significant correlation. The NE-SWER showed an even higher correlation with the MAP at 0.91. Although this might validate the conclusion that WER is a good measure for predicting relative IR performance, there is more to this.

The ASR components of all systems that took part in this evaluation were optimized for the same evaluation metric:

WER, a measure that does not differentiate between errors on content words or on stopwords. Since all ASR systems had the same basic layout, it could be argued that the performance of these systems will not differ very much in a qualitative way, so pure quantitative analysis could be sufficient. Comparing WER with NE-SWER, the *relative* performance of all of the systems stayed the same, except for one. This was precisely the system that had shown lower relative MAP scores than would have been predicted from its WER, but it showed an NE-SWER that was in line with its MAP score. This increased the overall correlation coefficient and supports the notion that a qualitative measure may be useful for ASR evaluation in an IR context.

Quantitative analysis of ASR performance is only indicative of retrieval performance if this was also the criterion used for optimizing the ASR system, as is the case in most dictation type applications. When optimizing an ASR system for a different application, leading for example to an increased performance on named entities at the cost of performance on stopwords, the WER may no longer be a good indication of relative retrieval performance. When an index is built using n-best or lattice output, the WER cannot even be calculated as such. This is further reason to conclude that different ASR performance metrics are required for SDR. The following paragraphs will introduce three such measures.

3.2.1 Boolean Index Accuracy

In a Boolean retrieval system, the index *is* the system, since queries are simply a way of selecting documents from a combination of sets that are entirely defined by the index. In the context of such a system, measuring the quality of the index is a matter of calculating the TER and is therefore quite straightforward.

When an index is created based on n-best or lattice ASR output, the number of terms that are associated with a document becomes quite variable. When only words for which confidence in the ASR correctness is very high are included, this leads to a relatively small number of associated terms, while inclusion of several alternatives for some sentence-positions will increase the amount of terms.

In practice, due to the possibility of creating relatively complex indexes from lattices or n-best lists, the TER (or UTER) value may become much larger than 1 (or 100%), making it difficult to interpret unambiguously. For example, is an 'empty' index with a TER of 1 better than a relatively large index with a TER value of 1.5? It would be preferable to always indicate the performance with a number between 0 and 1, where 0 would mean no match between hypothesis and transcription, while 1 would indicate that the hypothesized index is equal to the reference index. The Boolean Index Accuracy (BIA) is such a measure:

$$BIA = \left(1 - \frac{D}{N_{ref}}\right) * \left(1 - \frac{I}{N_{index}}\right) \quad (1)$$

Where D is the number deletions, meaning terms that are in the reference, but not in the hypothesis, while I is the number insertions, meaning terms that are in the hypothesis but not in the reference. N_{ref} is the number of terms in the reference, while N_{index} contains the number of terms in the index. Terms are considered unique for a particular story (or retrieval unit). Equation 1 is made up of two parts: the first bracketed part indicates the coverage of the index, i.e. the fraction of the words in the reference transcription that can be found in the index. The second bracketed part indicates its correctness, i.e. the fraction of the words in the index that is also found in the reference transcription.

3.2.2 Ranked Index Accuracy

In a system of ranked retrieval, the index contains weights for each indexable term in each document. These weights determine the ranking and therefore define the system. Measuring the similarity in weights between the hypothesized index and the reference index can be done using the standard VSM [17]. In this model, the index can be represented as a vector, with the indexed terms as vector dimensions and the weighting scores as vector lengths. By calculating the vector inner product of the normalized vectors, the similarity of two indexes can be determined. This property can be expressed in the RIA measure that is calculated as follows:

$$RIA = \frac{\sum_{k=1}^{m} d_k \cdot q_k}{\sqrt{\sum_{k=1}^{m} (d_k)^2} \cdot \sqrt{\sum_{k=1}^{m} (q_k)^2}} \quad (2)$$

Where m is the combined number of terms in the indexes and d_k and q_k represent the weight of term k in the reference index d and the hypothesis index q. The Ranked Index Accuracy (RIA) represents the similarity between two indexes on a scale of 0 to 1, with 1 meaning that the indexes are identical.

3.2.3 ROOV

In optimizing ASR, the lexicon and language model are of vital importance. It is therefore useful to measure OOV rate, i.e. the percentage of words in the audio that is not included in the lexicon of the speech recognition system. In principle, the number of OOV terms is independent of the ASR output and also independent of the type of index that is made. However, since in SDR the ASR system is no longer a dictation machine and not all terms are treated equally, this measure should be adapted somewhat.
Traditionally, the OOV rate is calculated by dividing the number of OOV terms by the total number of terms in the reference.

$$OOV = \frac{\#OOV\ terms}{\#terms} * 100\%$$

Within a Boolean retrieval environment, one only needs to divide the number of unique OOV occurrences by the number of unique indexable terms to calculate the unique OOV (UOOV):

$$UOOV = \frac{\#unique\ OOV\ terms}{\#unique\ indexable\ terms} * 100\%$$

Within a ranked retrieval environment, the OOV can be calculated by dividing the total mass of all weights of the OOV terms by the sum of all the weights of the (reference) index, resulting in the retrieval OOV (ROOV):

$$ROOV = \frac{\sum Weight_{OOV\ terms}}{\sum Weight_{index}} * 100\% \quad (3)$$

When optimizing the ASR lexicon for minimal OOV, the best strategy is to include only the most frequent words in the lexicon, either estimated on a subset of the collection

or on an external text corpus. When the lexicon is being optimized for ROOV, a different strategy must be chosen: for example including those words that have the highest expected weights.

3.2.4 Example of evaluation measures

When lattice or n-best ASR output is used for generating an index, the size or complexity of the index is variable: it is possible to include more or less terms from the lattice or n-best list in the generation of the index[2]. Inclusion can be done on the basis of many criteria, and weights may be adjusted accordingly. Figure 2 shows the values of the various performance measures for indexes where a variable number of terms from a lattice ASR output are included. The inclusion criterion in this case was the posterior probability.

The collection used for generating this graph contains radio recordings with (Dutch) noisy spontaneous speech, hence the relatively poor absolute performance when compared to typical results on BN type data (as found in Table 2). The total duration of the audio was approximately 220 minutes, divided into 34 stories containing an average of 1154 words per story. ASR was performed in a single pass, using BN optimized acoustic and language models and a lexicon of 65k words. This led to a WER of around 55%. For the RIA results, a $tf * log(idf)$ score was used for calculating weights. The BIA and RIA scores were calculated with an index based on a human-made transcription of the audio as the reference.

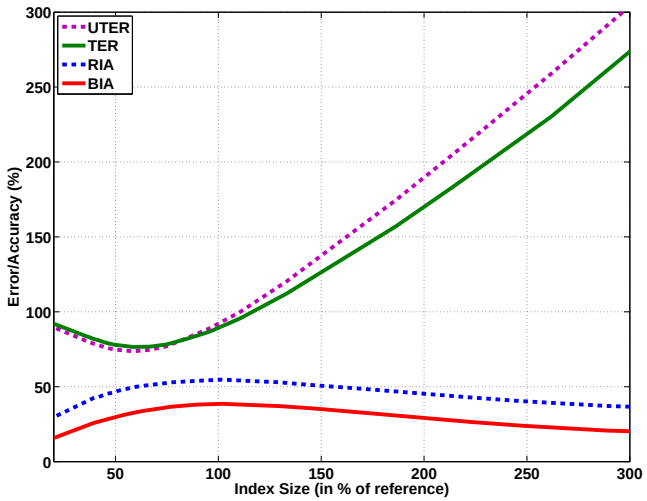

Figure 2: Comparison of performance measures for various index sizes. The index size is shown relative to the size of the reference index.

When an index is created for a certain collection, it makes sense to select the size that results in the highest similarity to the reference index. The performance of the system can thus be characterized in a quantitative manner by the maximum value of the BIA curve and in a qualitative manner by the maximum value of the RIA curve, in this case 0.39 and 0.55 respectively.

4. IR STRATEGIES AND THE ASR BASED INDEX

The evaluation measures introduced in Section 3 can be used to generate a performance number for both quantitative(BIA) as well as qualitative(RIA) evaluation. In order to show how these numbers are affected by more or less standard IR techniques, some experiments were performed with stopword filtering and stemming. The same data as in Section 3.2.4 was used.

4.1 Stopwords

In IR applications it is standard practice to filter stopwords from the index. These are words that are very common, have little or no meaning by themselves and will therefore not help in identifying relevant documents. Stopping of the most frequent words leads to a reduction in index size of up to 50% without impacting retrieval performance [18]. There is no real consensus as to what is the best size for the stopword list, but for Dutch, lists in the range of 50 to 1500 words can be found.

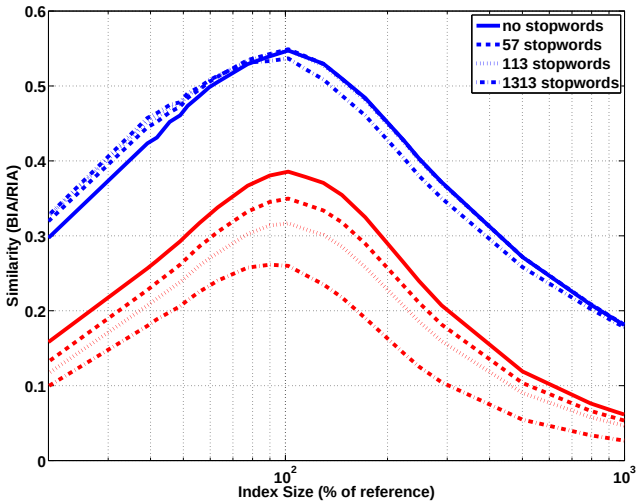

Figure 3: Index similarity after application of various stopword lists. Blue (upper) curves show RIA scores, red (lower) curves show BIA scores.

Figure 3 shows the performance curves of an ASR lattice-based index on a collection where stoplists of various sizes have been applied. The horizontal axis (index size) has been plotted on a logarithmic scale for clarity reasons. The graph clearly shows how the BIA value is impacted by using a stopword list, indicating that in this collection, ASR performance on stopwords is different (better) than on content words. The RIA measure is relatively stable, confirming that the stopwords have low relevance and ASR errors on these words may therefore have a limited impact on retrieval performance. Optimal index size seems to be at around 100% of the size of the reference index. This was to be expected, since the criterion for inclusion of terms in the index – the posterior probability – was the same criterion that was used by the ASR engine for selecting the 1-best path. The ASR engine was setup for minimal WER, which generally occurs

Stop-words	Indexable terms	U-Indexable terms	max. BIA	max. RIA	OOV (%)	UOOV (%)	ROOV (%)
0	39237	13442	0.39	0.55	4.0	7.3	11.3
57	23190 (-41%)	11980 (-11%)	0.35	0.55	5.9	7.9	11.5
113	17969 (-54%)	10811 (-20%)	0.32	0.55	7.6	8.7	11.8
1313	10488 (-73%)	6819 (-49%)	0.26	0.54	13.0	13.8	15.5

Table 1: Stopword statistics and index quality.

when the 1-best transcription is roughly the same length as the reference transcription.

Table 1 shows some statistics for this collection before and after applying the stopword lists. Although the total number of indexable terms in the transcriptions can easily be reduced by more than 50%, the number of unique indexable terms reduces much more slowly, so the reduction in index size will be less dramatic. The table also shows the various OOV measures as described in Section 3.2.3. The traditional OOV value of 4%, though not low, seems acceptable. However, when a stopword list is applied, it becomes clear that OOV rate of potential query terms in this particular SDR system is relatively high. ROOV seems to be the most robust measure, indicating more or less how much 'information' from the audio cannot be retrieved due to OOV issues. More on OOV rates and specific issues for Dutch can be found in [15].

4.2 Stemming

Both [6] and [12] found that using a Porter Stemmer [16] for Dutch did not significantly improve IR performance, but did not reduce performance either. [12] showed that a performance increase could be obtained by using more advanced algorithms, including compound splitting. It is not the aim of these experiments to build an optimal stemmer/splitter for Dutch, but merely to investigate the impact of such techniques on the quality of an index derived from an ASR run. The impact on the quality of the index as measured with BIA and RIA, using an implementation of the Porter Stemmer for Dutch is evaluated here.

Figure 4 shows the performance curves for an index based on the same ASR lattices, with and without stemming applied. Although previous studies indicated that the Porter stemmer may not improve IR performance for Dutch textual documents, these results show an increased similarity between the ASR based index and the reference index. Applying the stemmer increased RIA by 3.3% and BIA by 9.7% relative. It would therefore be interesting to further investigate whether the Porter stemmer can be beneficial for Dutch SDR, even though it is not for traditional Dutch IR.

5. COMPARISON TO OTHER MEASURES

To investigate whether the RIA and BIA measures are indeed useful for predicting retrieval performance in an SDR system, a complete IR evaluation platform must be used. Evaluations should be done based on several distinct ASR outputs. For the WER measure, this has been done for the TREC9 SDR track [3] by both Cambridge University (CU) and the University of Sheffield for seven different ASR runs. The results can be found in Table 2. Their retrieval results as well as the ASR outputs are publicly available from NIST

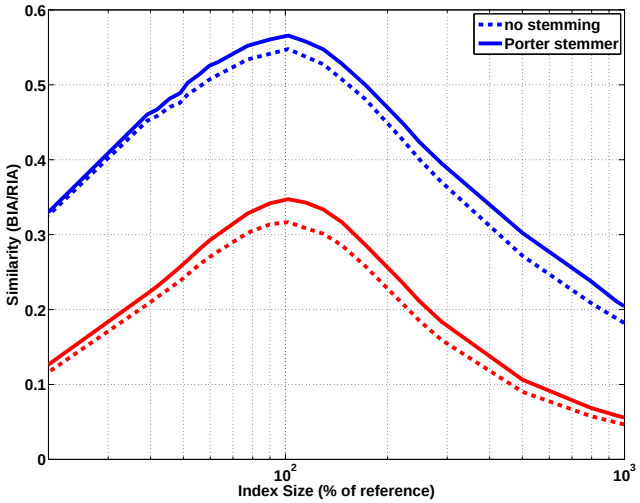

Figure 4: Performance with and without stemming applied. Blue (upper) curves show RIA scores, red (lower) curves show BIA scores.

and could therefore be used for comparisons with the values of RIA and BIA.

RIA and BIA scores were calculated after applying a stopword list and a Porter stemmer to the data, as was done by both CU and Sheffield systems. RIA scores were based on weights that were calculated using the method described in [10] with constants set to the values that were reported by those labs. Table 2 shows only the RIA values based on the CU weight calculation settings, the Sheffield RIA values (not shown) were very similar. Some complications arose, possibly leading to a suboptimal calculation:

- Reference transcriptions with >10% WER were used to calculate RIA, whereas WER was estimated on a 10h subset of checked references with 0% WER
- No lab-specific normalization scripts were available, only the supplied TRANFILT tool could be applied
- Postprocessing techniques could only be approximated from the system descriptions; actual stemmers and stopword lists were not available
- Indexes were generated assuming the Story Known condition, while recognizer results and MAP scores for cross-site evaluations were only available for the Story Unknown condition

transcription	WER	SWER	RIA	BIA	CU	Sheffield
human ref	10.3	11.0			0.4402	0.4180
cuhtks1p1u	27.6	25.1	0.695	0.500	0.4044	0.3576
cuhtks1u	22.0	19.6	0.732	0.549	0.4299	0.3727
limsi1u	22.8	19.7	0.726	0.540	0.4019	0.3862
limsi2u	22.3	18.8	0.736	0.546	0.4162	0.3968
nist2000b1u	27.3	23.6	0.699	0.505	0.4075	0.3837
shef1u	33.1	28.3	0.674	0.452	0.3958	0.3919
shef2u	30.4	25.6	0.693	0.478	0.3983	0.3931

Table 2: TREC cross system results; RIA scores are based on CU parameters, the final two columns show MAP scores.

	CU	Sheffield
WER	-0.760	0.133
SWER	-0.721	-0.043
RIA	0.759	0.036
BIA	0.769	-0.132

Table 3: Correlation coefficients for MAP vs. ASR performance metrics.

Sheffield performed worse than CU on all transcription sets, however, they seemed to perform relatively well on their own transcriptions as compared to those from other sites. The best transcription set as determined by WER (cuhtks1u) led to the second worst retrieval result for this group, while the best ASR set according to SWER (limsi2u) gave the best retrieval performance. In general, no correlation between any of the measures used here and the Sheffield scores was found, nor does there seem to be any obvious correlation between the Sheffield scores and the CU scores (see Table 2).

The CU system performance showed a significant correlation with ASR quality (see Table 3). Still, the differences in retrieval performance were quite small, indicating that much of the reductions in SWER are negated by retrieval techniques such as query expansion. When Story ACCuracy (SACC) is defined as 100-SWER, its relative improvement between the best and worst transcriptions is 13.2%, the improvement in RIA is 9.2% but final retrieval performance only improves by 5.2%. RIA therefore seems to be a better predictor of retrieval performance than SACC.

All the ASR error measures used here are highly correlated (not shown). BIA is highly correlated with WER, because the index size is always within 10% of the reference. Correlation of RIA and BIA with retrieval performance is similar to their traditional error measure counterparts. The items mentioned earlier prevented the calculation of more precise RIA values, something that should not be a problem if the actual indexing software were available, as would be the case when developing ones own SDR system.

Although the CU retrieval results showed a significant correlation with WER, the Sheffield results did not. A possible cause for this lack of correlation for the Sheffield system might be that their IR component was specifically optimized for use on their own output, for example through tuning of the query expansion to the ASR lexicon or through certain post-processing techniques.

To neutralise for the effects of a better match between IR and ASR through circumstances that could not be reproduced in our calculation of RIA, comparisons were made between two different ASR runs that were produced by the same site. Three sites submitted an alternative ASR run: Cambridge, Sheffield and Limsi. When comparing retrieval performance on two ASR runs that were generated within the same site, the 'best' transcription scored consistently higher in both IR systems. Table 4 shows the performance difference in the CU and Sheffield systems between two transcriptions from the same lab.

The CUHTKS1P1U transcription from CU had an accurracy that is 7.2% lower than their CUHTKS1U version. The RIA value was 5% lower while the MAP reduced by 5.9%. The difference in RIA value for the Sheffield system in this case was 5.1%, slightly higher than for the CU system, and the MAP reduced by 4.1%. A similar trend can be found when comparing ΔACC and ΔRIA in Table 4 for the other lab's transcriptions. ΔRIA turned out to always be a better predictor of ΔMAP score than ΔACC. When the ΔRIA is compared to $\Delta SACC$ the difference was smaller, but overall still favored RIA as a predictor for MAP.

Finally, RIA10h was calculated on a ten hour subset of the reference transcription that was manually corrected (the same subset that was used to calculate the ACC numbers). It proved to be a slightly better predictor than RIA in this comparison for MAP of the CU system, but slightly worse for the Sheffield system. This indicates that RIA can also be used if a reference transcription is available for only a relatively small part of the collection.

If more details had been available of the systems that were used in this comparison, a better estimation of the RIA score could have been made, possibly leading to a higher correlation between RIA and MAP scores.

6. CONCLUSION & FUTURE WORK

In this paper, the issue of how to evaluate ASR output for use in SDR systems was raised. To avoid the use of a prohibitively expensive full IR evaluation platform, the suggestion was made to evaluate just the ASR-derived index by comparing it against an index made on a reference transcription. Three evaluation measures were introduced: (i) BIA for evaluating the errors in a purely quantitave manner, (ii) RIA for a weighted evaluation and (iii) ROOV for a weighted measure of OOV rate.

			CU			Sheffield		
Site	ΔACC	$\Delta SACC$	ΔRIA	$\Delta RIA10h$	ΔMAP	ΔRIA	$\Delta RIA10h$	ΔMAP
Cambridge	7.2	6.8	5.0	**5.5**	5.9	**5.1**	5.6	4.1
Limsi	0.6	1.1	1.4	**1.6**	3.4	**1.5**	1.0	2.7
Sheffield	3.9	3.6	**2.7**	**2.7**	0.6	**2.7**	**2.7**	0.3

Table 4: Predicted and actual performance difference of CU/Sheffield system between two ASR transcriptions from the same site; RIA is calculated on the reference transcription, while RIA10h is calculated on the manually checked 10h subset. All values are percentages.

These measures were applied to a test set with noisy spontaneous Dutch speech. Results were encouraging and in line with expectations both for performance with stopword lists as well as for performance with stemming. When a comparison was made between RIA and the more traditional WER on a set of BN data from the TREC SDR benchmarks, the new measure was significantly better at predicting changes in retrieval performance, despite the fact that its calculation was hampered by a lack of details about the IR system used.

RIA, BIA and ROOV scores can be calculated on a subset of an audio collection as is usually also done for WER estimation. The most important limitations are that the test collection must be large enough for accurate weight estimation and that the audio included in the test collection is representative for the ASR performance of the full set.

As future work, to better estimate the correlation between ΔRIA and ΔMAP, a comparison should be made in the context of an SDR system that includes a full evaluation platform. The various measures can then be compared at more stages of ASR optimization than was the case with the BN data from TREC. It would also be interesting to see if it is possible to somehow include the effects of query expansion into the measure.

7. ACKNOWLEDGMENTS

The research reported on here was funded by the research project CHoral[1], part of the NWO-CATCH[2] program.

8. REFERENCES

[1] C. Chelba and A. Acero. Position specific posterior lattices for indexing speech. In *Proceedings of ACL*. Microsoft Research, 2005.

[2] C. Chelba, J. Silva, and A. Acero. Soft indexing of speech content for search in spoken documents. *Computer Speech and Language*, 21(3):458–478, 2007.

[3] J. Garofolo, J. Lard, and E. Voorhees. Spoken document retrieval track slides. 2000.

[4] J. S. Garofolo, C. G. P. Auzanneic, and E. M. Voorhees. The trec spoken document retrieval task: A success story. In *Proceedings of RIAO*, 2000.

[5] Y. Gong. Speech recognition in noisy environments: a survey. *Speech Communication*, 16(3):261–291, 1995.

[6] V. Hollink, J. Kamps, C. Monz, and M. de Rijke. Monolingual document retrieval for european languages. *Information Retrieval*, (7):33–52, 2004.

[7] S. Johnson, P. Jourlin, K. S. Jones, and P. Woodland. Spoken document retrieval for trec-8 at cambridge university. In *NIST Special Publication 500-246*, pages 197–206, 2000.

[8] S. Johnson, P. Jourlin, G. Moore, K. S. Jones, and P. Woodland. The cambridge university spoken document retrieval system. In *Proceedings of ICASSP '99*, pages 49–52, 1999.

[9] S. Johnson, P. Jourlin, G. Moore, K. S. Jones, and P. Woodland. Spoken document retrieval for trec-7 at cambridge university. In *Proceedings of TREC-7*, pages 191–200, 1999.

[10] S. E. Johnson, P. Jourlin, K. S. Jones, and P. C. Woodland. Spoken document retrieval for trec-9 at cambridge university. *NIST Special Publication 500-249*, pages 117–126, 2000.

[11] K. S. Jones, S. Walker, and S. E. Robertson. A probabilistic model of information retrieval: development and comparative experiments. parts 1 and 2. *Information Processing and Management*, 36(6):779–840, 2000.

[12] W. Kraaij and R. Pohlman. Viewing stemming as recall enhancement. In *Proceedings of ACM SIGIR*, pages 40–48, 1996.

[13] V. Levenshtein. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics - Doklady 10, vol. 10*, pages 707–710, 1966.

[14] X. Li, R. Singh, and R. Stern. Lattice combination for improved speech recognition. In *Proceedings of ICSLP*. CMU, 2002.

[15] R. Ordelman, A. van Hessen, and F. de Jong. Compound decomposition in dutch large vocabulary speech recognition. In *Eurospeech 2003*, 2003.

[16] M. Porter. An algorithm for suffix stripping. *Program 14(3)*, pages 130–137, 1980.

[17] G. Salton and M. J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, 1983.

[18] P. Schauble. *Multimedia Information Retrieval: Content-Based Information Retrieval from Large Text and Audio Databases*. Kluwer Academic Publishers, 1997.

[19] M. Siegler. Integration of continuous speech recognition and information retrieval for mutually optimal performance. In *PhD Thesis*. CMU, 1999.

[20] L. van der Werff, W. Heeren, R. Ordelman, and F. de Jong. Radio oranje: Enhanced access to a historical spoken word collection. In *Proceedings of CLIN17*, 2007.

[21] E. M. Voorhees. Overview of the sixth text retrieval conference (trec-6). Department of Commerce, National Institute of Standards and Technology, 1997.

[1] http://hmi.ewi.utwente.nl/choral

[2] http://www.nwo.nl/catch

Supporting radio archive workflows with vocabulary independent spoken keyword search

Martha Larson
ISLA, University of Amsterdam
Kruislaan 403
1098 SJ Amsterdam,
The Netherlands
larson@science.uva.nl

Stefan Eickeler
Fraunhofer IAIS
Schloss Birlinghoven
53754 Sankt Augustin,
Germany
stefan.eickeler@
iais.fraunhofer.de

Joachim Köhler
Fraunhofer IAIS
Schloss Birlinghoven
53754 Sankt Augustin,
Germany
joachim.koehler@
iais.fraunhofer.de

ABSTRACT

Archive departments of large radio archives stand to benefit greatly from speech recognition technology and other audio processing techniques. One of the reasons why automatic digital audio processing has not yet realized its full potential is that it remains unclear how to integrate automatic techniques into existing archive workflows. In order to move towards a practical understanding of how automatic techniques can be used to support archive staff, two large German radio broadcasters, Deutsche Welle and Westdeutscher Rundfunk, commissioned Fraunhofer IAIS to build a German-language radio archive prototype. This paper discusses the development and assessment of the spoken keyword search module of this prototype. The difference between the radio archive prototype discussed in this paper and existing systems for spoken document retrieval is that the prototype was designed and tested in a project group consisting of both multimedia researchers and archive professionals. As a result, the prototype and its evaluation is tuned to the explicit needs of archivists working at large radio archives. First, the paper discusses the special needs of radio archive staff and how they were accommodated in the design of the keyword search capacity. In particular, the archive staff required a vocabulary-independent search facility. This facility was implemented by a fuzzy-matching algorithm that performs a similarity search on syllable transcripts generated by the speech recognizer. Then, the paper presents the results of an evaluation designed to assess whether or not the radio archive prototype fulfilled the needs of archivists.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: Content Analysis and Indexing; H.4 [**Information Systems Applications**]: Miscellaneous; H.5 [**Information Interfaces and Presentation**]: Multimedia Information Systems

General Terms

Algorithms, Design, Performance

Keywords

spoken document retrieval, spontaneous speech, audio archive, speech recognition, radio broadcaster

1. INTRODUCTION

In 2000, spoken document retrieval was declared a solved problem [6]. Yet we arrive in 2007 and audio search systems are still not in widespread use to either retrieve or archive materials in large audio archives, such as those maintained by radio broadcasters. This paper reports on work in the area of spoken document retrieval carried out within a project commissioned from Fraunhofer IAIS by two large German radio broadcasters, Westdeutscher Rundfunk and Deutsche Welle. The goal of the project was to produce a German-language prototype radio archive that would be used to investigate the practical aspects of integrating automatically generated metadata into archive systems. The project wished to investigate the source of the lag between the availability of audio search technology and its integration into archiving workflows and to clarify which performance factors or design factors might be responsible.

Radio archive departments are eager to explore the possibilities for content-based spoken document retrieval offered by speech recognition technologies because they are faced with the task of annotating more material than would be possible by conventional (exclusively human-based) methods alone. Previously, archive departments stored radio content in the short term primarily for legal reasons. For long term storage, the best segments were removed and painstakingly annotated and archived as historical record or as cultural documents. Currently, the cost of storage medium has fallen far enough that it is becoming feasible to simply store everything broadcast. In other words, it not necessary to choose content for long term storage. Archives are increasingly being called upon to supply journalists and editors with content for reuse. Today's radio programs are enriched by an increasing number of sound bites drawn from past programs. This development has been characterized in the literature as archives moving "out of the basement" [8] to play a key role in production. In short, the trend is towards archiving as much material as possible.

In order to set a clear focus on the practical aspects of integrating automatic digital audio processing technology into archive work flows, the radio archive prototype was designed and tested by a project group consisting of both multimedia researchers, who were experts in the field of audio processing and speech recognition, and archive professionals, who were experts in the field of archiving and retrieval of radio content. Although the project group built on familiarity with previous research in the area of spoken document retrieval and on knowledge of existing systems and prototypes such as [9, 14, 7], an effort was made to eschew preconceptions and to build a prototype explicitly tailored to the needs of the archive departments of the two broadcasters. The prototype would thus offer a clear demonstration of the concrete potential of automatic digital audio processing to support the existing workflows of archive staff. The project group realized that in two areas the prototype radio archive needed to go above and beyond current practices in broadcast news retrieval. First, only a subset of the content of the radio archive is broadcast news or planned speech. Spontaneous speech, such as occurs in interviews, makes up a large portion of the archive content. Second, the information needs of archivists are specialized. Research in information retrieval places a focus on "aboutness", in other words in determining the topic of a document. User queries are considered to be requests for documents dealing with a certain topic. Although staff at radio archives needs to search for documents on certain topics, their information needs tend to transcend "aboutness." Archives receive requests for quotations spoken by certain prominent figures, for excerpts in which politicians express particularly strong negative opinions on popular issues and for segments in which interviewees use particular buzz-words or phrases. The project group also realized that in order to integrate well into the existing workflow, the radio archive prototype must be designed building on established archiving practices. The prototype needed to allow archivists to continue to use tried-and-true archiving conventions and familiar search strategies that allow them to combine detailed world knowledge and knowledge of the archive with a search interface.

This paper recounts the development, implementation and test of the spoken keyword search module of the radio archive prototype designed by the project group and implemented at Fraunhofer IAIS. It begins by a discussion of the needs of the archive staff and how these were incorporated into the functionality and interface design of the radio archive prototype. Then it discusses the testing of the prototype in order to determine if it met the needs of the archivists. Finally, it concludes with comments about lessons learned.

2. NEEDS OF ARCHIVE STAFF

In order to understand the radio archive domain and the broadcast news domain, the project group made a thorough investigation of the workflow and needs of archive staff. Both the content of the archive and the needs of the users (i.e. of the archive staff) turned out to constitute important differences between the radio archive domain and the broadcast news domain.

2.1 Radio archive content

Radio content is archived for several reasons. Broadcasters are typically legally required to keep a record of what they broadcast for a specified period of time. Also, a broadcaster, especially a public broadcaster, may have a mandate to preserve culturally relevant modern recordings and to curate a collection of historical audio recordings. Finally, a broadcaster archives material as resources to be rebroadcast or to be used in future productions.

The broadcasters who commissioned the radio archive prototype maintain broadcast news in their archives. But they also store a wealth of other content including documentaries, interviews and talk shows. Archive professionals pointed out that the most pressing need for content-based retrieval of radio content was for those programs for which there was little or no formal metadata. A news show typically has a minimal description of each of the report segments that was used to produce the show and which, if all goes well, follows it from production to the archive. Interview talk shows, however, are largely unplanned. In fact, their appeal to the listening public lies exactly in the spontaneous and free form discussion between show host and guest. An interview talk show arrives at the desk of the archivist with long sections which are unsegmented and not described in any way. The greatest potential for content-based retrieval for archive staff is being able to access these sorts of segments. Often, there are insufficient resources available to allow human annotation of interview talk shows and they are stored in the archive without annotation and are effectively lost. In sum, although radio archives contain planned speech such as broadcast news, archive departments need speech retrieval systems in order to retrieve programs containing unplanned speech, since these have no production data.

2.2 Archiving workflow

The first step in the archive workflow is for the archive staff to screen radio material for selection. They select which recordings, or which subsections of recordings, will be archived and for how long. They also select the level of detail at which each radio recording will be annotated. Naturally, material selected for long-term archival will be earmarked for a high level of annotation detail. The next step is to annotate each recording that is selected for archiving. In this step, the archivist produces, by hand, a description of the recording that will make it possible to find the recording in the archive. This description takes the form of a summary or a list of keywords. Depending on the level of granularity required, this description can include a division of the recording into topical sections, each marked with a time stamp. Each topical selection is then annotated separately. Sometimes a list of program segments is available from the production metadata. If such a list is available, the archivist is able to use it as a skeleton on which to build up the description. A final responsibility of the archive staff is to maintain the archive, protecting it against inconsistencies.

The radio archive prototype was conceived to support the archiving workflow of archive staff. Automatically generated metadata such as segmentations and speech recognition transcripts are to be used to aid conventional archiving practices. Annotation becomes a collaboration between archivist and machine. In contrast to broadcast news systems that fully automate the generation of metadata, the radio archive prototype, aims to support the workflow of archive staff.

2.3 Retrieval from the archive

The archive staff is responsible for responding to requests for recordings from the archives. Emphasis in retrieval from the archive is on precision and not recall, since archivists' task is to find something that is suitable and not necessarily to find everything that is suitable. The project group surveyed the types of requests generally received and assessed the response potential of the content-based search functionality of the prototype radio archive. The types of requests received by archive departments can be grouped into different groups with respect to the support that content-based keyword search potentially provides for responding to them.

In the first group are the requests that can be handled by using only the formal metadata of a recording, namely information such as the title of the series, the title of the particular program and the date of first broadcast. This group includes requests for a particular program by title, title words, producer, or by broadcast date.

In the second group are requests that are abstract and require knowledge of the archivist about the archive, about the programming of the broadcaster and about the world in general. Examples of highly abstract queries would include requests as, "Find statements of politicians who are pessimistic about the economy," "Find segments discussing prejudices of and against Germans," and "Find excerpts on the negative side of being famous." Archive staff see very limited potential for keyword search for these types of requests. They rely largely on knowledge accumulated during annotation work about which shows might include such fragments and which politicians or public figures might be inclined to make such statements.

The third group of requests can be approached by using full-text search of speech recognition transcripts. The most straight-forward cases are when the archivist is searching for a known quotation, such as the famous 1997 quote of Roman Herzog, President of Germany, "Durch Deutschland muß ein Ruck gehen." In the case of the original speech, the archive staff have probably annotated it already with a transcription of this quote. However, full text search on speech recognition transcripts makes it possible to find which politicians have quoted Roman Herzog since the original speech. Speech recognition transcripts are also invaluable to find buzzwords or currently important phrases spoken in different contexts by different people, such as "Harz IV", the labor reform. Also, many topics have indicative keywords which can be assumed to appear in the transcripts when these topics are discussed.

3. RADIO ARCHIVE PROTOTYPE

During the course of the project, the project group met on a regular basis to define the functional specifications of the prototype and to design the prototype interface. This section gives information about the prototype definition process and about the resulting system.

3.1 Data

The radio archive prototype needed to contain the full range of types of radio programs that the broadcaster archives must handle. Four programs were chosen to cover these types, two from each radio broadcaster. Deutsche Welle contributed approximately 80 hours of material from *Funkjournal* and *Wiso* two programs containing news reports and interviews. If the interview is not conducted in German, a clip of the interviewee responding in the interview language is played before the German translation is blended in. If the interviewee is speaking in English, the whole answer is played and then repeated translated into German. WDR contributed 40 hours from *Der Tag*, which contains reports, interviews, opinions and music. WDR also contributed 40 hours from *Montalk*, a two hour interview talk show featuring prominent figures from media, sports, politics and culture. Each show includes surprise guests, some physically present in the studio and some participating in the show via telephone. It also contains collages or short recordings made of people answering interview questions on the street as well as music. *Montalk* is the most challenging of the 4 programs because it contains nearly exclusively conversational speech characterized by laughter, interruptions and speakers with regional and/or colloquial speech. The radio archive prototype contained 160 hours of total material broadcast in 2005. As is discussed later in the paper, 12 hours was chosen to be annotated as a test set for the evaluation of the performance of the system.

3.2 Prototype functionality

The archive professionals in the project group defined the functionality that was most critical for the radio archive prototype. In this section, three of these functionalities that are related to spoken keyword search are discussed. In the following section implementation of these functionalities in the user interface is discussed.

First, it was deemed essential that the audio archive prototype not be dependent on a speech recognition vocabulary. The requests for information that archivists must respond to deal with a disproportionately large number of proper names. Archive staff need to be able to respond to requests concerning a rare name or a previously rarely mentioned place. The speed with which new proper names can break into the media was dramatically illustrated by the data set which included material broadcast in the weeks after the 2004 tsunami. This material contained many names of smaller places in Thailand and of people who were feared to be missing. Even the largest vocabulary cannot guarantee complete coverage of the large variety of human names. Archives are interested in exploiting marginal topics in radio collections, such as people or places that were mentioned only fleetingly. Such sound bites are of great value if the mentioned people or places subsequently achieve a high public profile.

Second, it was important that the prototype allow intelligent skimming of audio. Currently, archivists screening a radio program need to make best guess jumps when they fast forward. Intelligent skimming means offering the archivists signposts so that jumps can be informed. These signposts can take the form of segment boundaries or of segment labels. Archivists are then able to skip over music within a radio program and listen in only to speech. Signposts can also take the form of keywords. Archivists wanted to be able to type in a keyword and see at which places it is spoken within a radio program. In this way, an archivist can use

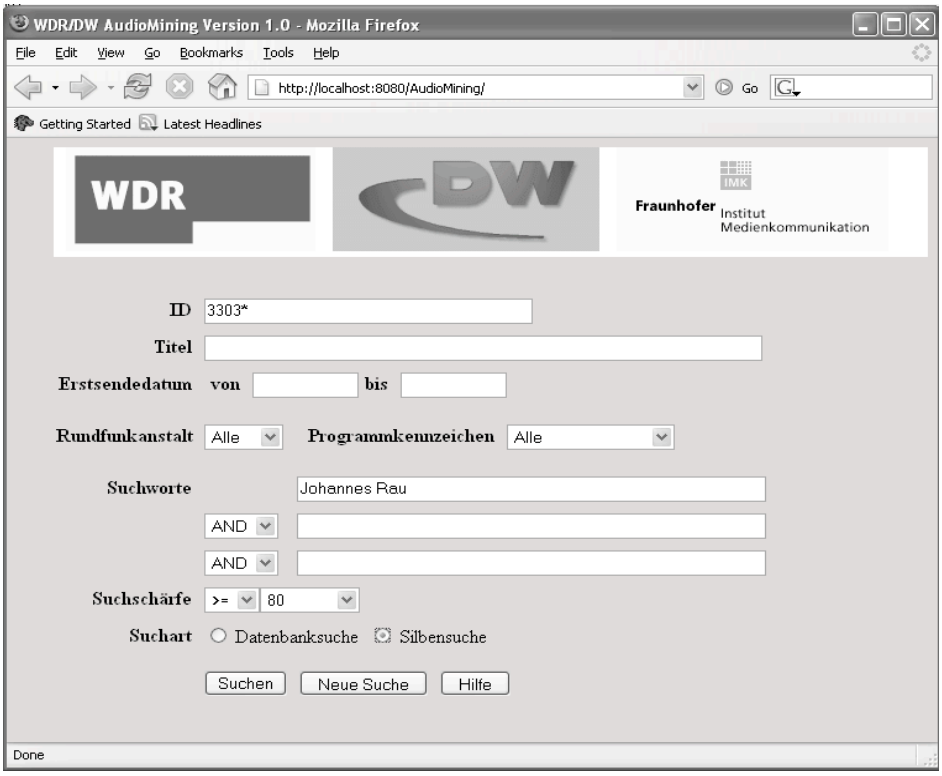

Figure 1: The search interface for the radio archive prototype

previously existing notes or titles to chose keywords and perform a search using these keywords to localize certain topics. In order to skim through an interview, archivists can jump from question to question by tracking audio segments containing the voice of the host through the audio file.

Third, it was important that the archive allow combined search on formal data and on spoken audio content. As mentioned above, many topics have indicative keywords such as "Theo van Gogh" or "Arctic Monkeys." Using the formal metadata to restrict search to shows broadcast around the time of van Gogh's death is a form of integration of world knowledge that archivists often exploit. Archivists would also tend to expand a query for information on Arctic Monkeys with the names of the band members. Archivists use outside sources or their own knowledge to implement these refinements. The project group realized how important it is to maintain possibilities of combining formal metadata with content for search and of maintaining the possibility to use familiar search strategies involving the integration of outside information. Exploring the possibilities of automatic query expansion fell outside the scope of the project.

3.3 Prototype interface

The prototype interface was a joint design created by the archive professionals and the multimedia researchers in the project group. This section discusses how the required functionality was integrated into the user interface.

Figure 1 depicts the search mask of the radio archive prototype. This mask adopts the search fields for formal metadata currently used for search in the archive metadata database. The fields are ID, title, broadcast date (specified as a range), broadcaster and station. These fields are augmented with fields that make possible search in the spoken content. In the area labeled "Suchwort," three terms can be input and joined by and- or or-operators. Archivists input orthographic words, but phonetic transcriptions of words are also accepted as input for system test and development purposes. In the field labeled "Suchschärfe," the user can input the degree of acoustic match required between the query and the transcript. This degree is a match-score which represents the distance between the input string and the syllable transcript, as will be explained in detail later. For the purpose of the interface, the match-score is chosen on a scale of 1-100, although it is not technically a percent. The match-score does not have a direct relation to the underlying algorithm, but was chosen to resemble a percent because users had best intuition of its purpose this way. Finally, the user can specify the kind of search. "Silbensuche" is syllable search in the syllable transcripts and "Datenbanksuche" is search in the database in which popular syllable searches have previously been stored. At the bottom are buttons labeled "search," "new search," and "help."

In Figure 1, the query that is being entered is for "Johannes Rau" and the system is being constrained to operate on particular archive numbers beginning with "3303." The possibility to constrain the system in this way is necessary so that archivists can continue to integrate their knowledge about the world and about the contents of the archive. Some of the functionality requested by the archive staff initially did not

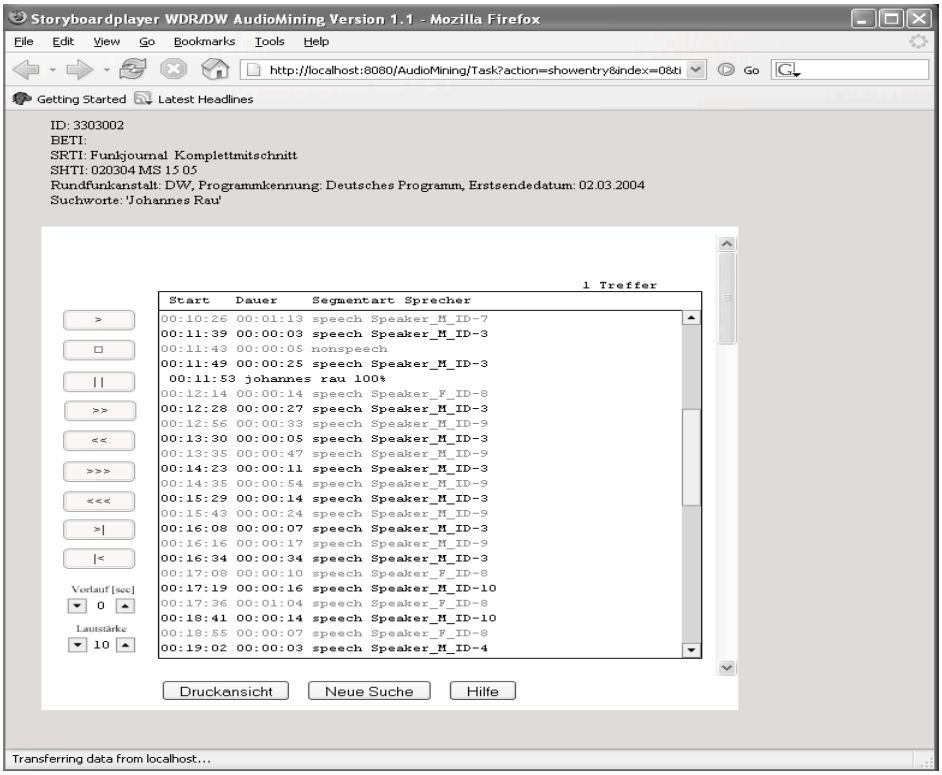

Figure 2: Display of a radio program with a keyword hit

seem particularly user-friendly to the multimedia researchers in the project group. As the researchers learned more about archivist workflows, however, it became clearer why some of the features were recommended. In this case the possibility of inputting a truncated archive number relates to the semantic structure of the archive IDs within the archive, which archive staff are thoroughly familiar with and use daily.

Figure 2 displays the structured audio player in which the retrieved radio programs are played. The programs are depicted as a list of segments. Keywords are depicted below the segments containing them together with their match scores. Note that the interface represents hits as orthographic words. The audio file can be played back using the buttons at the left, which allow for normal playback as well as accelerated playback at two different speeds. Clicking on a keyword jumps into the audio file and starts playback at the moment when the keyword is spoken. If the user wishes to add lead time, the button "Vorlauf" makes it possible to set the number of seconds before the keyword that the playback starts. The user can also adjust volume with "Lautstärke." This list can be exported for print via the button "Druckansicht". Archive staff indicated that a possible integration into the currently existing workflow would be to transfer this list per cut-and-paste from the print view to another application. This list can then form the basis for the hand generated annotations, eliminating the tedious work of setting time codes leaving archivists free to concentrate on creating high level summaries. The structured audio player is described in detail in [12].

In general, archive staff recommended a minimalist interface. Archive staff prefer to see as much information as possible on a single screen and avoid clicking "next" or scrolling to see additional hits or cuts. When the system displays results to the user query, the individual fields are marked with standard abbreviations, as can be seen at the top of Figure 2. These abbreviations are part of the daily vocabulary of (German-speaking) archivists, but appear cryptic to lay persons. In sum, the radio archive system was designed to be used on a regular basis by highly trained experts with a very different view on what is intuitive and user-friendly than a non-specialist or an occasional user.

The balance of this paper focuses on the keyword search functionality of the radio archive prototype, explaining the algorithm used to perform the fuzzy match between query and transcript and presenting the results of the evaluation of the system.

4. VOCABULARY INDEPENDENT SEARCH

The vocabulary independent search calculates a distance score between the user query and transcripts generated by automatic speech recognition (ASR). This score is calculated in a way that is intended to capture acoustic similarity. The system is required to find places in the speech recognition transcripts that "sound" the same as the user query. Search by acoustic similarity rather than by word match has the benefit of freeing the system from dependence on the vocabulary of the speech recognition. The approach promises another advantage as well. Retrieval systems that perform exact match in word transcripts are sensitive to speech rec-

ognizer errors. If a "sounds like" match is performed rather than an exact match, it opens the possibility that a spoken word or phrase is identified correctly despite the presence of speech recognition errors. The corrective power of the fuzzy-match technique applied by the radio archive prototype relies on the insight that speech recognizer errors are often caused by acoustic confusion. This section first describes the generation of the ASR transcripts and then details the distance calculation.

4.1 Syllable transcripts

The radio archive prototype implements vocabulary independent keyword search on the basis of syllable level speech recognition transcripts. The syllable constitutes a basic building block of speech. Words that are not contained in the training data can be reconstructed from the syllable transcripts by searching for the appropriate sequence of component syllables. Approaches using linguistic units at the phoneme level have long been popular [4, 3]. These units have the advantage that they form a very small and closed set, but have the disadvantage that they are too small to provide the large acoustic contexts needed for optimal speech recognition performance. Larger units, such as morphemes and in our case syllables are also popular [2, 13, 1]. Although such units do not form a closed vocabulary, it is possible to attain good coverage of a language with a relatively restricted inventory. The project also aimed to explore other advantages of syllables such as the potential for smaller, faster language models that require less training data.

The speech recognition transcripts used in the radio archive prototype are generated by the ISIP speech recognition system HMM-based speech recognition toolkit[5]. Instead of a word-level vocabulary, however, a syllable-level vocabulary is used. The language model is trained on a corpus consisting of 64 million running words from German newswire. A tri-gram syllable language model is trained by decomposing the word level text into a syllable level text using the transcription module from a speech synthesis system [15]. The syllable vocabulary contains the top 5000 most frequent syllables. Previous work has shown that at this vocabulary size, the performance of syllable recognition levels off [11]. Previous work has also shown that this tri-gram syllable model attains a syllable accuracy of 75% on studio quality speech, which is the same syllable rate achieved by our 91k word-level bi-gram language model. A 75% syllable rate was estimated to correspond to a 68% word accuracy [10].

4.2 Fuzzy syllable search

The algorithm that matches query words with acoustically similar points in the syllable transcripts is based on a two-stage Levenshtein distance. First, the query word is decomposed into syllables using the same transcription module that decomposed the training data for the syllable language model. Then, the fuzzy match algorithm calculates the Levenshtein distance between the query syllable string and each position in the syllable transcript. This Levenshtein distance is weighted using a acoustic similarity score between syllables. The acoustic similarity score is itself another weighted Levenshtein distance between the strings of phonemes that compose the syllables. The weights are calculated using confusion information derived from analyzing the performance of the speech recognizer. Substitutions between phonemes easily confused by the recognizer receive a lower penalty than substitutions between phonemes rarely confused by the recognizer. Finally, positions in the syllable transcript that receive a similarity score above a certain threshold are hypothesized by the system to be hits for the query word. This threshold is determine empirically and reflects the "fuzziness" or "exactness" of the match between the query and the hit. The interface gives the user the ability to adjust the match-score threshold, providing control over the tradeoff between precision and recall.

5. EVALUATION

The system was evaluated by using 213 queries that were chosen by archive professionals to reflect the kinds of information requests they receive. The queries consisted of both single words and multi-word phrases and the system was required to return the positions in the audio files at which the word or phrase was spoken. A lot of effort was devoted to creating a representative and well-distributed query list, since the performance of the system was to be evaluated on the basis of whether or not it was able to provide archivists with appropriate responses.

First and foremost the system was evaluated on speech recorded in studio conditions. The project group placed primary emphasis on attaining adequate performance under studio conditions since the group was pessimistic about potential retrieval performance on telephone speech, speech recorded on the street, or speech with music or foreign speech background. Notice that studio speech comprises the greater portion of the interview talk show *Montalk*. Recall that *Montalk* was particularly important for the radio archive prototype since it contains long expanses for which no production metadata are available. *Montalk* stands to benefit greatly from being made accessible to archivists through content-based keyword search. *Montalk* is also the most challenging of the 4 programs contained in the archive because it contains nearly exclusively spontaneous speech.

Retrieval performance was tested on 12 hours of material from the radio prototype archive, 4 hours each from *Montalk* and *Der Tag* and 2 hours each from *Funkjournal* and *Wiso*. Roughly estimated, the test material contains 50% spontaneous speech and 10% music and commericals. The data was annotated with segment boundaries between speakers and between speech and non-speech. Each segment was assigned a label relating to the acoustic quality of the audio in that segment. The data were transcribed by a professional transcription service and then automatically aligned with the audio files, so that each individual word spoken was associated with a time code. Then, human annotators listened to all 12 hours and checked the time codes of the words, correcting the alignment when the ASR alignment software had committed an error. The remainder of this section details the tests that were performed on this test set using the archivist-defined queries.

5.1 The effects of fuzzy match

As previously mentioned, the user interface provides the user with control over how much acoustic mismatch the system should admit between the query and the syllable transcripts. Table 1 reports precision, recall and F1-Value for five differ-

Table 1: System Performance at different match-score levels

Level	Precision	Recall	F1-Value
60	0.28	0.72	0.41
70	0.45	0.62	0.52
80	0.56	0.51	0.54
90	0.5	0.39	0.44
100	0.45	0.31	0.37

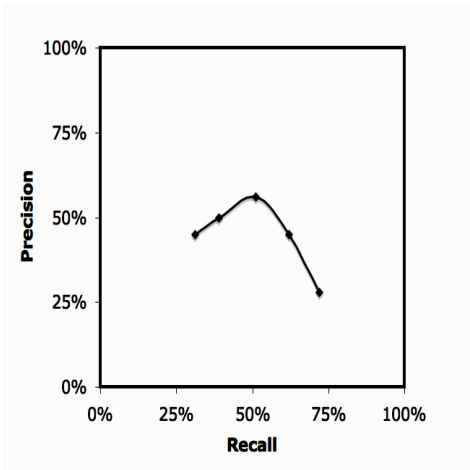

Figure 3: Recall vs. Precision on studio quality audio

ent settings of the match-score parameter. At low match scores, precision is low and recall is high. An interesting effect is that precision hits a peak and then deteriorates as the match is forced to be more and more exact. This effect is due to the fact that at high levels of exactness, many correct matches are not longer contained in the list of the hits returned by the system and false positives returned due to ASR errors dominate the shorter hit list. Figure 3 shows plots precision vs. recall, illustrating clearly the rise and fall of precision. After the radio archive prototype had been implemented, archivists experimented with the system and discovered that the most useful operating point was one at which the precision and the recall were approximately matched, i.e. the break-even point. This operating point occurs at match-score 80. In the remainder of this discussion, the system is evaluated at match-score 80, since this point had the highest utility for supporting archive workflows. Although it was disappointing that the system did not achieve a higher level of performance, the advantages of the fuzzy-match approach are clear, since the system at match-score level 80 clearly outperformed exact match (i.e. system at match-score 100).

5.2 The effects of implicit decomposition

A side-effect of the fuzzy-match approach used by the radio archive prototype is to introduce an implicit decomposition into the keyword search. The speech recognition transcripts are strings of syllables and do not represent word boundaries, which are not hypothesized by the ASR system. For this reason, a position in the speech transcripts that has a high match score with respect to the user query might actually correspond to a partial word in the spoken audio. There is no simple strategy for implementing a way to "turn off" this effect. The implicit decomposition means that a user query *Kinder* ("children") will return points in audio files at which the word *Kindergarten* ("kindergarten") is spoken. For this example, returning a compound containing the query word is probably not going to hinder the archivists' work. Indeed *Kindergarten* does have relevance to children. However, the situation is different if the original query was for *Garten* ("garden"). Here the system also returns points in audio files at which *Kindergarten* is pronounced. Such hits are clearly semantically far afield from the original query and effectively lower the precision of the system.

Table 2: Effects of implicit decompounding at match-score 80

Type	Precision	Recall	F1-Value
Admit partial word matches as correct	0.56	0.51	0.54
Exclude partial word matches as correct	0.52	0.53	0.53

Table 3: Effects of adding out-of-studio, telephone quality and music-background material at match-score 80

Type	Precision	Recall	F1-Value
Studio quality only	0.56	0.51	0.54
All except music	0.51	0.46	0.48
All audio qualities	0.51	0.45	0.48

The project group decided to evaluate the performance of the system under the stringent requirement that only exact matches be counted as correct hits. The purpose of this evaluation was to determine to what degree the return of compound words containing the query word lowered the precision of the system. The results on studio speech at match-score level 80 are reported in Table 2. If partial words (i.e. compound sub-units) are excluded as correct hits the precision declines somewhat. At the same time, recall improves slightly, since the system no longer was required to find every instance of an acoustic match. The over-all effect was only a slight, possibly insignificant, deterioration of system performance. If the query list compiled by the project group is taken to be representative of the queries that an archivist would submit to the system during the normal course of responding to information requests, it can be concluded that the implicit compounding of the system is not an aspect of the prototype that will cause an increased burden on archivists in their work.

5.3 The effects of including non-studio speech

At the end of the project, the performance of the radio archive prototype was evaluated on all speech types in order to ascertain what kind of deterioration of performance could be expected. Table 3 provides a comparison of system performance on studio speech with performance on studio plus non-studio speech and with performance on all audio qualities, including spoken audio with music background. It can be seen that when the system moves beyond its self-imposed restriction to studio speech, performance deteriorates. The

same pattern of deterioration was observed for all levels of match-score, although it is reported here only for match-score 80. The level of deterioration is not such that it would motivate the exclusion of non-studio speech from the system. Indeed the performance of the system on all audio qualities introduced a tolerable drop in system precision.

6. CONCLUSIONS

The radio archive prototype discussed in this paper was built with the goal of acquiring a concrete, practical understanding of how automatic digital audio processing can be integrated into the existing workflows in archive departments at large radio broadcasters to support the work of archive staff. This focus of this paper was vocabulary-independent keyword search. It was shown that the syllable based fuzzy search algorithm delivers tolerable performance without relying on a pre-defined vocabulary and is not derailed by acoustically challenging audio. Although the prototype was designed to retrieve keywords from German audio only, the fact that the system occasionally returns a proper name from the small fraction of English audio in the archive suggests that the method holds promise for keyword search in a multilingual archive.

The tests performed on the system demonstrate that higher precision rates can only be achieved with significant sacrifices in the area of recall. Archivists do not consider the system at its current level of performance to provide significant support to their workflow. This conclusion must be seen against the backdrop of the fact that many requests for information can be satisfied by searching formal metadata only or are so abstract that they could not be met using keyword search, even if the performance were perfect.

The keyword search functionality implemented in the radio archive prototype demonstrates three aspects which confirm its clear potential in the future for archive staff support. First, vocabulary independent archive access is indeed possible. Second, the implicit de-compounding that is a by-product of the fuzzy-match search approach has a very limited negative impact on precision. Third, inclusion of all types of audio in the archive and not just audio recorded under studio conditions does cause deterioration of system performance, but not to an extreme degree. In sum, the vocabulary independent keyword search method implemented in the radio archive prototype continues to hold promise for the future, even in the face of the challenges offered by a collection containing a large amount of spontaneous speech such as occurs in interview talk shows.

7. ACKNOWLEDGMENTS

We would like to acknowledge Deutsche Welle and WDR, who commissioned the project and provided the data. Thank you to the members of the archives departments who worked in the project group. The final work for this paper was carried out by the first author while being supported by the EU IST programme of the 6th FP for RTD under project MultiMATCH contract IST-033104.

8. REFERENCES

[1] G. Choueiter, D. Povey, S. Chen, and G. Zweig. Morpheme-based language modeling for Arabic LVCSR. *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, 1, 2006.

[2] M. Elbeze and A. Derouault. A morphological model for large vocabulary speech recognition. *Acoustics, Speech, and Signal Processing, 1990. ICASSP-90., 1990 International Conference on*, pages 577–580, 1990.

[3] A. Ferrieux and S. Peillon. Phoneme-level indexing for fast and vocabulary-independent voice/voice retrieval. *ESCA ETRW on Accessing Information in Spoken Audio*, pages 60–63, 1999.

[4] J. Foote, S. Young, G. Jones, and K. Sparck Jones. Unconstrained keyword spotting using phone lattices with application to spoken document retrieval. *Computer Speech & Language*, 11(3):207–224, 1997.

[5] A. Ganapathiraju, N. Deshmukh, J. Hamaker, V. Mantha, Y. Wu, X. Zhang, J. Zhao, and J. Picone. ISIP Public Domain LVCSR System. *Proceedings of the Hub-5 Conversational Speech Recognition (LVCSR) Workshop*, 1999.

[6] J. Garofolo, C. Auzanne, and E. Voorhees. The TREC spoken document retrieval track: A success story. *Text Retrieval Conference (TREC)*, 8:16–19, 1999.

[7] J. Gauvain, L. Lamel, and G. Adda. The LIMSI Broadcast News Transcription System. *Speech Communication*, 37(1-2):89–108, 2002.

[8] N. Hans and J. de Koster. Taking care of tomorrow before it is too late: A pragmatic archiving strategy. *116th convention of the Audio Engineering Society*, May 2004.

[9] A. Hauptmann and M. Witbrock. Informedia: News-on-demand multimedia information acquisition and retrieval. *Intelligent Multimedia Information Retrieval*, pages 215–239, 1997.

[10] M. Larson and S. Eickeler. Using syllable-based indexing features and language models to improve German spoken document retrieval. *Eurospeech'03*, pages 1217–1220, 2003.

[11] M. Larson, S. Eickeler, K. Biatov, and J. Köhler. Mixed-unit language models for German language automatic speech recognition. *Elektronische Sprachsignalverarbeitung, Tagungsband der*, 13:127–134, 2002.

[12] M. Larson and J. Köhler. Structured Audio Player: Supporting radio archive workflows with automatically generated structure metadata. *Proceedings of RIAO 2007*, 2007.

[13] K. Ng. *Subword-based approaches for spoken document retrieval*. PhD thesis, Massachusetts Institute of Technology, 2000.

[14] G. Rigoll. The ALERT system: Advanced broadcast speech recognition technology for selective dissemination of multimedia information. *Automatic Speech Recognition and Understanding, 2001. ASRU'01. IEEE Workshop on*, pages 301–306, 2001.

[15] K. Stöber, P. Wagner, J. Helbig, S. Köster, D. Stall, M. Thomae, J. Blauert, W. Hess, R. Hoffmann, and H. Mangold. Speech synthesis by multilevel selection and concatenation of units from large speech corpora. *W. Wahlster (ed.), Verbmobil: Foundations of speech-to-speech translation*, 2000.

Advances in SpeechFind: CRSS-UTD Spoken Document Retrieval System

Wooil Kim, Murat Akbacak, and John H. L. Hansen
Center for Robust Speech Systems (CRSS)
Erik Jonsson School of Engineering and Computer Science
University of Texas at Dallas, Richardson, Texas, USA
{wikim, murat.akbacak, John.Hansen}@utdallas.edu

ABSTRACT

This paper presents our recent advances with SpeechFind and the U.S. based Collaborative Digitization Program (CDP). SpeechFind [1] is a spoken document retrieval (SDR) system consisting of two phases which include enrollment of audio material and online retrieval. A proto-type of SpeechFind for the CDP is currently serving as the search engine for 1,300 hours of CDP audio content. Analysis on the CDP corpus shows that the audio content includes a wide range of acoustic conditions, vocabulary selection, and topics. In an effort to determine the amount of user corrected transcripts needed to impact automatic speech recognition (ASR), a web-based online interface for verification of the ASR-generated transcript was developed. In this study, we also present two advanced fusion approaches to merge subword and word based retrieval methods within a multilingual SDR system. We focus on creating robust multilingual SDR systems employing both word-based and subword-based retrieval methods. Presented algorithms employ an OOV-word detection module to generate hybrid transcripts/lattices. In the Dynamic Fusion (DF) approach, hybrid transcripts/lattices are used to assign dynamic fusion weights to each subsystem. In the Hybrid Fusion (HF) approach, queries are searched through hybrid lattices. Experimental results on the CDP corpus demonstrate that acoustic model adaptation using the verified transcripts is effective in improving recognition accuracy. Through combining several methods, 16.5% relative improvement in ASR was obtained on relatively low SNR audio documents. The fusion algorithms are evaluated in a proper name retrieval task within the Spanish Broadcast News domain, where the presented algorithms yield improvements over traditional fusion methods.

Keywords

SpeechFind, spoken document retrieval, CDP, multilingual, audio indexing, subword

1. INTRODUCTION

As available online collections drastically increase, the need for automatic and efficient information retrieval continues to expand, placing demands on advances in technology including computational power and storage capacity. Recently, there has been growing interest in retrieving information, especially, online for multimedia data consisting of rich information such as audio, video and speech. Today, multimedia information collections include radio/television broadcast news, interviews, entertainment content, User Generated Content (UGC), and others. This increasing demand has drawn remarkable attention to research on Spoken Document Retrieval (SDR).

SpeechFind is a SDR system serving as the platform for several programs across the United States for audio indexing and retrieval including the National Gallery of the Spoken Word (NGSW) and the Collaborative Digitization Program (CDP) [1, 2, 3]. The system consists of two main phases; (i) enrollment and (ii) online search retrieval. Our recent work on SpeechFind has included an effort to improve performance by addressing band-limited speech among wide range of acoustic conditions [4].

This paper provides an overview of recent advances in the SpeechFind system and collaboration with the CDP. A prototype of SpeechFind for the CDP has been established to serve as the search engine for the CDP corpus which presently contains 1,300 hours of audio documents. An online system for verification of the ASR-generated transcripts has been developed to improve the speech recognition engine and evaluate overall transcript generation performance. The corrected transcripts from the verification process were used for acoustic model enhancement by applying model adaptation algorithms.

In this study, we also focus on rapid transition to resource-limited target languages within the context of multilingual audio indexing and retrieval. Our goal therefore, is to develop robust retrieval methods for new target languages. Different tiers might result from changing lexicon coverage (poor coverage at the beginning, with an evolving lexicon as more resources are employed to obtain better coverage), or data sparseness or mismatch during acoustic model training.

To address the problem of misrecognition (both in-vocabulary word and out-of-vocabulary word errors) during SDR, previous studies have employed fusion methods [5, 6, 7, 8, 9] to

recover from recognition errors during retrieval. In these methods, fusion weights are optimized for specific tasks. More importantly, these methods assume a homogeneous audio collection where the ASR system achieves similar performance in each audio document. Results from these studies show small improvements over word-based-only retrieval approaches as the number of documents increases.

In our work, in order to use fusion methods more effectively for changing tiers within SDR, for each utterance we first run an OOV (Out-Of-Vocabulary) word and misrecognition detection module. Based on the output of this module, the first fusion algorithm, Dynamic Fusion (DF) approach, decides how to merge subword and word based retrieval scores dynamically. The second algorithm, Hybrid Fusion (HF) approach, searches query words through the hybrid lattice. These methods could be considered as dynamic back-off strategies where subword-based retrieval scores are merged with word-based retrieval scores according to the performance level of the word-based recognizer.

This paper is organized as follows. We review the SpeechFind system and recent collaboration with the CDP in Sec.2-4. In Sec.3, we discuss the structure of the audio materials from the CDP corpus. Sec.4 presents development of the transcript verification process including online web-interface. In Sec.5, we present the fusion methods for multilingual SDR system including Dynamic Fusion (DF) and Hybrid Fusion (HF). Representative experimental procedures and their results are presented and discussed in Sec.6. Finally, in Sec.7, we summarize and provide conclusions.

2. OVERVIEW OF SPEECHFIND

SpeechFind is a spoken document retrieval system developed to serve as the search engine for the National Gallery of the Spoken Word (NGSW) [1, 2]. The system includes the following modules: i) an audio spider and transcoder, ii) spoken documents transcriber, iii) transcription database, and iv) an online public accessible search engine. The audio spider and transcoder are responsible for automatically fetching available audio archives from a range of available servers and converting the incoming audio files into the designed audio formats for processing. This module also parses the metadata and extracts relevant information into a "rich" transcript database to guide future information retrieval.

The spoken document transcriber includes an audio segmenter and transcriber. The audio segmenter partitions audio data into manageable small segments by detecting speaker, channel, and environmental change points. The transcriber decodes every speech segment into text using a large vocabulary continuous speech recognition (LVCSR) engine.

The online search engine is responsible for information retrieval tasks, including a web-based user interface as the front-end and search and index engines at the back-end. The web-based search engine responds to a user query by launching back-end retrieval commands, formatting the output with the relevant transcribed documents that are ranked by relevance scores and associated with timing information, and provides the user with web links to access the corresponding audio clips.

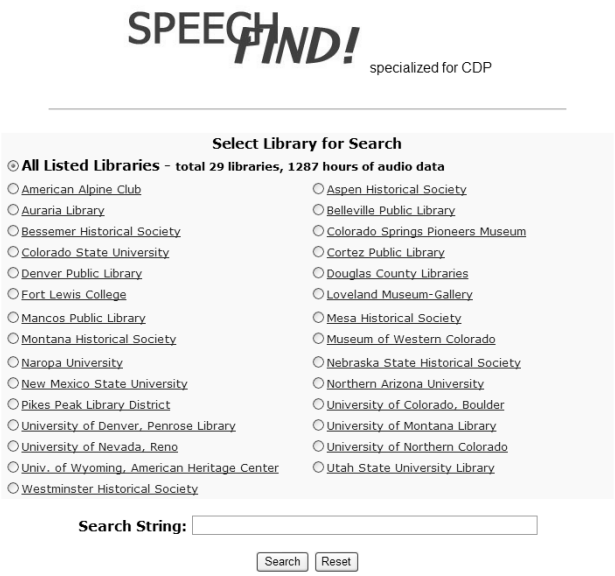

Figure 1: Main page of SpeechFind for CDP.

The SpeechFind system is also currently serving as the search engine for the CDP audio corpus, which has been established via a collaboration between CRSS and CDP program. Fig.1 shows the main page of SpeechFind specialized for the CDP corpus (http://SpeechFind.utdallas.edu/index_cdp.html).

3. STRUCTURE OF CDP AUDIO CORPUS

In this section, we discuss the structure of the CDP corpus. From the available limited metadata, it is known that the CDP audio files include interviews, discussions/debates, and lectures, each with 2-5 speakers participants. The recorded audio documents are spontaneously articulated with many overlapping speakers, and burst noise events such as clapping, laughing, etc. which make speech recognition challenging. The content of the speeches include speakers' personal experience and opinions on social issues such as Word War II, Red Cross, civil rights, feminist activity, and other topics. The speakers are reported to be leaders in local communities including senators, professors, activity group leaders, etc. Recordings were conducted from the 1960s to 2000s and held at library offices, classrooms, homes, etc. Depending on the documents, there exists background noise which would occur due to recording media or transmission.

The audio corpus from a total 29 participants (libraries, societies, museums, etc.) are currently available on SpeechFind for search and retrieval, which have approximately 1,300 hours and 150 GB as shown in Table 1. They were automatically transcribed by our speech recognition engine for online document retrieval. Here, 5% of the total ASR-generated transcripts were verified by CDP participants via online correction phase for performance improvement and evaluation. Table 2 shows details on the CDP corpus which has been verified for evaluation. Although verified transcripts make up about 5% of the entire CDP corpus, they are expected to represent the characteristics of the entire corpus because they were evenly selected from across each

Table 1: Entire CDP corpus and verified parts for evaluation.

Entire Corpus	29 participants 1,286.5 hours (148.2GB)
Verified Parts	70.6 hours 18,651 audio segments 5.5 % of entire

Table 2: Details on CDP corpus; verified parts for evaluation.

Total number of words	512,435
Total number of vocabulary	20,003
OOV rate	2.34 %
Average SNR	23.65 dB
Perplexity	18.98
Entropy	4.23

document. Perplexity, entropy and OOV in Tables were obtained using CMU-Cambridge Statistical Language Modeling Toolkit [10]. OOV rates were calculated based on Broadcast News vocabulary consisting of 64K words which is employed for the current LVCSR engine. Total number of OOVs are 6,487 which mostly include name entities and some amount of miss-prints by proof-readers. The detected OOVs are used for updating the acoustic model and language model. We also plan to identify and correct the miss-prints in the transcripts to enhance the transcripts for model adaptation. SNRs were calculated using NIST Speech Quality Assurance software [11].

4. TRANSCRIPT VERIFICATION PROCESS WITH CDP

We recently established a proto-type of the transcript verification process with CDP. An online web-interface was developed in order to improve the quality of the ASR-generated transcripts. The transcript verification process is as follows:

(1) Automatic Transcription: the audio documents delivered by CDP participants are automatically transcribed via our speech recognition engine. The original audio data with format of stereo, 44.1kHz, 24bit PCM are converted into single channel, 16kHz, 16bit PCM for processing. Every audio document with a length of approximately 15-40 min is segmented into small segments (15-30 sec) using our developed segmentation algorithm [12]. Each segment is automatically transcribed by large vocabulary continuous speech recognition (LVCSR) engine currently employing SPHINX3.

(2) Online Verification: from each audio document, approximately 5% of the segments are selected in an approximate uniform manner across each file. The transcripts of the selected segments were uploaded to the online system where participants would log-in using their accounts for verification work. They would listen to audio clips and correct the uploaded transcripts via the online web-interface. The words newly appearing in the verified transcripts which are out of vocabulary employed by ASR are automatically detected and stored in separate files for future processing. Several types of transcription conventions are allowed when correcting transcripts such as *(unknown)*, *(noise)*, *(clapping)* and *(laughter)*. Fig. 2 shows the online web-interface for CDP

Figure 2: Online web-interface for transcript verification.

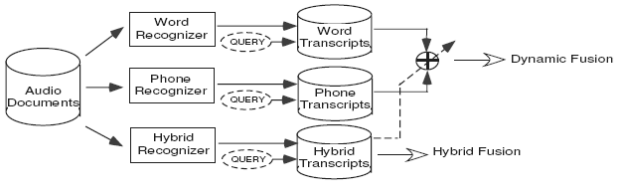

Figure 3: An overview of our work employing subword and word based retrieval systems.

user transcript verification.

(3) Model Enhancement: the corrected transcripts are used for model enhancement. Model enhancement was applied only to several participants in the current system and it will be applied to the entire audio corpus eventually. Our preliminary experimental results on model enhancement using the verified transcripts will be discussed in Sec.6. As shown in Table 1, 70.6 hours of speech segments have been verified via the online process, which is about 5% of the entire corpus.

5. FUSION METHODS FOR MULTILINGUAL SDR

This section focuses on the retrieval scheme employing our fusion algorithms for robust multilingual SDR system. Fig. 3 illustrates an overview of the SDR system with the fusion methods which will be presented in this section.

5.1 Baseline System

In our baseline system, we employ conventional fusion methods where subword and word based retrieval system outputs are merged with fixed weights for in-vocabulary query

words. For OOV query words, only subword-based retrieval method is employed. In the current system, only phonemes are used as subword units. In addition to 1-best recognition hypothesis, we also use N-best transcripts and lattices for our word-based retrieval and phoneme-based retrieval systems, respectively. In our word-based retrieval engine, we use a modified version of the MG [13] retrieval system. In this version, the *tfidf* weighting scheme is replaced with Okapi as explained in [1]. Stop-word removal and stemming are applied to the resulting ASR transcripts. For phoneme-based retrieval, we use the system developed in our previous study [14] where Finite State Transducers (FST) are used to index phone lattices, and to retrieve query words using confusion-embedded pronunciations.

5.2 Employed Fusion Schemes

We employ a hybrid-recognition based OOV detection module which is similar to the one proposed in [15]. Different from the word-based recognizer used in the baseline system, we use a generic word model (i.e., every OOV word is mapped to the generic word model) which allows arbitrary phoneme sequences during recognition. In our word lexicon and word-based language model, we use *UNK* tag for the generic word model. N-gram word language model treats *UNK* just like any other word in the lexicon. During decoding, at the end of every word hypothesis, we allow transitions into generic word model *UNK*, and within the generic word model *UNK*, the recognizer switches to the monophone language model and considers the phoneme set as the active lexicon. The module output is then used in the following algorithms.

5.2.1 Dynamic Fusion Approach

In this method, we assign dynamic weights to phoneme-based and word-based retrieval scores for each utterance [16]. First, we calculate the OOV detection/misrecognition probability (e.g., probability of having OOV-word or misrecognition in the utterance), and use this probability to decide when/how much to employ subword-based retrieval in addition to word based retrieval. For example, when we search for OOV words in a given set of utterances/documents, we employ subword based retrieval more in utterances where it is more likely to have an OOV word. To calculate the probability of having OOV-word or misrecognition in a given utterance/document, we take the ratio of the number of occurrences of the generic word UNK over the total number of words in the utterance assuming that OOV/misrecognition detection module is performing at a reasonable level.

5.2.2 Hybrid Fusion Approach

In this approach, we perform retrieval through hybrid recognition lattices generated by OOV detection module [16]. The motivation behind this algorithm can be explained with the following example. When the query word is OOV word, it is not the best approach to search for the query word in a monophone-only lattice since phoneme-based search is not discriminative enough to yield the desired performance. This will increase the recall rate but will have a negative impact on precision. In this method, we search for OOV words in parts of the lattice where the hybrid recognizer generates monophone sequences assuming that those parts of the lattice will correspond to in vocabulary words. In other words, OOV words are being searched in a smaller document space which is assumed to correspond to misrecognized words. For the case of searching in-vocabulary query words, we can use this algorithm in a fusion scheme whenever we want to back-off to subword-based retrieval (e.g., for a small number of returned hits using word-based retrieval).

6. EXPERIMENTAL RESULTS

6.1 Acoustic Model Enhancement

The transcripts corrected via online verification process are used to improve the performance of speech recognizer by enhancing the acoustic model. In this section, our efforts to improve the acoustic model for speech recognition utilizing the verified transcripts are presented. To evaluate the performance, the database for test and adaptation were selected among the CDP audio documents which have the verified transcripts. Table 3 shows the configuration of the database used for acoustic model enhancement.

A single number of audio documents from each library was selected for recognition test and model adaptation. From each audio document, approximately 5% of segments are selected as adaptation data (Adapt1) with uniform distribution of location across the audio. These adaptation data are used for updating the acoustic model to match the actual test condition. The remaining segments are used for recognition testing: total 60.2 min. The audio segments from the other documents are also used for adaptation (Adapt2) which are expected to modify the model to match the general acoustic conditions across the audio documents from the same library.

Table 4 shows the performance of baseline, spectral subtraction (SS) and MLLR (Maximum Likelihood Linear Regression) adaptation. MLLR is applied using the adaptation data selected from parts of test data (Adapt1) as presented in Table 3. The transcripts for adaptation are available, so the adaptation methods here were applied in a supervised style. By applying spectral subtraction and MLLR, relatively low SNR audio documents (e.g., DCL and UNC) show considerable improvement. We found that the test audio document from DCL and UNC include relatively high energy background noise which is considered due to recording media. However, in cases of AL and MESA which have relatively high SNR and low WER as baseline, the performance decreases even by applying spectral subtraction and MLLR. Spectral subtraction used in our study employs a noise estimation algorithm based on minimum statistics, which is known to be robust to slowly changing background noise. Failure to correctly estimate the burst noise such as laughing, clapping, etc. would result in degraded recognition performance. The speakers in the audio documents are not identified nor classified for model adaptation in our work, which would be another cause of performance decrease.

In the experimental results in Table 5, the speech samples from the other audio documents (Adapt2) are used for model adaptation. They have relatively large number of samples, therefore MAP (Maximum *A Posteriori*) adaptation is conducted. Similar to Table 4, test documents from DCL and UNC consistently show significant improvement by employing several combinations of MAP, spectral subtraction and MLLR. The audio documents from those two libraries are

Table 3: CDP database used for model enhancement.

Data	Number of segments (min.)
Test	532 (60.2)
Adapt1	30 (7.9)
Adapt2	3,186 (733.9)

Table 4: Model enhancement using Adapt1 data (WER, %).

Library	SNR (dB)	Baseline	SS	MLLR	SS +MLLR
AL	39.0	**41.1**	44.6	42.0	43.9
DCL	18.6	74.9	71.6	70.0	67.8
MESA	22.3	51.9	55.7	51.8	55.3
PPLD	22.5	75.4	74.8	**71.6**	73.6
UDPL	20.8	59.1	57.8	56.9	**56.7**
UNC	17.2	75.1	68.8	69.7	64.2
Avg.	**23.4**	**63.0**	**62.0**	**60.3**	**60.0**

found to have similar background noise across the comparable amount of other audio documents used for MAP.

The last column in Table 5 shows the cases for best performance of each library among the several combinations of model enhancement methods. For DCL and UNC which have relatively low SNR and baseline performance in WER, we obtained 16.8%[†] of averaging relative improvement by applying a combination of spectral subtraction, MLLR and MAP methods. A more elaborately designed algorithm is necessary to detect and estimate burst noises in the audio recordings for speech enhancement. It is also required to apply model adaptation in a discriminative way to different speakers of the same audio document by employing model classification and clustering techniques.

6.2 Evaluation of Fusion Methods for Multilingual SDR

We evaluated the fusion algorithms presented in Sec.5 on a a proper name retrieval task in the Spanish Broadcast News (Spn-BN) corpus [16]. In this task, we used Latin-American Spanish (LAS) as the target language, and focused on proper name retrieval within a broadcast news domain. It is important to note that sufficient resources clearly exist for Spanish based ASR development. While other languages (e.g., Dari, Pashto, Somalian, etc.) are possible, we selected Spanish to be able to intentionally limit the available resources to see what performance can be achieved as further data/resources are available. In other words, using a language such as LAS allows us to select the tier level of resources. In our experiments, we intentionally restrict the following resources: lexicon coverage in spoken documents as well as in query words. We denote OOV rate in spoken documents and OOV rate in queries as OOV_{doc} and OOV_{query} respectively.

We trained microphone speech models (Spn-Mic) and broadcast news models (Spn-BN) from Latino40 corpus and Spanish Broadcast News speech corpus [17]. We applied a bootstrapping approach to train microphone speech models for

[†]This value is calculated considering the amount of test data from each library, DCL (4.9 min.) and UNC (13.4 min.).

Table 5: Model enhancement using Adapt1 & Adapt2 data (WER, %).

Library	MAP	MAP +MLLR	SS+MAP +MLLR	**Best (relative%)**
AL	45.5	45.0	50.0	**41.1 (0.0)**
DCL	63.0	61.2	**59.4**	**59.4 (20.7)**
MESA	51.3	**50.6**	56.2	**50.6 (2.5)**
PPLD	74.0	72.5	75.8	**71.6 (5.0)**
UDPL	61.9	58.0	58.9	**56.7 (4.1)**
UNC	70.9	68.7	**63.8**	**63.8 (15.5)**
Avg.	**61.7**	**59.8**	**60.9**	**57.5 (8.8)**

Spanish by using English microphone models via a phone mapping during initial alignment, and then iteratively perform alignment and training steps with updated Spanish acoustic models as explained in [14]. Next, Spanish microphone models (Spn-Mic) are used to initially align the Spanish Broadcast News speech corpus, and then train Spanish Broadcast News models (Spn-BN) using 20 hours of speech corpus. Speech corpus used during retrieval experiments was kept separate.

Different lexicons were created for evaluation purposes: L_{45K}, L_{50K}, L_{51K}. L_{45K} was obtained from Callhome Spanish lexicon [17], and L_{50K} was created with an additional most frequently occurring 5K words from Spn-BN corpus. L_{51K} was created to contain all query words that are used in retrieval experiments. OOV_{doc} and OOV_{query} values for these lexicons are shown in the Table 6. N-gram ($N = 3$) language models at the monophone and word level were trained using Spanish Newswire Text corpus [17] consisting of 5 Million words. Bigram and trigrams occurring less than 4 times are pruned during N-gram counting. Sentences having high OOV rates (in our experiments sentences with more than 40% OOV rate) are also discarded in our language model training to prevent spelling errors, as well as high unigram probability for generic word *UNK*.

During recognition, we apply single-class MLLR adaptation. We report recognition results in terms of PER (Phone Error Rate), and SSF-WER (Stemmed and Stop-word-filtered WER) as illustrated in Table 7. Rather than WER results, to be consistent with our retrieval engine, here we report SSF-WER since our wordbased retrieval engine, MG, removes stop words and applies Porter stemming to the resulting transcripts, which is very common in text retrieval applications. We report recognition performance in terms of oracle performance for different N-best sizes since we perform lattice-based search during retrieval.

Table 8 describes the spoken document and query sets used in our evaluation. The test queries were designed to simulate a known item retrieval task. For each query, there is only one document considered relevant for the purposes of this evaluation. While other documents may have some relevance to the query, only the document it was designed to retrieve was scored as a correct retrieval. To reflect the nature of this task, we used Inverse Average Inverse Rank (IAIR) as a performance criteria. One characteristic of the IAIR is that it rewards correct documents near the top more than

Table 6: OOV_{doc} and OOV_{query} for different Spanish lexicons used proper name retrieval experiments.

	L_{45K}	L_{50K}	L_{51K}
OOV_{doc}	3.80%	1.15%	0.98%
OOV_{query}	100%	100%	0.00%

Table 7: Oracle error rates of (a) monophone and (b) word based ASR in Spn-BN corpus.

	$N=1$	$N=20$	$N=100$	$N=500$
monophone	24.39	22.10	19.64	17.46
word-based	33.03	30.45	28.72	28.20
	29.58	26.62	25.23	23.91
	29.31	25.92	24.82	23.65

documents in the middle or towards the end of the rankings:

$$IAIR = \frac{1}{\sum_{i=1} rank_i^{-1}} \quad (1)$$

where $rank_i$ is the rank of document i.

Proper name retrieval results are shown in Table 9. The presented fusion methods (DF approach and HF approach) perform better than the baseline system employing fusion approach with fixed weights. Another observation is that for changing lexicon sizes, our fusion methods yield more robust and consistent performance than the baseline. In other words, when lexicons with better coverage are used, baseline system performance does not change much (e.g., L_{45K}). On the other hand, our methods benefit from better lexicon coverage. This is mostly due to the fact that employing lexicons with better coverage does not guarantee less errorful ASR word transcripts, especially when less frequent words are under consideration as in our case where we try to retrieve proper names.

7. CONCLUSIONS

In this paper, we presented our recent advances in SpeechFind and collaboration with the CDP. A proto-type of SpeechFind for the CDP has been established serving as the search engine for about 1,300 hours of the CDP audio content. The web-based online interface for verification of the ASR-generated transcript has been developed for use in improving and evaluating the speech recognition engine. In this study, we also focused on audio indexing and retrieval having tiered resources (e.g., lexicon coverage, acoustic model accuracy, etc.). We presented two methods to obtain more robust retrieval performance for systems employing recognition systems operating at changing performance levels due to tiered structure. The first algorithm, Dynamic Fusion (DF), employed a hybrid recognizer to calculate OOV-and-misrecognition-detection probabilities to assign dynamic weights to each subsystem (subword and word based retrieval). In the second algorithm, Hybrid Fusion (HF), we used hybrid lattices, and performed searches through these lattices. Experimental results on the CDP corpus demonstrate that the acoustic model adaptation using the verified transcripts is effective in improving recognition accuracy. Through combining several methods, a 16.5% relative improvement was obtained on relatively low SNR audio documents. The framework and results here help suggest an effective process to provide transcription support to libraries with limited ASR/search expertise. Evaluation on the fusion methods showed that both DF and HF approaches perform better than the baseline system employing traditional fusion methods with fixed fusion weights in a proper name retrieval task.

Table 8: Description of document and query sets in Spanish BN.

Number of Documents	5,000
Average Length of Documents	9 sec.
Average # of Words per Documents	11 words.
Number of Queries	100
Average Length of Queries	6 phonemes
Number of Relevant Documents	100
Average Relevant Documents per Query	1 doc

Table 9: Inverse average inverse rank (IAIR) for proper name retrieval task within Spn-BN domain for different lexicons.

	Baseline	Dynamic Fusion	Hybrid Fusion
L_{45K}	1.69	1.53	1.48
L_{50K}	1.69	1.49	1.45
L_{51K}	1.68	1.48	1.45

8. REFERENCES

[1] J. H. L. Hansen, R. Huang, B. Zhou, M. Seadle, J. R. Deller, Jr., A. R. Gurijala, M. Kurimo, P. Angkititrakul, "SpeechFind: Advances in Spoken Document Retrieval for a National Gallery of the Spoken Word," *IEEE Trans. SAP*, 13(5):712-730, 2005.

[2] http://www.ngsw.org.

[3] http://cdpheritage.org.

[4] W. Kim and J. H. L. Hansen, "Missing-Feature Reconstruction for Band-Limited Speech Recognition in Spoken Document Retrieval," *Interspeech2006*, pp.2306-2309, Sept. 2006.

[5] M.G. Brown, J.T. Foote, G.J.F. Jones, K.S. Jones, and S.J. Young, "Open-vocabulary speech indexing for voice and video mail retrieval," *ACM Multimedia'96*, pages 307316, Boston, 1996.

[6] K. Ng and V.W. Zue, "Subword-based Approaches for Spoken Document Retrieval," *Speech Communication*, 32(3), pp.157-186, 2000.

[7] D.A. James and S.J. Young, "A Fast Lattice-Based Approach to Vocabulary Independent Word spotting," *ICASSP2000*, pp.1029-1032, 2000.

[8] M. Saraclar and R. Sproat, "Lattice-based search for Spoken Utterance Retrieval," *HLT-NAACL* 2004.

[9] C. Allauzen, et al., "General Indexation ofWeighted Automata - Application to Spoken Utterance Retrieval," *HLT-NAACL* 2004.

[10] http://www.speech.cs.cmu.edu/SLM/toolkit.html

[11] NIST SPeech Quality Assurance (SPQA) package version 2.3, http://www.nist.gov/speech.

[12] R. Huang and J. H. L. Hansen, "Advances in Unsupervised Audio Classification and Segmentation for the Broadcast News and NGSW corpora," *IEEE Trans. on ASLP*, 14(3):907-919, 2006.

[13] I.H. Witten, A. Moffat, and T.C. Bell, *Managing Gigabytes: Compressing and Indexing Documents and Images*, Morgan Kaufmann Publishing, 2nd Edition, 1999.

[14] M. Akbacak and J.H.L. Hansen, "Spoken Proper Name Retrieval in Audio Streams for Limited-Resource Languages via Lattice Based Search Using Hybrid Representations," *ICASSP-06*, 2006.

[15] T.J. Hazen and I. Bazzi, "A comparison and combination of methods for OOV word detection and word confidence scoring," *ICASSP2001* 2001.

[16] M. Akbacak and J.H.L. Hansen, "A Robust Fusion Method for Multilingual Spoken Document Retrieval Systems for Employing Tiered Resources," *Interspeech2006*, 2006.

[17] Linguistic Data Consortium http://www.ldc.upenn.edu

Examining the Contributions of Automatic Speech Transcriptions and Metadata Sources for Searching Spontaneous Conversational Speech

Gareth J. F. Jones Ke Zhang Eamonn Newman Adenike M. Lam-Adesina
Centre for Digital Video Processing
Dublin City University
Dublin 9, Ireland
{gjones,kzhang,enewman,adenike}@computing.dcu.ie

ABSTRACT

The searching spontaneous speech can be enhanced by combining automatic speech transcriptions with semantically related metadata. An important question is what can be expected from search of such transcriptions and different sources of related metadata in terms of retrieval effectiveness. The Cross-Language Speech Retrieval (CL-SR) track at recent CLEF workshops provides a spontaneous speech test collection with manual and automatically derived metadata fields. Using this collection we investigate the comparative search effectiveness of individual fields comprising automated transcriptions and the available metadata. A further important question is how transcriptions and metadata should be combined for the greatest benefit to search accuracy. We compare simple field merging of individual fields with the extended BM25 model for weighted field combination (BM25F). Results indicate that BM25F can produce improved search accuracy, but that it is currently important to set its parameters suitably using a suitable training set.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing—*Indexing Methods*; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Algorithms, Experimentation

Keywords

searching spontaneous speech transcriptions, metadata, data fusion, field combination

1. INTRODUCTION

Spontaneous speech forms a natural and often almost unconscious means of communicating information between individuals. Increasing archives of digitally recorded spontaneous speech are creating new opportunities to access information contained in this data. However, retrieving relevant content presents significant challenges. Previous research in spoken document retrieval (SDR) for broadcast news, notably the TREC-8 and TREC-9 tasks, has demonstrated that, when handled appropriately, there is little difference in retrieval effectiveness between errorful transcriptions generated using automatic speech recognition (ASR) and a near accurate[1] manual transcription [3]. However, much of this data is read speech and well defined distinct document units are generally easily identifiable. Data nearer to spontaneous speech was used in the Video Mail Retrieval using Voice (VMR) project [5], where there was degradation in retrieval performance for automatically indexed data compared to manual transcriptions, but in this case the manual transcriptions were completely accurate. While the documents in the VMR collection generally comprise spontaneous speech, they are still distinct individual documents.

Spontaneous conversational speech, where document boundaries are often not well defined, raises a number of new issues for search. The CLEF Cross-Language Speech Retrieval (CL-SR) uses data from the Malach oral history collection to explore retrieval of spontaneous speech with significant conversational elements in the context of cross-language information retrieval [9]. These collections can of course also be used to explore monolingual search without the additional complexities associated with cross-language search. An interesting feature of this collection is that ASR document transcriptions are accompanied by several automatically and manually derived metadata fields. Results from CLEF workshops held in 2005 and 2006 show that retrieval effectiveness using only the ASR fields is poor, while using metadata gives much better performance. It is not however clear exactly why this is the case, this topic is explored in more detail in Sections 2 and 3. Retrieval is clearly shown to be improved by combining metadata fields, with manual metadata being considerably more useful than automatic metadata. While the utility of field combination is clear, it is important to consider how these fields should best be combined for best results, we explore this issue in Section 4 with specific reference to the BM25 model, based on the analysis in [8], and report experimental results using the CLEF CL-SR collection in Section 6. Note that while metadata is clearly important for retrieval of spontaneous speech in the

[1]Word Error Rate (WER) $\approx 10\%$.

case of these CLEF collections, it is not possible to explore whether it might also improve retrieval effectiveness results for the earlier SDR tasks reviewed above beyond their already high values, since comparable metadata fields do not exist for these collections.

The remainder of this paper is organised as follows: Section 2 discusses the nature of spontaneous conversational speech, Section 3 examines searching this spontaneous speech and associated metadata, Section 4 explores issues in field combination and the BM25F model, Section 5 outlines the CLEF CL-SR test collections, Section 6 gives our experimental results and analysis, and finally Section 7 concludes the paper.

2. SPONTANEOUS VS SCRIPTED SPEECH

Unlike more deliberately generated written text communication or speech read from a script, when speaking spontaneously a person will often convey many details in an informal and unstructured way, and frequently make considerable use of the context in which they are speaking and the background of the audience which is being addressed, whether it be an individual, a business meeting, a class of students or a general interest grouping.

The degree of genuine spontaneity will depend on the circumstances in which they are speaking, and their experience in ensuring that what they say is unambiguous and will not return to haunt them in the future. Contrast the implications of a slip of the tongue in a social gathering between close friends or a business meeting with regular colleagues, and a live radio or TV interview of a leading politician or a contract negotiation meeting between companies. In the former cases perhaps a simple clarification or apology will often suffice if a slip is made, or perhaps no one will even notice and the exchange can proceed without interruption, in the latter cases there may be significant long term implications of using a certain expression or even implying something unintentionally. While the political interview is spontaneous in the sense that it is not scripted, the interviewer will often have an agenda of points that they wish to raise and include sufficient context in their questions and responses to inform the listener of the topic under discussion, and the politician will typically respond carefully, for the reasons stated above. In a social gathering or local business meeting such contextual information will generally be missing from the exchanges, since the participants are familiar with each other or the subject under discussion and many details to not need to be stated. Essentially there is a large degree of *tacit knowledge* at play in such exchanges.

Words spoken may make reference to what is known to the participants to establish or maintain common understanding of the point under discussion, but important words related to the topics may often be new or very specific ones important in conveying new information. Such words will often be outside the vocabulary of an ASR system meaning that they cannot be recognised correctly, and will therefore not appear in such automatically generated transcriptions and in consequence not be available for search. In the case of ASR transcriptions of spontaneous speech we might well expect them to contain words with low average specificity in terms of identifying relevant documents, i.e. recognised words may appear in a higher average proportion of document transcriptions, making it harder to rank relevant documents reliably. We demonstrate that this can indeed be the case in Section 6.

For more structured interviews the need to establish the context for listeners will mean that a greater number of topic related words appear and that these additional words are more common in the language as a whole. These more common topics words are more likely to be within the vocabulary of an ASR system, particularly if it has been adapted to the domain of interest. Obviously examining this hypothesis fully would require access to suitable corpora of accurately transcribed speech. However, if it is found to be even partially correct, the success of TREC SDR may be to some degree attributable to these words which can be recognised correctly, as well as redundancy and term co-occurrence effects. Searching spontaneous conversational speech may thus be an intrinsically much more difficult task.

3. SEARCHING SPONTANEOUS SPEECH

These observations potentially have significant implications for searching of spontaneous conversational as speech. If the words are not articulated between participants while expressing an opinion, developing an idea or clarifying some point, since they are already common knowledge, then searching an audio recording to find material containing content pertaining to these details is clearly going to present problems, since many of the obvious search words are just not present in the speech. This problem would be significant in itself if the details of the conversations were accurately transcribed. However, the volume of speech means that it is only practical to perform transcription using ASR which inevitably introduces introduces errors arising from various sources. Thus the issue of the absence of important context descriptive content in conversational speech is compounded by the presence of errors in the transcription. As has been observed previously [3], the issue of transcription errors has not proven to be a significant problem for searching spoken segments which can be broken into distinct documents, such as the easy segmentation of a news broadcast, and which are scripted to explain the context of the material covered, again as exemplified by broadcast news stories. However, conversational speech represents a new search problem combining the previously described problems of the absence of contextual review, ASR errors and also uncertainty in topic boundaries, and indeed even the scope of topics within the data to which boundaries might be assigned. Where there is a lower density of topic specific words being spoken, recognising individual spoken words correctly becomes more important. This lack of redundancy means that failure to correctly recognise individual useful words may have apparently disproportionately significant implications for retrieval.

In order in facilitate effective search of this errorfully transcribed data where topic boundaries are unclear and which lacks articulation of much of the associated knowledge, it would seem obvious to suggest that the content should be annotated with terms useful for improving search reliability. The question then arises how should such annotation, or descriptive metadata, be assigned to the speech transcription? One option obviously is to enter this manually, although this will often be extremely expensive, and will only be justified in limited cases. The other option is to seek to assign

metadata automatically or possibly semi-automatically. The availability of suitable metadata will depending on the type of data under consideration.

In the field of education there is growing interest in recording of lectures. These can then be made available for download for students for private study to reinforce lectures or distance learning. Beyond this basic use, lectures recordings are also potentially a very valuable new resource, since they are often sources of the lecturers tacit knowledge of a subject which they fail to include in written materials associated with the course, or which arise unexpectedly during the lecture, possibly promoted by questions from the audience. Whilst a student taking a particular course will be able to identify the lecture recording that they wish to access, as such archives grow it will clearly become impractical to search lecture archives manually. This will be true for small archives in the case of distance learning students or those searching a remote archive, when the student doesn't know exactly where the information they are interested in is located. Thus we should seek to make lectures searchable. Importantly making them searchable also significantly increases their value as a knowledge source for students wishing to learn about a subject. The first stage in making a lecture recording searchable is to transcribe the content using ASR. However, even after the correct lecture has been located, viewing a complete lecture takes a considerable amount of time, and efficiency in locating relevant sections of a lecture can be improved by adding structure to the lecture. Associated with a lecture there will often be a set of electronic slides.

In previous work we demonstrated that where such slides are available, even highly errorful lecture transcriptions can be segmented and assigned to their related slide with a high degree of reliability [4]. Associating relevant manually created metadata with each section of a noisy lecture transcription has several positive advantages. Since they are created manually, the contents of the slides are accurate, and since they are slides designed to support a lecture presentation, they are likely to contain concise statements of the key points to be raised in the lecture, and to do so using carefully selected vocabulary used to describe the topic under discussion. By contrast the ASR transcription of the lecture will contain mistakes and will almost certainly fail to recognise important domain specific words which are outside the vocabulary of the ASR system. In addition, the lecturer may fail to use accepted domain specific vocabulary in their description while extemporizing on the subject under discussion[2]. Thus annotating the transcription with the slides can improve the indexing and search of this content. Annotating spontaneous speech in this way is only possible if there is high quality descriptive content available that can be associated with the transcription. The structure of lecture presentation means that the problem is generally one of alignment within a limited search space. Other environments will constitute a much more challenging metadata association task.

While the contents of a formal lecture are generally spontaneous, they are not often truly conversational, unless the lecturer chooses to engage in extensive interaction with the class. Within education a small group tutorial forms a better example of a spontaneous conversational speech environment. Such sessions may possibly be even more valuable than formal lectures. The discussions will be largely unstructured with many unanticipated comments from the tutor and the students, with much greater potential for the expression of ideas that are not available in formal instruction associated with the course. This environment introduces the problems associated with searching spontaneous conversational speech discussed earlier. A key question if the speech is to be augmented with metadata for searching, where might this metadata come from? Research at IBM has explored the automated delivery of information associated with a meeting [1]. The *Meeting Miner* system performs live ASR on the audio stream emerging from a meeting, and analyses the resulting transcription to form questions or queries to archives related to the meeting, and returns items from the archive to the participants in an attempt to provide them with additional information that they may find useful to enhance their participation in the meeting. Information gathered in this way might potentially be used to annotate the meeting transcription, to more fully describe the topic under discussion in the meeting and thus potentially facilitate improved search. The key question here is whether materials can be chosen with sufficient selectivity and reliability to give improved search.

[2]This of course assumes that the lecturer is not reading from a script!

4. INFORMATION RETRIEVAL AND FIELD COMBINATION

Assuming that annotations can be suitably selected, there is the further important question of how the ASR transcription might be combined with metadata fields to provide most effective search. Two methods are typically used to process documents with multiple fields in retrieval. The simplest approach is simply to merge all the data for a document into a single vector losing the document structure, and then perform standard information retrieval. The alternative is to perform separate retrieval runs for the individual search fields, and then form a sum of the resulting ranked lists to produce a single combined document list for output. In this latter method, often referred to as *data fusion* the lists may be weighted prior to merging.

In this section we examine these methods in more detail in the context of the BM25 retrieval model [7] based on the review of this topic and proposed a simple multi-field extension model (BM25F) appearing in [8].

BM25 is a very successful weighting scheme based on the probabilistic model of information retrieval. The model was developed for standard single field documents such as those used in early TREC ad hoc search tasks. The standard model does not allow for exploitation of the structure of multi-field documents. However, as illustrated later, this approach can lead to problems in term weighting when we attempt to take account of the field structure in multi-field documents, due to the nonlinear treatment of within document term frequency ($tf(i,j)$) in the BM25 function.

4.1 The Problem

Consider an unstructured document j belonging to a collection J, where j can be regarded as a vector
$j = \{tf(1,j), tf(2,j), \ldots, tf(V,j)\}$ where $tf(i,j)$ is the *term*

frequency of the i term in j, and V is the total vocabulary. Documents can be scored against a query using a ranking function such as BM25, where BM25 is defined as follows,

$$cw(i,j) = \frac{tf(i,j) \times (k_1+1)}{k_1((1-b)+b \times ndl(j)) + tf(i,j)} cfw(i),$$

where

$$cfw(i) = \log \frac{N - n(i) + 0.5}{n(i) + 0.5}, \quad (1)$$

$cw(i,j)$ is the combined term weight of i in j, N = total number of documents in the collection, $n(i)$ = number of documents in the collection containing term i, $cfw(i)$ = collection frequency weight, $ndl(j) = dl(j)/\text{ave } dl$ = normalised document length, $dl(j)$ = length of document j, ave dl = average document length across the collection, and k_1 and b are scalar parameters. The standard document matching score $ms(j,q,J)$ is computed by summing the $cw(i,j)$ of terms matching a query q also represented by a vector and assumed to be unweighted $q = \{q(1), q(2), \ldots, q(V)\}$.

Consider a collection with a set of field types $T = \{1, \ldots, f, \ldots, K\}$, e.g. $f = 1$ ASR transcription, $f = 2$ assigned keywords, etc, and assum that the fields are non-repeatable and non-hierarchical.

A structured document $\mathbf{j}$ can be written as a vector of fields: $\mathbf{j} = \{j[1], j[2], \ldots, j[k], \ldots, j[K]\}$. Each $j[k]$ can be seen as a vector of term frequencies $(tf(i,j[k]))_{i=1,\ldots,V}$ similar to a standard unstructured document. $\mathbf{j}$ is thus a matrix, note any field may be empty for an individual document. Let $\mathbf{J}$ refer to the collection of structured documents. In order to weight the fields differently, define the field weight vector of each document as $\mathbf{v} \in R^K$. Without loss of generality, set one field weight, e.g. the ASR transcription, equal to 1.

When scoring a structured document for query q we want to take account of the document contents and the collection, but also the field structure and the relative weight vector $\mathbf{v}$. The problem is therefore how to extend a standard ranking function $ms(j,q,J)$ into a new function $ms(\mathbf{j},q,\mathbf{J},\mathbf{v})$. The extension model proposed in [8] basically assumes that similar words appear in different fields, although probably with different distributions.

Most modern term weighting functions, including BM25, have a nonlinear $tf(i,j)$ component. This is desirable since the information gained on observing a term the first time in a document is greater than that of each subsequent occurrence. In BM25 the term frequency saturates after a few occurrences, which is fine for simple single field short documents, such as published new stories, for which it was originally developed, but may not be so for more complex "structured" documents. The rate at which the saturation point is reached is controlled by the k_1 factor, and this needs special consideration for such documents.

The simple linear summation of scores across multiple fields breaks the nonlinear $tf(i,j)$ relation. For example, for a query term in a document with metadata ASR $tf(i,j[2]) = 2$ and $tf(i,j[1]) = 1$. For a standard unstructured document these will be combined to give an overall $tf(i,j) = 3$ in a single BM25 combined weight for this term i in document j. If we weight the metadata $v[f] = 2$ and the ASR $v[f] = 1$. This should boost the weight of this term somewhat overall in the matching score of the document, but not in a simple linear fashion. The linear combination of scores in simple data fusion would give a rather higher value than this, equivalent to an effective $tf(i,j)$ contribution of $2 \times f(BM25_{metadata}(tf(i,j[1]) = 1) + f(BM25_{ASR}(tf(i,j[2]) = 2)$, i.e. almost double the expected BM25 $tf(i,j)$ function value for a single field document. This would mean that a document matching a single query term over several fields could score much higher than a document matching several terms in one field only.

4.2 Developing a Solution

If all the field weights v_f are set to 1, it is reasonable that the document and retrieval result should revert to the unstructured case (equivalent to merging all the fields). However, this is not the case with a non-linear tf function with linear summation of the field scores, i.e.

$$ms(j,q,J) \neq \sum_f ms(j[f],q,J)$$

Instead, we get a score that is very hard to interpret and no longer satisfies the properties of the original ranking function. In this case, setting weights becomes a hard problem.

BM25 requires the two parameters k_1 and b to be tuned for each collection to which it is applied. k_1 controls the nonlinear $tf(i,j)$ effect, b the effect of length normalization. The simple linear sum of scores method requires separate parameters to be set for each field. The values of a field weight vector $\mathbf{v}$ would also have to be set empirically, $K-1$, since one field can be set to 1. Thus for BM25 the total number of tuning parameters to be set is $2K + (K-1) = 3K - 1$.

The method proposed in [8] is based on weighting term frequency combination at indexing time. In doing this it seeks to modify standard ranking functions to exploit multiple weighted fields, while satisfying the following requirements:

- **preserve term frequency non-linearity** which has been shown repeatedly to improve retrieval performance.
- **give a simple interpretation** to collection statistics and to document length incorporating field weights.
- **revert to the unstructured case** when field weights are set to 1.

The method combines the term frequencies of the different fields by forming a linear combination weighted by the corresponding field weights,

$$\mathbf{j}' = \sum_{f=1}^{K} v_f . j[f]$$

and $\mathbf{J}'$ is a new collection of documents. Note that $\mathbf{j}'$ and $\mathbf{J}'$ are both dependent on the values in the field weight vector $\mathbf{v}$.

Documents are then scored using the resulting term frequencies,

$$ms_2(\mathbf{j}, q, \mathbf{J}, \mathbf{v}) = ms(\mathbf{j}', q, \mathbf{J}')$$

In this scenario the term weighting and scoring functions are applied only once to each document.

From the earlier example, combining the term frequencies and field weights would give $2 + 2 \times 1 = 4$, resulting in a slight boost to the weight of the term in each field, while term dependence is maintained. The resulting boost is sufficiently small that matching several terms remains more significant than matching the same individual term in several fields. This is equivalent to mapping the structured document collection into a new unstructured collection with modified term frequencies.

Although developed for BM25, this method is generally applicable for different ranking functions for non-structured documents. However, the benefits of using it may vary for different functions.

A few issues of interpretation need to be considered in the case of the extended multi-field BM25 model.

Document Length. There are various different ways of counting the document length. The simplest is to count the number of words in the document, considering only those words that are indexed. Thus the length of the document is the sum of the term frequencies. This definition applies naturally to the modified documents of J': the modified term frequencies are simply summed.

k_1 *and* b. Since the merging method substantially changes the $tf(i, j)$ values, it can also be expected to change the optimal value of k_1. [8] proposes a method for estimatiing k_1 and b based on values derived empirically for an unweighted merged collection. However, in experiments we found this approach to be unreliable and instead set them empirically for the each modified weighted collection itself.

5. CLEF CL-SR TEST SET

This section summaries the design and features of the CLEF CL-SR test collections, further detail is contained in the original track report [9]. The collection is based on digitized interviews with Holocaust survivors, witnesses and rescuers made by the Survivors of the Shoah Visual History Foundation (VHF). A very large collection (116,000 hours) of interviews was collected. One 10,000 hour subset of this collection was extensively annotated. A project funded by the U.S. National Science Foundation focused on Multilingual Access to Large Spoken Archives (MALACH) has produced ASR systems for this collection to foster research on access to spontaneous conversational speech [2].

5.1 Document Test Set and Related Metedata

The objective of a ranked retrieval system is to sort a set of "documents" in decreasing order likelihood of relevance. This makes the implicit assumption that clearly defined document boundaries exist. The nature of oral history interviews means that document boundaries are less clearly defined. The average VHF interview lasts more than 2 hours. It is not realistic to browse spoken units of this size spoken. Therefore it is more useful to retrieve relevant passages rather than entire interviews. The annotated 10,000 hour subset of the VHF collection is provided manually segmented by subject matter experts into topically coherent segments. Segments from these recordings were selected as the "documents" for the CLEF 2005 and CLEF 2006 CL-SR evaluations.

The document set used for the CLEF evaluations was selected as follows. Roughly 10% of the dataset, comprising 403 interviews (totaling roughly 1,000 hours of English speech) were selected. Of these interviews, portions of 272 were digitized and processed by two ASR systems for the CLEF 2005 CL-SR test collection. A total of 183 of these are complete interviews; for the other 89 interviews ASR results were available for at least one, but not all, of the 30-minute tapes on which the interviews were originally recorded. Finally, some further sections involving brief discussion of visual objects were eliminated from the collection. The resulting test collection comprised 8,104 segments from 272 interviews totaling 589 hours of speech. Thus each segment ("document") has an average duration of about 4 minutes (503 words) of recognized speech. A collection of this size is very small from the perspective of contemporary text information retrieval experiments, such as those as TREC, but is comparable to the 550 hour broadcast news collection used in the TREC 8 and TREC 9 SDR evaluations [3]. For the retrieval evalation each segment was uniquely identified by a DOCNO based on the recording from which it was taken.

For each segment a number of fields, including the ASR transcriptions, were created by VHF subject matter experts while viewing the interviews. The following fields were included in the test collection:

- NAME: contains the names of persons other than the interviewee that are mentioned in the segment.
- MANUALKEYWORDS: The MKW field contains thesaurus descriptors selected manually from a large thesaurus that was constructed by VHF. Two types of keywords are present, but not distinguished: (1) keywords that express a subject or concept; and (2) keywords that express a location, often combined with time in one pre-coordinated keyword. On average about 5 manually thesaurus descriptors were manually assigned to each segment, at least one of which was typically a pre-coordinated location-time pair (usually with one-year granularity)
- SUMMARY: contains a three-sentence summary in which a subject matter expert used free text in a structured style to address the following questions: who? what? when? where?

The following fields were generated fully automatically by systems that did not have access to the manually assigned metadata for any interview in the test collection. These fields can therefore be used to explore the potential of different techniques for automated processing:

- ASRTEXT fields contain words produced by an ASR system. The speech was automatically transcribed by ASR systems developed at the IBM T. J. Watson Research Center. For CLEF 2005, two ASR transcriptions were generated. The ASRTEXT2004A field contains a transcription using the best available ASR system, for which an overall mean word error rate (WER) of 38% and a mean named entity error (NEER) rate of 32% was computed over portions of 15 held-out interviews. The recognizer vocabulary for this system was primed on an interview-specific basis with person names, locations, organization names and country names mentioned in an extensive pre-interview questionnaire. The ASRTEXT2003A field contains a transcription generated using an earlier system for which a mean WER of 40% and a mean NEER of 66% was computed using the same held-out data. The ASRTEXT2006A ASR field was created for CLEF 2006 with mean word error rate of 25%. This was not available for all segments, where the ASRTEXT2004A field was inserted instead to form the ASRTEXT2006B field, further details are contained in [6].

- Two AUTOKEYWORD fields contain thesaurus descriptors, automatically assigned by using text classification techniques. The AUTOKEYWORD2004A1 (AKW1) field contains a set of thesaurus keywords that were assigned automatically using a k-Nearest Neighbor (kNN) classifier using only words from the ASRTEXT2004A field of the segment; the top 20 keywords are included. The classifier was trained using data (manually assigned thesaurus keywords and manually written segment summaries) from segments that are not contained in the CL-SR test collection. The AUTOKEYWORD2004A2 (AKW2) field contains a set of thesaurus keywords that were assigned in a manner similar to those in the AKW1, but using a different kNN classifier that was trained (fairly) on different data; the top 16 concept keywords and the top 4 location-time pairs (i.e., the place names mentioned and associated dates) were included for each segment.

5.2 Topics and Relevance Assessment

For the CLEF 2005 CL-SR task, a total of 75 requests felt to be representative of the form and subjects real search requests were selected from those created by users of the VHF collection. These were formed into standard TREC style topic statements consisting of a title, a short description and a narrative. Only topics for which relevant segments exist can be used as a basis for comparing the effectiveness of ranked retrieval systems. The developers sought to choose a set of topics and interviews for which the number of relevant segments was likely to be sufficient to yield reasonably stable estimates of mean average precision (30 relevant segments was chosen as the target, but considerable variation was allowed). A total of 12 topics were excluded, 6 because the number of relevant documents turned out to be too small to permit stable estimates of mean average precision (fewer than 5) or so large (over 50% of the total number of judgments) that the exhaustiveness of the search-guided assessment process used was open to question. The remaining 6 topics were excluded because relevance judgments were not ready in time for release as training topics and they were not needed to complete the set of 25 evaluation topics. The 63 topics developed in CLEF 2005 were thus available as a training set for CLEF 2006. 30 additional topics were created for the CLEF 2006 task. These were combined with 12 topics developed in 2005, but for which relevance data was not released, to form a test topic set of 42 topics. Following analysis of the results of participants submission 33 topics from the 42 topic released as the test set were selected as the 2006 evaluation set. Full details of the topics and relevance assessment procedures adopted are given in [9] and [6].

6. EXPERIMENTAL INVESTIGATION

In this section we give experimental retrieval results for the individual metadata fields of the CLEF CL-SR task and give some analysis of these results, and then report results for experiments combining ASR transcriptions and metadata fields. The basis of our experimental system is the City University research distribution version of the Okapi system [7]. The documents and search topics are processed to remove stopwords from a standard list of about 260 words, suffix stripped using the Okapi implementation of Porter stemming and terms are indexed using a small standard set of synonyms. None of the indexing procedures were adapted for the CLEF CL-SR test collections. All experiments are for the 63 English language training topics from CLEF 2006 using the combined TD topic fields[3]. $k1$ and b were tuned empirically for each experiment. Standard Okapi pseudo relevance feedback (PRF) [7] is used in all cases with an empirically determined upweighting of the original topic terms in each case. Results here thus represent an upper bound on expected performance for this system. The following metrics are shown: Recall in terms of total number of relevant documents retrieved for topics, standard TREC mean average precision (MAP), and precision at rank cutoffs of 5, 10 and 30.

6.1 Individual Field Retrieval Runs

Table 1: Retrieval results for individual document fields with CLEF 2005 CL-SR test topics.

	Recall	MAP	P5	P10	P30
MKW	2274	0.225	0.444	0.381	0.296
Summary	2157	0.234	0.422	0.384	0.285
ASR2006B	1488	0.071	0.215	0.200	0.131
AKW1	1451	0.047	0.149	0.138	0.106
AKW2	625	0.039	0.102	0.094	0.064

Table 1 shows retrieval results for individual fields[4]. Looking at these results for individual fields we can observe a number of interesting points. The good result for the Summary field is perhaps not surprising since these descriptions are constructed manually by domain experts. However, the result for MKW is only slightly lower. Our indexed MKW fields had an average of about 22 terms, similar to the number of terms in each of the AKW fields. This indicates that if a set of keywords related to the specific contents of a document can be assigned, then useful retrieval performance can

[3]The CLEF 2006 test set was not used since it is in use as test data in CLEF 2007

[4]No result is shown for the Name field since it is empty for many documents

be achieved without the need for extensive manual descriptions. Retrieval performance based on the ASR and AKW1 and AKW2 fields is much lower. Without access to full accurate transcriptions of the speech and AKWs assigned based on such transcriptions, it is not clear to what extent poor retrieval performance is due only to errors in the ASR transcriptions, and consequentially the assigned keywords. Or the extent to which the failure of important words to be articulated in the speech at all, means that even with perfect transcription relevant documents cannot be reliably retrieved at high ranks. However, even without this information we can perform some interesting analysis of spoken transcriptions in relation to indexing and search.

Table 2: Term occurrence statistics for TREC 8 and TREC 9 SDR Text and Speech collections.

	Text	Speech
No. of Unique Terms	78611	23316
Terms $n(i) = 1$	46626	4444
Terms $n(i) > 1$	31985	18872

One interesting feature to consider is the coverage of the vocabulary appearing in spoken documents against the vocabulary of the ASR system. There is no ground truth of the contents of CLEF CL-SR collections. However, we performed an analysis of the vocabulary of the spoken document collection used for the TREC-8 SDR task [3]. This data set comprises around 22,000 broadcast news documents. A baseline ASR transcription is provided along with a rough manual transcription of the data. The results of this analysis are shown in Table 2. It can be seen that the total number of unique terms appearing in the ASR transcription is about one third of those appearing in a manual transcription, while the vocabulary of the manual transcriptions is somewhat inflated by the presence of typos which will not be present in the ASR transcription, there is a clear trend. While the frequency of many of these additional terms in the manual transcription is very low, the data associated with speech segments must be mapped to within-vocabulary words, and evidence suggests that this set of words is drawn from a subset of the recognition vocabulary of the ASR system. This means that the frequency of recognised words will be higher in the ASR transcriptions than in accurate transcriptions. While the OOV rate is overall probably less than 10% for the ASR system in this news domain, a great many rare words are missing from the transcription, either because they are outside the vocabulary, or because the ASR system is "reluctant" to use them, possibly because of problems in statistical estimation in the language model associated with rare words in the training set. Whatever the reason for this, their absence from the transcription means they are not available for search.

The BYBLOS recognition system used to generate this transcription is quite well suited to the data to be recognised. Given the training of the ASR used to generate the CL-SR transcriptions described in Section 5.1, while the TREC SDR corpus is read rather than spontaneous speech, a similar trend is likely to occur for the ASR output of spontaneous speech in terms of vocabulary coverage in terms of vocabulary coverage.

Table 3: Average $cfw(i)$ and topic coverage values for CLEF 2006 CL-SR Title field. Total no of non-stopword terms = 74.

Field	Mean	Std Dev	Terms Present
MKW	5.24	1.76	48
Summary	6.47	1.89	66
ASR2006B	5.35	1.57	66
AKW1	4.17	2.02	47
AKW2	4.23	2.64	39

Table 4: Results for combination of ASR2006B with various metadata fields.

		Recall	MAP	P5	P10	P30
+AKW1	Unwgt	1584	0.077	0.248	0.219	0.144
	Wgt	1641	0.086	0.254	0.237	0.156
+AKW2	Unwgt	1665	0.086	0.238	0.210	0.149
	Wgt	1663	0.088	0.244	0.211	0.140
+AKW1 +AKW2	Unwgt	1717	0.092	0.241	0.233	0.157
	Wgt	1778	0.097	0.264	0.221	0.164
+MKW	Unwgt	2129	0.225	0.417	0.370	0.273
	Wgt	2334	0.255	0.432	0.419	0.313
+SUMM	Unwgt	2166	0.213	0.415	0.363	0.270
	Wgt	2252	0.242	0.454	0.405	0.292

Given that we expect many rare search terms will be missing from the ASR transcriptions, we might expect that the terms which do appear will have lower discriminative ability, i.e. lower than expected $cfw(i)$ values. Table 3 shows mean and standard deviation cfw(i) values for the document fields calculated for non-stopwords in the Title field of the CLEF 2006 CL-SR topics with non-zero $cfw(i)$ values. It can be seen that mean $cfw(i)$ values for the ASR fields are lower than for the Summary field, both of which have the same coverage of the search terms. The keyword fields have significant numbers of topic search terms missing. We can again see that the mean $cfw(i)$ value for manual MKW fields is higher than those for the AKW fields. While the differences in $cfw(i)$ values may not appear large, these actually correspond to very large variations in the numbers of document in which a search term appears. Thus we can see that manual fields have greater discrimination than the automatically generated ones.

6.2 Field Combination Experiments

We now report results for merging of the ASR transcription field with metadata fields as described in Section 4. Table 4 shows combination of the automatically generated ASR2006B field with AKW1, AKW2, AKW1 and AKW2, MKW and Summary fields. Two combination schemes are compared in this experiment: simple merging of the fields and weighted field merging using BM25F. The field weights for the weighted runs and BM25 parameters were based on training using the CLEF 2005 data sets. Fields weights are based on the relative average precisions for the individual fields on the training set. Weights were set as follows: ASR2006B *times* 2, AKW1 *times* 1, AKW2 *times* 1, MKW *times* 4, and Summary *times* 4. A number of observations

can be made about these results. Use of BM25F by weighting the fields improves retrieval performance with respect to all metrics in nearly all cases. Comparing with the results for the individual fields in Table 1 it can be seen that the weighted combination results are in all cases better than those of any one of the individual component fields. Similar comparison reveals simple merge of ASR2006A with either the MKW or Summary field reduces performance compared to the individual manual fields, while simple combination of the automated fields still produces an improvement in effectiveness compared to individual fields, albeit a small one than with the weighted combination. The improvement in effectiveness for MKW and Summary when using weighted combination with ASR2006B is interesting since it indicates that while the transcription is noisy, it is still able to contribute useful information does not appear in the manual fields. These results should be treated with some caution since the parameters have been optimised for the individual runs on the test collection. We will be conducting further evaluations of field combination scenarios as part of our participation in the CLEF 2007 CL-SR track, and it will be interesting to see whether the trends retrieval effectiveness observed in this paper in are preserved for a set of search topics for which the algorithms have not been tuned.

7. CONCLUSIONS

Searching spontaneous conversation speech is a challenging problem raising more significant research challenges than earlier work on retrieval from read speech news collections. This paper has explored some of these problems, and examined the potential utility of related metadata to spoken content to enhance search effectiveness. We then examined the issue of field combination in multi-field documents. Experimental results using the CLEF CL-SR data sets illustrate that combination of ASR transcripions with metadata fields can enhance retrieval effectiveness. Further work is required in examining data combination for search, if performance on unseen search topics is to be made reliable. Examination of $cfw(i)$ values for automatically and manually fields show that automatic fields have lower term specificity indicating that this is one of the reasons for poor document ranking using these features.

Overall the results indicate that spontaneous speech search can benefit from the use of high quality metadata. Generating manual metadata is time consuming and expensive, although as demonstrated in our experiments it can be much more effective that automatically generated material. A research challenge then is to improve the quality of automatically generated metadata. In some domains, such as education, useful metadata is often easily available and relatively simple to associate with spoken content, in other domains, automatically locating and assigning precise metadata to associate with spoken segments for search will prove very challenging.

8. ACKNOWLEDGEMENT

Work partially supported by European Community under the Information Society Technologies (IST) programme of the 6th FP for RTD - project MultiMATCH contract IST-033104. The authors are solely responsible for the content of this paper. It does not represent the opinion of the European Community, and the European Community is not responsible for any use that might be made of data appearing therein.

9. REFERENCES

[1] E. Brown, S. Srinivasan, A. Coden, D. Ponceleon, J. W. Cooper, and A. Amir. Towards speech as a knowledge resource. *IBM Systems Journal*, 40(4):985–1001, 2001.

[2] W. Byrne, D. Doermann, M. Franz, S. Gustman, J. Hajic, D. Oard, M. Picheny, J. Psutka, B. Ramabhadran, D. Soergel, T. Ward, and W.-J. Zhu. Automatic recognition of spontaneous speech for access to multilingual oral history archives. *IEEE Transactions on Speech and Audio Processing, Special Issue on Spontaneous Speech Processing*, 12(4):420–435, 2004.

[3] J. Garofolo, G. Auzanne, and E. Voorhees. The trec spoken document retrieval track: A success story. In *RIAO 2000 Conference: Content-Based Multimedia Information Access*, pages 1–20, Paris, France, 2000.

[4] G. J. F. Jones and R. J. Edens. Automated alignment and annotation of audio-visual presentations. In *ECDL '02: Proceedings of the 6th European Conference on Research and Advanced Technology for Digital Libraries*, pages 276–291, London, UK, 2002. Springer-Verlag.

[5] G. J. F. Jones, J. T. Foote, K. S. Jones, and S. J. Young. Retrieving spoken documents by combining multiple index sources. In *19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 96)*, pages 30–38, Zurich, Switzerland, 1996.

[6] D. W. Oard, J. Wang, G. J. F. Jones, R. W. White, P. Pecina, D. Soergel, X. Huang, and I. Shafran. Overview of the clef-2006 cross-language speech retrieval track. In *CLEF 2006: Workshop on Cross-Language Information Retrieval and Evaluation*, Alicante, Spain, 2007.

[7] S. E. Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu, and M. Gatford. Okapi at TREC-3. In *Third Text REtrieval Conference (TREC-3)*, pages 109–126.

[8] S. E. Robertson, H. Zaragoza, and M. Taylor. Simple BM25 extension to multiple weighted fields. In *13th ACM International Conference on Information and Knowledge Management*, pages 42–49, Washington D.C., U.S.A., 2004.

[9] R. W. White, D. W. Oard, G. J. F. Jones, D. Soergel, and X.Huang. Clef 2005 cross-language speech retrieval track overview. In *CLEF 2005:Workshop on Cross-Language Information Retrieval and Evaluation*, pages 744–759, Vienna, Austria, 2006.

An Analysis of Sentence Segmentation Features for Broadcast News, Broadcast Conversations, and Meetings

Sebastien Cuendet[1]
cuendet@icsi.berkeley.edu

Elizabeth Shriberg[1,2]
ees@speech.sri.com

Benoit Favre[1]
favre@icsi.berkeley.edu

James Fung[1]
jgf@icsi.berkeley.edu

Dilek Hakkani-Tur[1]
dilek@icsi.berkeley.edu

[1]ICSI
1947 Center Street
Berkeley, CA 94704, USA

[2]SRI International
333 Ravenswood Ave
Menlo Park, CA 94025, USA

ABSTRACT

Information retrieval techniques for speech are based on those developed for text, and thus expect structured data as input. An essential task is to add sentence boundary information to the otherwise unannotated stream of words output by automatic speech recognition systems. We analyze sentence segmentation performance as a function of feature types and transcription (manual versus automatic) for news speech, meetings, and a new corpus of broadcast conversations. Results show that: (1) overall, features for broadcast news transfer well to meetings and broadcast conversations; (2) pitch and energy features perform similarly across corpora, whereas other features (duration, pause, turn-based, and lexical) show differences; (3) the effect of speech recognition errors is remarkably stable over features types and corpora, with the exception of lexical features for meetings, and (4) broadcast conversations, a new type of data for speech technology, behave more like news speech than like meetings for this task. Implications for modeling of different speaking styles in speech segmentation are discussed.

General Terms

Prosodic Modeling, Sentence Segmentation

Keywords

Spoken Language Processing, Sentence Segmentation, Broadcast Conversations, Spontaneous Speech, Prosody, Word Boundary Classification, Boosting.

1. INTRODUCTION

We investigate the role of identically-defined lexical and prosodic features when applied to the same task across three different speaking styles—broadcast news (BN), broadcast conversations (BC), and face-to-face multi-party meetings (MRDA). We focus on the task of automatic sentence segmentation, or finding boundaries of sentence units in the otherwise unannotated (devoid of punctuation, capitalization, or formatting) stream of words output by a speech recognizer.

Sentence segmentation is of particular importance for speech understanding applications, because techniques aimed at semantic processing of speech input—such as machine translation, question answering, information extraction—are typically developed for text-based applications. They thus assume the presence of overt sentence boundaries in their input [9, 7, 3]. In addition, many speech processing tasks show improved performance when sentence boundaries are provided. For instance, speech summarization performance improves when sentence boundary information is provided, as observed in [2]. Similarly, named entity extraction and part-of-speech tagging in speech is improved using sentence boundary cues in [4], and the use of sentence boundaries for machine translation is shown to be beneficial for machine translation in [8] . Sentence boundary annotation is also important for aiding human readability of the output of automatic speech recognition systems [5], and could be used for determining semantically and prosodically coherent boundaries for playback of speech to users in tasks involving audio search.

While sentence segmentation of broadcast news, and to some extent of meetings, has been studied in previous work, little is known about broadcast conversations. Indeed, data for this task has only recently become available for work in speech technology. Studying the properties of broadcast conversations and comparing them with those of meetings and broadcast news is of interest both theoretically, and also practically, especially because there is currently less data available for broadcast conversations than for the other two types studied here. For example, if two speaking styles share characteristics, one can perform adaptation from one to another to improve the performance of the sentence segmentation, as proved previously for meetings by using conversational telephone speech [1].

The goal of this study is to analyze how different sets of features, including lexical features, prosodic features, and

their combination, perform on the task of automatic sentence segmentation for different speaking styles. More specifically we ask the following questions:

1. How do different feature types perform for the different speaking styles?
2. What is the effect of speech recognition errors on performance, and how does this effect depend on the feature types or on the speaking style?
3. For this task, are broadcast conversations more like broadcasts or more like conversations?

Results have implications not only for the task of sentence boundary detection, but more generally for prosodic modeling for natural language understanding across genres.

The next section describes the data set, features, and approach to sentence segmentation. Section 3 reports on experiments with prosodic and lexical features, and provides further analysis and a discussion of usage of various feature types (or groups) and comparison across speaking styles. A summary and conclusions are provided in Section 4.

2. METHOD

2.1 Data and annotations

To study the differences between the meetings, BN and BC speech for the task of sentence segmentation, we use the ICSI Meetings (MRDA) [12], the TDT4 English Broadcast News [15], and the GALE Y1Q4 Broadcast Conversations corpora.

The ICSI Meeting Corpus is a collection of 75 meetings, including simultaneous multi-channel audio recordings, word-level orthographic transcriptions. The meetings range in length from 17 to 103 minutes, but generally run just under an hour each, summing to 72 hours. We use a 73 meeting subset of this corpus that was also used in the previous research [12] with the same split into training, held-out and test sets. TDT4 Corpus was collected by LDC and includes multilingual raw material, news wires and other electronic text, web audio, broadcast radio and television. We use a subset of TDT4 English broadcast radio and television data in this study. The GALE Y1Q4 Broadcast Conversations Corpus, also collected by LDC, is a set of 47 in-studio talk shows with two or more participants, including formal one-on-one interviews and debates with more participants. Three shows last for half an hour and the rest of the shows run an hour each, for a total of about 45 hours.

In the experiments to follow, classification models are trained on a set of data, tuned on a held-out set, and tested on an unseen test set, within each genre. The corpora are available with the words transcribed by humans (reference) and with the words output by the speech recognizer (STT). For the reference conditions, word start and end times are obtained by using a flexible alignment procedure [14]. Reference boundaries in speech recognizer output and flexible alignments are obtained by aligning these with manual transcriptions with annotations. Statistics on these data sets are shown in Table 1 for the STT conditions.

Note that the three different speaking styles differ significantly in mean sentence length, with sentences in meetings being only about half the length on average as those in broadcast news. Meetings (and conversational speech in general) tend to contain syntactically simpler sentences and significant pronominalization. News speech is typically read from a transcript, and more closely resembles written text. It contains for example appositions, center embeddings, and proper noun compounds, among other characteristics, that contribute to longer sentences. Discourse phenomena also obviously differ across corpora, with meetings containing more turn exchanges, incomplete sentences, and higher rates of short backchannels (such as "yeah" and "uhhuh") than speech in news broadcasts and in the broadcast conversations.

	MRDA	TDT4	BC
Training set size	456,486	800,000	270,856
Test set size	87,576	82,644	40,598
Held-out set size	98,433	81,788	37,817
Vocabulary size	11,894	21,004	12,502
Mean sentence length	7.7	14.7	12.6

Table 1: Data set statistics. Values are given in number of words, based on the output of the speech recognizer (STT).

Sentence boundary locations are based on reference transcriptions for all three corpora. Sentences boundaries are annotated in BN transcripts directly. For the meeting data, boundaries are obtained by mapping dialog act boundaries to sentence boundaries. The meetings data are labeled according to 5 classes of dialog acts: backchannels, floor-grabbers and floor-holders, questions, statements, and incompletes. In order to be able to compare the three corpora, all dialog act classes are mapped to the sentence boundary class. The BC Corpus frequently lacked sentence boundary annotations, but included line breaks and capitalization as well as dialog act tag annotations. In order to use this data, we implemented heuristic rules based on human analysis to produce punctuation annotations. For BC data, we similarly mapped the 4 types of dialog acts defined (statements, questions, backchannels, and incompletes) to sentence boundaries. Note that we have chosen to map incomplete sentence boundaries to the boundary class, even though they are not "full" boundaries. This is because the rate of incompletes, while not negligible, was too low to allow for adequate training of a third class in the BC data given the size of the currently available data. We thus chose to group it with the boundary class, even though incompletes also share some characteristics with non-boundaries. (Namely, material to the left of an incomplete resembles non-boundaries, whereas material to the right resembles boundaries).

2.2 Automatic speech recognition

Automatic speech recognition results for the ICSI Meetings data, the TDT4 data and the BC data were obtained using the state-of-the-art SRI conversational speech recognition system [17], BN system [16], and BC system [14], respectively. The meetings recognizer was trained using no acoustic data or transcripts from the analyzed meetings corpus. The word error rate for the recognizer output of the complete meetings corpus is 38.2%. Recognition scores for the TDT4 corpus is not easily definable as only closed captions are available that frequently do not match well with the actual words of the broadcast news shows. The estimated word error rate lies between 17% and 19%. The word error rate for the recognizer output of the BC data is 16.8%.

2.3 Features

Sentence segmentation can be seen as a binary classification problem, in which every word boundary has to be labeled as a sentence boundary or as a non-sentence boundary[1]. We define a large set of lexical and prosodic features, computed automatically based on the output of a speech recognizer.

Lexical features.

Previous work on sentence segmentation in broadcast news speech and in telephone conversations has used lexical and prosodic information [13, 6]. Additional work has studied the contribution of syntactic information [10]. Lexical features are usually represented as N-grams of words. In this work, lexical information is represented by 5 N-gram features for each word boundary: 3 unigrams, 2 bigrams and 1 trigram. Naming the word preceding the word boundary of interest as the *current* word, and the preceding and following words as the *previous* and *next* word respectively, the 5 lexical features are as follows:

- unigrams: {previous}, {current}, {next},
- bigrams: {current, next},
- trigram: {previous, current, next}.

Prosodic Features.

Prosodic information is represented using mainly continuous values. We use 68 prosodic features, defined for and extracted from the regions around each inter-word boundary. Features include pause duration at the boundary, normalized phone durations of the word preceding the boundary, and a variety of speaker-normalized pitch features and energy features preceding, following, and across the boundary. Features are based in part on those described in [13]. The extraction region around the boundary comprises either the words or time windows on either side of the boundary. Measures include the maximum, minimum, and mean of pitch and energy values from these word-based and time-based regions. Pitch features are normalized by speaker, using a method to estimate a speaker's baseline pitch as described in [13]. Duration features, which measure the duration of the last vowel and the last rhyme in the word before the word boundary of interest, are normalized by statistics on the relevant phones in the training data. We also include "turn" features based on speaker changes.

2.4 Boosting Classifiers

For classification of word boundaries, we use the AdaBoost algorithm [11]. Boosting aims to combine weak base classifiers to come up with a strong classifier. The learning algorithm is iterative. In each iteration, a different distribution or weighting over the training examples is used to give more emphasis to examples that are often misclassified by the preceding weak classifiers. For this approach, we use the BoosTexter tool described in [11]. BoosTexter handles both discrete and continuous features, which allows for a convenient incorporation of the prosodic features described above (no binning is needed). The weak learners are one-level decision trees (stumps).

[1]More detailed models may distinguish questions from statements, or complete from incomplete sentences.

2.5 Metrics

The quality of a sentence segmentation is usually computed with F-measure and NIST error. The F-measure is the harmonic mean of the recall and precision measures of the sentence boundaries hypothesized by the classifier to the ones assigned by human labelers. The NIST error rate is the ratio of the number of wrong hypotheses made by the classifier to the number of reference sentence boundaries. In this work, we report only the F-Measure performances.

2.6 Chance performance computation

What is of interest in the following experiments is the performance gain obtained by the classifier towards the baseline performance that one would achieve without any knowledge about the data but the prior of the classes. The easiest way of doing so is to compute the prior probability $p_t(s)$ of having a sentence boundary on the training set, and classify each word boundary in the test set as a sentence boundary with probability $p_t(s)$. Concretely, the chance score is evaluated by computing the probability of each error and correct class (true positives, false positives, and false negatives) and the ensuing value for the F-Measure computation. The final chance performance only depends on the prior probabilities of having a sentence boundary on the training set and on the test set. Therefore, the chance performance can differ slightly on the reference and STT experiments, due to word insertion and deletion errors introduced by automatic speech recognition.

3. RESULTS AND DISCUSSION

This section discusses results for the three different corpora, in two subsections. The main section, Section 3.1, presents results for feature groups (e.g., pitch, energy, pause) and for combinations of the groups. Section 3.2 examines feature subgroups within the pitch and energy features (for example, pitch reset versus pitch range features), to gain further understanding of which features contribute most to performance.

3.1 Performance by feature group

Performance results for experiments using one or more feature types are summarized for reference in Table 2. To convey trends, they are plotted in Figure 1; lines connect points from the same data set for readability. The feature conditions on the X-axis are arranged in approximate order of performance for the best-performing condition, i.e. for MRDA using reference transcriptions.

Although chance performance is higher for MRDA than for the broadcast corpora, consistent with the shorter average sentence length in meetings, all corpora have low chance performance. While chance performance changes slightly for reference versus automatic transcriptions, they are close within a corpus. As a consequence, one can compare the F-Measure results almost directly across conditions. To simplify the discussion, we define δ the relative error reduction. Since the F-Measure is a harmonic mean of two error types, one can compute the relative error reduction for a model with F-Measure F and the associated chance performance c as:

$$\delta = \frac{(1-c)-(1-F)}{1-c} = \frac{F-c}{1-c} \quad (1)$$

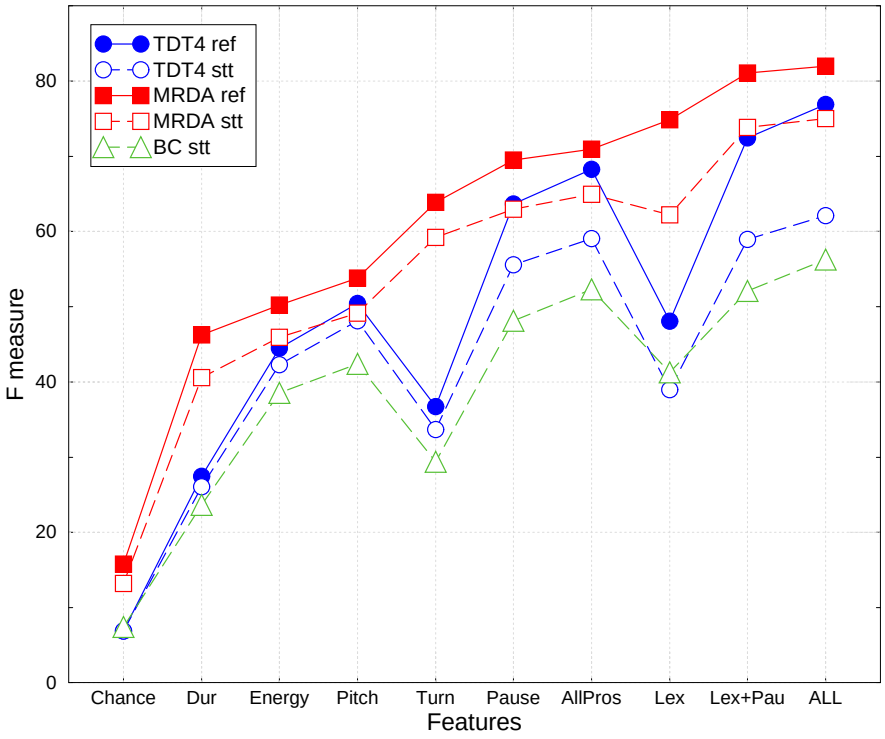

Figure 1: F-measure Results by Condition and Features Included in Model. Dur = duration features, AllPros = all prosodic feature types, Lex = lexical features, Lex+Pau = lexical feature plus pause features, ALL = lexical plus all prosodic features.

Features	TDT4 ref	TDT4 STT	MRDA ref	MRDA STT	BC STT
Chance	6.9	6.9	15.8	13.2	7.4
Duration	27.5	26.0	46.3	40.6	23.6
Energy	44.5	42.3	50.3	46.0	38.5
Pitch	50.5	48.1	53.8	49.2	42.4
Turn	36.7	33.7	63.9	59.2	29.4
Pause	63.7	55.6	69.5	63.0	48.1
All pros	68.3	59.1	71.0	65.0	52.3
Lex only	48.1	39.0	74.9	62.2	41.3
Lex+pause	72.4	59.0	81.1	73.9	52.1
ALL(lex+pros)	76.9	62.1	82.0	75.0	56.3

Table 2: Overall results: F-Measure scores for each group of features showed in the first column, for all combinations of corpus/conditions.

Since chance error rates are near 100%, the relative reduction in error after normalizing for chance performance is nearly the same value as the F-Measure itself. That is, an F-measure of 70 corresponds to a relative error reduction δ of about 70% for the data sets considered.

When comparing performance within a corpus (TDT4 or MRDA) for reference versus automatic transcripts, results show a remarkably consistent performance drop associated with ASR errors. This implies that all feature types are affected to about the same degree by ASR errors. The interesting, clear exception is the performance of lexical features for the meeting data, which degrades more in the face of ASR errors than do other conditions. For example, relative to the all-prosody features condition, lexical features in this corpus give better performance for reference transcripts, but worse performance for automatic transcripts. In contrast, TDT4 shows about the expected drop for lexical features from ASR features. One possible explanation is that in MRDA, there is a high rate of backchannel sentences (such as "uh-huh") which comprise a rather small set of words, sometimes with fairly low energy, that are more prone to recognition errors or that cause class errors when misrecognized. The same argument could be made for other frequent words in MRDA that are strong cues to sentence starts, such as "I" and various fillers and discourse markers. Further analysis, in which selected dialog acts such as backchannels are removed from the train and test data, could shed light on these hypotheses.

If we consider that the BC data set is much smaller than the other two sets, and thus the training material for the

classifier smaller, all three corpora are quite similar in performance in both energy and pitch features (although we will see in the next section that within these feature classes, there are some corpus differences). The corpora also share the trend that duration features are less useful than pitch or energy features, and that pause features are the most useful individual feature type. Interestingly, duration features alone are more useful in MRDA than in either of the broadcast corpora. A listening analysis using class probabilities of errors from the model revealed a possible explanation. In broadcast speech, speakers do lengthen phones before sentence ends, but they also lengthen phones considerably in other locations, including during the production of frequent prominences, and at the more frequent sub-sentential syntactic boundaries found in news speech. Both characteristics appear to lead to considerable false alarms on duration features in the broadcast corpora.

Another noticeable difference across corpora is visible for turn features. Here again, the meeting data differs from the broadcast data. This result reflects both the higher rate of turn changes in the meeting data, especially for short utterances such as backchannels, and the way that the data is processed. As already mentioned in Section 2, the turn is computed differently in the meetings than in the two other corpora. In the broadcast data, the turn is only estimated by an external diarization system that may introduce errors, whereas in the meetings the turn information is the true one since each speaker have their own channel. Furthermore, while turns in both broadcast and meetings data are broken by 0.5 second pauses, the meeting pauses are derived from the reference or STT transcript while the broadcast data pauses come from the less-sophisticated speech/non-speech preprocessor of the diarization system.

A final observation from Figure 1 concerns the patterns for the BC data. This is a newer corpus in the speech community and little is understood about whether it is more like broadcast news speech or more like conversational speech. The results here, both for chance performance and for performance across feature types, clearly indicate that in terms of sentence boundary cues, broadcast conversations are more like broadcast news, and less like conversations. The overall lower results for BC data are as noted earlier, likely explained simply by the smaller set of training data available. The one exception visible from Figure 1 in this trend is in the condition using lexical features only. We would expect the BC result here to be lower than that for TDT4 STT, given the overall lower performances for BC than TDT4. But instead we see a higher-than-expected result for BC in this condition, similar in trend to the pattern seen for MRDA STT for lexical features. We hypothesize that BC shares with conversational data the added utility of lexical features from either backchannels (that start and end sentences) or from words like fillers, discourse markers, and first person pronouns (that tend to start sentences). Further analyzes of the BC data suggest that while backchannels may not play a large role in broadcast conversations, the second class of words, i.e. those that tend to start sentences, are fairly frequent and thus probably aid the lexical model.

3.2 Performance by feature subgroup

A further feature analysis step is to look more closely at the two feature types that capture frame-level prosodic values, namely pitch features and energy features. These are also the two feature types that are normalized for the particular speaker (or speaker estimated via diarization) in our experiments. Our feature sets for each of these two feature types consisted of three subgroups.

Subgroup 1 uses an estimate of "baseline" pitch or energy, intended to capture the minimum value for the particular talker. Features in this subgroup reflect pitch or energy values in the word or short time window preceding the boundary in question, and compares those values to the estimated baseline value for the talker. The idea is to capture how high or low the pre-boundary speech is, relative to that speaker's range, with lower values correlating with sentence ends. We refer to these features as *range* features, since they capture the speaker's local value within their range. Note that because these features look at information prior to the boundary itself, they could be used for online processing, i.e. to predict boundary types before the following word is uttered.

Subgroup 2 looks *across* the boundary, i.e. at words or time windows both before and after the boundary in question. These features compare various pitch and energy statistics (for example maximum, mean, minimum) of the preceding and following speech, using various normalizations. The idea is to capture "resets" typical of sentence ends, in which the pitch or energy has become low before the end of a sentence, and is then reset to a higher value at the onset of the next sentence. Such features are defined only when both the preceding and following speech is present (within a time threshold for any pause at the boundary). We refer to these as *reset* features.

Subgroup 3, like subgroup 1, looks only at words or time windows at one side of the boundary. The idea is to capture the size of pitch or energy excursions by using the *slope* of regions close to the boundary. The slope is taken from linear fits of pitch and energy contours after various preprocessing techniques to remove outliers. Sentence ends in conventional linguistic studies of prosody are associated with a large "final fall", which these features are intended to capture. They may also however capture excursions related to prominent syllables at non-boundaries.

Results for the three feature types, for both energy and pitch, are shown in Figure 2. For ease of readability, lines connect points for the same condition. We look only at the three STT conditions, since reference results for TDT4 and MRDA show a similar pattern to their respective STT results, and using STT results allows us to compare all three corpora. To compare relative usage of the different feature subgroups directly, we look at relative error reduction results (see previous section), although as noted there the absolute F-measure results will look similar.

A first point to note about Figure 2, which can be construed by comparing to results in Figure 1, is that in all conditions, subgroups perform less well on their own than the all-energy and all-pitch groups. This indicates that the subgroups contribute some degree of complementary information. Second, across all conditions and across the two feature types, it is clear that the *reset* features perform better than features based on local *range* or *slope*. The interpretation is that it is better to use information from words or windows on both sides of a boundary in question than to look only at one side or the other. Third, pitch features are relatively more useful than energy features for all three corpora, but the largest differential is for the TDT4 data.

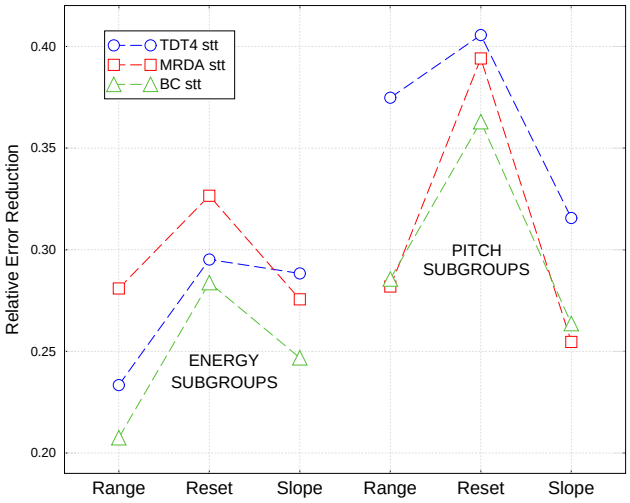

Figure 2: Relative error reduction for the three subgroups of features *range*, *reset* and *slope*, for both energy and pitch.

Note however that the strongest subgroup, i.e. pitch reset, shows nearly identical relative error reduction performance for both TDT4 and MRDA; BC is not far behind given the much smaller amount of data available for training.

Finally, these results show one example in which the BC data compares more closely with meeting data than with broadcast news data. TDT4, more so than the two conversational corpora, can make use of additional subtypes such as slope for energy features, and range for pitch features. Thus, although BC looks more like TDT4 than like MRDA when examining overall feature usage (see Figure 1), it shares with MRDA that certain feature subtypes, such as those based on looking at only one side of a boundary, are much less robust than the reset features that look across boundaries. This suggests that the conversational data may have greater range variation and less defined excursions than read news speech.

4. SUMMARY AND CONCLUSIONS

We have studied the performance of sentence segmentation across two spontaneous speaking styles—broadcast conversations and meetings— and a more formal one—broadcast news. The average length of sentences and comparison of the lexical and prosodic feature types performance showed that in terms of sentence boundary cues, broadcast conversations are more like broadcasts and less like meetings. However, the performance of the lexical features suggests that BC shares with meetings the added utility of lexical features from word like fillers, discourse markers, and first person pronouns. Other similarities between meetings and BC were also observed, such as the benefit of prosodic features that looking at characteristics of both sides (rather than only one side) of an inter-word boundary.

The three speaking styles showed similarities in the role of individual features. Pitch and energy features, as overall groups, perform surprisingly similarly in absolute terms for all three corpora. Also, for all corpora pause features are the most useful individual type of features, and duration features are less useful than energy and pitch. However, while the rank of the feature types was the same, the duration features were comparatively more useful in meetings than in the two other corpora, most likely because of the tendency of broadcast speakers to lengthen phones not only near sentence boundaries, but also in other locations.

A closer look at pitch and energy features in terms of feature subgroups revealed that subgroups provide complementary information, but some subgroups are clearly better than others. For all three corpora, there was greatest benefit from features that compare speech before and after inter-word boundaries. Broadcast news differed from the conversational corpora in being able to also take good advantage of features that look only at one side of the boundary, likely reflecting the more careful and regular prosodic patterns associated with read (as opposed to spontaneous) speech.

Comparisons of the reference and speech-to-text conditions showed, interestingly, that nearly all feature types are affected to about the same degree by ASR errors. The exception was lexical features in the case of the meetings, which degrade more than expected from ASR errors. Possible explanations for this are that sentence segmentation performance in meetings relies more heavily on certain one-word utterance like backchannels, as well as on a small class of highly predictive sentence onset words such as "I", fillers, and discourse markers.

In future work we plan to explore methods for improving performance on BC data, including adaptation and addition of similar data from other corpora. We also plan to study the impact of removing specific classes of dialog acts from the meetings, to determine the behavior of lexical features for this corpus, as just described above, is related to specific dialog acts or to some other phenomenon. Finally, we hope that further work along the lines of the studies described herein, can add to our longer term understanding of the relationship between speaking style and various techniques and features for natural language processing.

5. ACKNOWLEDGMENTS

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under contract No. HR0011-06-C-0023. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the view of DARPA. The authors would like to thank Matthias Zimmermann, Yang Liu, and Mathew Magimai Doss for their help and suggestions.

6. REFERENCES

[1] S. Cuendet, D. Hakkani-Tür, and G. Tur. Model adaptation for sentence unit segmentation from speech. In *Proceedings of SLT*, Aruba, 2006.

[2] S. Furui, T. Kikuchi, Y. Shinnaka, and C. Hori. Speech-to-text and speech-to-speech summarization of spontaneous speech. *Speech and Audio Processing, IEEE Transactions on*, 12(4):401–408, 2004.

[3] D. Hakkani-Tür and G. Tur. Statistical sentence extraction for information distillation. In *Proceedings of ICASSP*, Honolulu, HI, 2007.

[4] D. Hillard, Z. Huang, H. Ji, R. Grishman, D. Hakkani-Tur, M. Harper, M. Ostendorf, and W. Wang. Impact of automatic comma prediction on

pos/name tagging of speech. In *Spoken Language Technologies (SLT)*, 2006.

[5] D. Jones, W. Shen, E. Shriberg, A. Stolcke, T. Kamm, and D. Reynolds. Two experiments comparing reading with listening for human processing of conversational telephone speech. In *Proceedings of EUROSPEECH*, pages 1145–1148, 2005.

[6] Y. Liu, E. Shriberg, A. Stolcke, B. Peskin, J. Ang, D. Hillard, M. Ostendorf, M. Tomalin, P. Woodland, and M. Harper. Structural metadata research in the EARS program. In *Proceedings of ICASSP*, 2005.

[7] J. Makhoul, A. Baron, I. Bulyko, L. Nguyen, L. Ramshaw, D. Stallard, R. Schwartz, and B. Xiang. The effects of speech recognition and punctuation on information extraction performance. In *In Proc. of Interspeech*, Lisbon, 2005.

[8] E. Matusov, D. Hillard, M. Magimai-Doss, D. Hakkani-Tür, M. Ostendorf, and H. Ney. Improving speech translation with automatic boundary prediction. In *Proceedings of ICSLP*, Antwerp, Belgium, 2007.

[9] J. Mrozinski, E. W. D. Whittaker, P. Chatain, and S. Furui. Automatic sentence segmentation of speech for automatic summarization. In *Proc. ICASSP*, Philadelphia, PA, 2005.

[10] B. Roark, Y. Liu, M. Harper, R. Stewart, M. Lease, M. Snover, I. Shafran, B. Dorr, J. Hale, A. Krasnyanskaya, and L. Yung. Reranking for sentence boundary detection in conversational speech. In *Proceedings of ICASSP*, Toulouse, France, 2006.

[11] R. E. Schapire and Y. Singer. Boostexter: A boosting-based system for text categorization. *Machine Learning*, 39(2/3):135–168, 2000.

[12] E. Shriberg, R. Dhillon, S. Bhagat, J. Ang, and H. Carvey. The ICSI meeting recorder dialog act (MRDA) corpus. In *Proceedings of SigDial Workshop*, Boston, MA, 2004.

[13] E. Shriberg, A. Stolcke, D. Hakkani-Tür, and G. Tur. Prosody-based automatic segmentation of speech into sentences and topics. *Speech Communication*, 2000.

[14] A. Stolcke, B. Chen, H. Franco, V. R. R. Gadde, M. Graciarena, M.-Y. Hwang, K. Kirchhoff, A. Mandal, N. Morgan, X. Lin, T. Ng, M. Ostendorf, K. Sonmez, A. Venkataraman, D. Vergyri, W. Wang, J. Zheng, and Q. Zhu. Recent innovations in speech-to-text transcription at SRI-ICSI-UW. In *IEEE Trans. Audio, Speech and Language Processing*, volume 14, pages 1729 – 1744, 2006.

[15] S. Strassel and M. Glenn. Creating the annotated TDT-4 Y2003 evaluation corpus. In *TDT 2003 Evaluation Workshop, NIST*, 2003.

[16] A. Venkataraman, A. Stolcke, W. Wang, D. Vergyri, V. Gadde, and J. Zheng. SRIs 2004 broadcast news speech to text system. In *EARS Rich Transcription 2004 workshop*, Palisades, 2004.

[17] Q. Zhu, A. Stolcke, B. Chen, and N. Morgan. Using MLP features in SRIs conversational speech recognition system. In *Proceedings of INTERSPEECH*, pages 2141 – 2144, Lisbon, Portugal, 2005.

Results of the 2006 Spoken Term Detection Evaluation

Jonathan G. Fiscus, Jerome Ajot, John S. Garofolo
National Institute of Standards and Technology
Gaithersburg, Maryland, USA
{jfiscus,ajot,jgarofolo}@nist.gov

George Doddingtion
Orinda, CA, USA
george.doddington@comcast.net

ABSTRACT
This paper presents the pilot evaluation of Spoken Term Detection technologies, held during the latter part of 2006. Spoken Term Detection systems rapidly detect the presence of a *term*, which is a sequence of words consecutively spoken, in a large audio corpus of heterogeneous speech material. The paper describes the evaluation task posed to Spoken Term Detection systems, the evaluation methodologies, the Arabic, English and Mandarin evaluation corpora, and the results of the evaluation. Ten participants submitted systems for the evaluation.

Categories and Subject Descriptors
H.3.3 Information Search and Retrieval

General Terms
Measurement, Design, Theory.

Keywords
Speech Retrieval, Audio Indexing, Audio Mining, Multilingual, Speech Recognition

1. INTRODUCTION
Information processing has become a major activity in the world, and spoken communications is a major source of that information. This, coupled with growing computer-accessible volumes of audio data, has created an opportunity and a need for effective retrieval of information from archives of speech data. To support development of such technology, NIST created the Spoken Term Detection (STD) pilot evaluation initiative. This evaluation is structured as a collaborative research activity that is intended to foster technical progress in STD, with the goals of exploring promising new ideas in STD, developing advanced technology incorporating these ideas, measuring the performance of this technology, and establishing a community for the exchange of research results and technical insights. The evaluation supported experiments on three languages: Arabic (Modern Standard and Levantine), English, and Mandarin Chinese.

The evaluation task and evaluation infrastructure are documented in the STD 2006 Evaluation Plan which can be found on the NIST STD website [1]. Section 1 summarizes the evaluation plan which defines: the STD task and STD system architecture, the STD system output, the STD search terms, and the evaluation methodology. Section 2 covers the specifics of the STD 2006 evaluation including the test corpora and results.

1.1. STD Task and System Architecture
The goal of the STD evaluation task is to rapidly detect the presence of a *term* – a sequence of words consecutively spoken – in a large audio corpus of heterogeneous speech material. The effectiveness of a deployed STD system is a tradeoff between processing resource requirements and detection accuracy. The evaluation plan prescribes a generic system architecture (Figure 1) that systems must adhere to in order to participate in the evaluation. While NIST typically does not prescribe system-internal operations for its language technology evaluations, it was necessary to model two key application constraints so that the evaluation task was a good model of the intended application. First, search times for a given term must be small (within seconds). Therefore, systems must index the audio corpus before searching, rather than search the corpus directly for each search term. Second, the indexer does not have advance knowledge of the search terms and therefore cannot use that information during indexing. These imposed constraints effectively force system developers to address both the real-time challenge of pre indexing corpora without knowledge of the search terms and the challenge of rapidly returning search results.

A benefit of the prescribed architecture is to enable uniform operation resource measurements across systems, e.g., indexing speed, index size, search speed, etc.

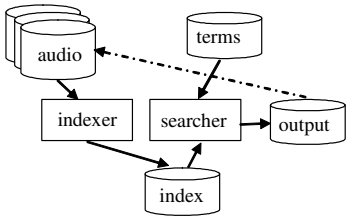

Figure 1: Generic STD System Architecture

Previous speech retrieval evaluations like TREC's Spoken Document Retrieval [7] (SDR), and Topic Detection and Tracking [8] (TDT) have investigated technologies similar to STD. However, they each addressed different problems. Source data robustness is a key component of STD whereas SDR and TDT focused on the broadcast news domain. The query for STD, a search term, is a markedly smaller unit than SDR's query definition which was a natural language description of an information need, and more specific than TDT's topic exemplar documents. A technology similar to STD is keyword spotting [10]. The main difference between keyword spotting and STD is the number of words in a search term.

1.2. STD Terms

STD terms are sequences of consecutively spoken words. They have no linguistically defined correlate, but range in grammatical scope from single words to phrases, e.g, /grasshopper/, /organizing/, /New York/, /Albert Einstein/, and /the coalition government/. STD terms are required to have a "recognizable, complete meaning" that a hypothetical user would want to find. For example, the trigram /crosby v. o./ is not a potential term because the /crosby/ is the name of a Voice Of America (/v. o. a./) reporter /Tom Crosby/ and therefore not complete. Further, it is not recognizable by itself. While this is a subjective definition, it models the information need of the searcher.

Native language orthography is the sole specification of a search term. Defining terms in this manner was a pragmatic decision. Ideally, each term would have a single specific interpretation or meaning. However, the contextual/phonetic definitions required to differentiate senses is beyond the term specification a hypothetical user will perform. Therefore, the term definitions for /wind/ (air movement) and /wind/ (twist) are indistinguishable. Systems must therefore handle pronunciation variations internally.

Terms include five or fewer "words." The concept of a "word" is not the same in all languages. In English, words include the morphological prefixes and suffixes in typical written text. Since articles, pronouns, and prepositions are separate words in English, they were not included. In Arabic, words are declared to be white space separated elements as typically used in Modern Standard Arabic (MSA). The Arabic terms included particles as part of the term since they are affixes and prefixes. For Mandarin, word segmentation was a product of the transcription process where the transcribers divided the character streams into word-like units.

Human annotators selected terms for the evaluation from a series of putative term lists derived from the evaluation corpus and from out-of-corpus sources. The in-corpus putative term lists included: tri-grams, bi-grams, uni-grams, and high frequency words. Bi-grams of all selected tri-grams and uni-grams of all bi-gram terms (including the bi-grams of the selected tri-grams) were added to the term lists so that constituent error rates for multi-word terms could be measured. Annotators added terms to the term lists that did not occur in the evaluation corpus. These out-of-corpus terms were used test the system's response to non-occurring terms.

Reference term occurrences are found automatically by searching high-quality transcripts. The following criteria were employed to determine the existence of a term; constituent words of a term must be adjacent, spoken by a single speaker, and within 0.5 second of each other. Sub-strings were not considered matches so an uttered word /grasshopper/ was not an occurrence of the term /grass/. Likewise, inflected forms were not considered matches so an uttered word /speaking/ was not an occurrence of the term /speak/. In a real applications, these forms could be sought simultaneously if that is what the user wishes.

1.3. STD System Inputs and Outputs

The ability of STD systems to process a variety of sources is an important factor of system performance, so the evaluation corpus contains as many sources as possible. STD systems index and search the complete test corpus with no *a priori* knowledge of the data. However, to make the first evaluation tractable for simple ports of existing technology, the audio files within the evaluation corpus included domain identifications, e.g., broadcast news (BNEWS), conversational telephone speech (CTS), or meeting room (MTG). Future STD evaluations will not provide this side information.

Systems process each term independently during the system's search phase. For each likely occurrence of a given term, the system is required to output a record that includes:

- the beginning and ending time of the term occurrence in the audio recording.
- a binary decision ("YES" or "NO") as to whether or not the system believes this putative occurrence is an occurrence of the term. This is called an "actual decision." Internal to the system, an actual decision threshold differentiates the YES/NO decisions[1].
- a detection score indicating how likely this putative term actually occurs (with more positive values indicating more likely occurrences.) The score for each term occurrence can be of any scale. However, the scores must be on a commensurate scale to permit the generation of pooled-term performance measurements.

Requiring systems to output both an actual decision and detection score for each putative term occurrence has a large benefit for system evaluation. Developers need a single metric to optimize system performance. However, *a priori* specification of an optimization criterion is dependent on the application: i.e. is high precision or high recall required. The actual decision provides the means to both optimize performance to a specific optimization criterion, via "YES" actual decisions, and over-generate putative occurrences, via "NO" actual decisions, to assess performance over a wide range of operating points. Section 1.4 covers this in more detail.

1.4. STD Evaluation Methodology

STD is a detection task – namely to detect all of the occurrences of each given term in the audio corpus. Two error types characterize STD performance: false alarms and missed detections.

Several NIST language evaluations have used the detection evaluation formalism, e.g., as in speaker recognition [1] [3]. Abstractly, detection systems answer the question: "Is this instance of data an example of the provided training data?" Each time the system answers this question, it is a called a "trial". The instance can be anything, a segment of speech for instance. The training data can be an exemplar of any form, a set of speech files for instance. Typically, the instances are discrete events or objects and therefore the trials are discrete. However, the STD task lacks the usual structure of discrete 'trials' necessary for computing normalized error rates, and therefore the evaluation methodology was adapted as follows.

[1] System performance is optimized by computing system performance based on the actual decisions.

- First, an estimate was required for the number of discrete trials in the reference. Unlike the speaker recognition evaluations, there are no discrete trials in continuous speech. Thus, part of the evaluation metric below specifies the number of trials as a constant.
- Second, an alignment between the system-detected occurrences and reference occurrences was needed in order to evaluate the system because systems are not given *a priori* knowledge of word/term boundaries in the speech. The Hungarian Solution to the Bipartite Graph [9] matching problem was used to compute the 1:1 mapping. The optimized objective function takes into account the temporal overlap of the system and reference occurrences (with a tolerance collar) and the term occurrence's detection score.
- Third, systems generate only a partial list of putative term occurrences[1] unlike speaker evaluations where systems provided decisions and scores for every trial.

System performance was evaluated using two methods: graphically with Detection Error Tradeoff (DET) curves [3] and for a particular operating point in the DET curve space using a Term-Weighted Value (TWV). The former provides an intuitive view of system performance for both high recall and high precision application needs, while the TWV provides developers with a single performance metric as a target for system optimization.

1.4.1. Detection Error Tradeoff Curves

Graphical performance assessment uses a detection error tradeoff (DET) curve that plots miss probability (PMiss) versus false alarm probability (PFA). Miss and false alarm probabilities are functions of the detection threshold, θ. This (θ) is applied to the system's detection scores, which are computed separately for each search term, then averaged to generate a DET line trace. The formulas for a single term's PMiss and PFA are:

$$P_{Miss}(term,\theta) = 1 - \frac{N_{correct}(term,\theta)}{N_{true}(term)}$$

$$P_{FA}(term,\theta) = \frac{N_{spurious}(term,\theta)}{N_{NT}(term)}$$

where:

$N_{correct}(term,\theta)$ is the number of correct (true) detections of *term* with a detection score greater than or equal to θ.

$N_{spurious}(term,\theta)$ is the number of spurious (incorrect) detections of term with a detection score greater than or equal to θ.

N_{true}(term) is the true number of occurrences of term in the corpus,

N_{NT}(term) is the number of opportunities for incorrect detection of term in the corpus (= "Non-Target" *term* trials).

Since there is no discrete specification of "trials", the number of Non-Target trials for a term, $N_{NT}(term)$, is defined somewhat arbitrarily to be proportional to the number of seconds of speech in the test set. Specifically:

$$N_{NT}(term) = n_{tps} \cdot T_{speech} - N_{true}(term)$$

where:

n_{tps} is the number of trials per second of speech (arbitrarily set to 1), and

T_{speech} is the total amount of speech in the test data (in seconds).

1.4.2. Term Weighted Value

To measure a system's "value" is to measure the usefulness of a system to a user. A perfect system always responds correctly to a stimulus, however an omitted response or a misleading response reduces the value of a system to a user. Thus, Term-Weighted Value (TWV) is one minus the average value lost by the system per term. The value lost by the system is a weighted linear combination of P_{Miss} and P_{FA} as defined above. The weight, β, takes into account both the prior probability of a term and the relative weights for each error type.

$$TWV(\theta) = 1 - \underset{term}{average}\{P_{Miss}(term,\theta) + \beta \cdot P_{FA}(term,\theta)\}$$

where:

$$\beta = \frac{C}{V} \cdot \left(Pr_{term}^{-1} - 1\right).$$

θ is the detection threshold.

For the current evaluation, the cost/value ratio, **C/V**, is 0.1, thus the value lost by a false alarm is a tenth of the value lost for a miss. The prior probability of a term, Pr_{term}, is 10^{-4}.

The maximum possible TWV is 1.0, corresponding to "perfect" system output: no misses and no false alarms. The TWV of a system that outputs nothing is 0.0 and negative TWVs are possible.

1.4.3. Actual vs. Maximum Term Weighted Value

While DET curves represent performance for all possible values of θ, two points on the DET curve are of interest because they determine if the system's actual decision threshold is optimal. The first is Actual Term-Weighted Value (ATWV) which is the TWV using the actual decisions. ATWV represents the system's ability to predict the optimal operating point given the TWV scoring metric. The second is Maximum Term-Weighted Value (MTWV). MTWV is the TWV at the point on the DET curve where a value of θ yields the maximum TWV. The difference between the values for ATWV and MTWV indicate the benefit of selecting a better actual decision threshold.

1.5. Processing Resource Measurements

Fielded STD technologies will process vast amounts of data. As such, "speed is important". Systems were required to record speed and resource measurements during processing. The

[1] The general application would also preclude generating exhaustive putative occurrences.

measurements allow both extrapolations to larger data sets and facilitate inter-system comparisons, i.e., comparing fast to slow systems would be unfair. The measurements are Index Size, Indexing Speed, Indexing Memory Usage, Search Speed, and Search Memory Usage.

Measurements such as these are often difficult to make during system execution when the processes are broken down into sub steps via UNIX shell scripts, (which the researchers predominately use.) To facilitate the measurements, NIST developed a new tool, ProcGraph [11], that tracks resource usage for UNIX shell scripts including subordinate processes.

2. STD 2006 EVALUATION

The 2006 evaluation was the first STD Evaluation. The process of designing the evaluation began in spring 2006. During the summer and fall of 2006, NIST assembled the evaluation infrastructure and developers built their systems. The evaluation occurred in November. NIST hosted the 2006 STD Evaluation workshop to discuss the results of the evaluation on December 14-15, 2006.

Ten sites participated in the evaluation: BBN Technologies (BBN), Brno Univ. of Tech. (BUT), Department of Defense (DOD), IBM, Institut Dalle Molle d'Intelligence Artificielle Perceptive (IDIAP), OGI School of Science and Tech. (OGI), Queensland Univ. of Tech. (QUT), SRI International (SRI), Stellenbosch Univ. (STELL), Technischen Universität Berlin (TUB). STELL and TUB collaborated to submit a system referred to a STBU.

The following sections provide summaries of the evaluation corpora, terms, and system performance measurements.

2.1. Evaluation Corpora

The evaluation made use of a small corpus of previously used Speech-To-Text evaluation test sets [4], [5], [6] which included high quality transcripts and automatically-derived time locations for each word. The word locations where computed with two methods. The first method, which was used for the English data, made use of Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur (LIMSI) speech recognition tools to align the reference transcript to the acoustic signal. (Forced word alignment is the common name for this process.) The second method, which was used for the Arabic and Mandarin data, inferred word locations from the output of an automatic speech recognizer (ASR) by finding a word alignment between the reference and ASR output words, then mapping the ASR word times onto the reference words. The time mapping procedure linearly interpolated times for reference words during regions of incorrectly recognized speech. As expected, they were not as accurate as the forced alignment-derived word times; however, the use of a temporal mapping tolerance collar reduces the impact of less accurate word times. Table 1 lists the data for each language and source type with the predominant dialect identified.

The Linguistic Data Consortium[1] transcribed all if the material according to high quality standards set by the speech recognition community. Appen[2] further enhanced the Arabic transcripts by correcting minor flaws and adding diacritics to the transcripts.

Table 1: STD 2006 evaluation corpus composition

	Arabic	Chinese	English
Broadcast News (BNEWS)	MSA ~1 hour	Mandarin ~1 hour	American~ 3 hours
Telephone Conversations (CTS)	Levantine ~1 hour	Mandarin ~1 hour	American ~3 hours
Roundtable Meetings (MTG)	None	None	American ~2 hours

2.2. Evaluation Terms

Nominally, 1100 terms were selected for each language with the following rough proportions: 10% tri-grams, 40% bi-grams, 50% uni-grams. For the Arabic data, the vast majority tri-grams were partial sentences and whole sentences. Since they were linguistically larger than phrases, Arabic tri-grams were not included in the term lists.

Table 2 shows the number of terms selected per language and the number of reference occurrences per source type. The terms selection protocol produced an English term list balanced by source type. However, the same is not true for Arabic and Mandarin. Subsequent evaluations will factor source type into the term selection protocol.

The evaluation used two forms of the Arabic terms, with and without diacritics – the former being posited as a means to better specify the terms thus accounting for dialectal variation. The diacritized terms were derived from the non-diacritized terms by a process that converted each term into a set of diacritized variants. The diacritized variants for each constituent word were limited to the variations found in the reference transcripts.

Table 2: Term Set Properties by Language

	Arabic		English	Mandarin
	Diacritized	Non-Diacritized		
Terms Selected	1101	937	1100	1120
Ref. Occ.	2433	2807	14421	3684
Reference Occurrences Per Source, Per Speech Hour				
BNEWS	1513	1749	2212	3070
CTS	557	638	1957	582
MTG			1750	

2.3. Arabic Results

BBN, BUT and DOD participated in the Arabic test. Table 3 summarizes their scores for both diacritized and non-diacritized terms. The highest ATWV for non-diacritized terms in the CTS domain was 0.34 by BBN. For the diacritized terms in the BNEWS domain, the highest ATWV was –0.06.

[1] See the LDC website www.ldc.upenn.edu

[2] See the Appen website www.appen.com.au

Although we wanted to test our hypothesis that diacritics would help searching, two difficulties emerged during the evaluation that prevented us from doing so. First, and foremost, diacritization is an inherently difficulty task for humans and therefore the reference transcripts contained diacritization errors. For example, Appen had two independent teams correct and diacritize 25 minutes of BNEWS and CTS data (50 minutes total). 12.5% of the BNEWS and 11% CTS words had at least one different diacritic after quality control passes. To put this in context, this is two-three times the error rate of human transcription of English. Second, building purely undiacritized systems is not possible because common Arabic transcription practices make use of diacritics to disambiguate word usage. Thus, the evaluation results between the two term sets are not directly comparable.

Table 3: Arabic Actual Term Weighted Values

Search Terms	Site	BNEWS	CTS
Non-diacritized	BBN		0.35
Diacritized	BUT	-0.09	0.00
	DOD		-6.57

2.4. English Results

All sites built systems for the English data. (BBN and DOD only built systems for the CTS portion of the test set.) Figure 2 presents the ATWVs for all the English tests by source type. The highest ATWVs were 0.85 for BBN's system on BNEWS data, 0.83 for BBN's system on CTS data, and 0.26 for SRI's system on MTG data. As expected, the order of difficulty by source type is BNEWS, CTS, MTG. This matches the source difficulty for speech recognition systems in the Rich Transcription evaluations.

Figure 4 contains the DET curves for all primary English systems on the CTS data. The graph shows the tradeoff between false alarms and missed detections. DET line traces for better performing systems, with regard to accuracy, have lines closer to the origin. The BBN system, which had the highest ATWV at 0.83, achieved a MTWV of 0.83 indicating a suitable actual decision threshold was chosen. At the MTW point, the false alarm rate was 0.005% and a missed detection rate was 11.9%. Note that no DET curve trace extends beyond 5% miss because the systems do not output a decision for every trial.

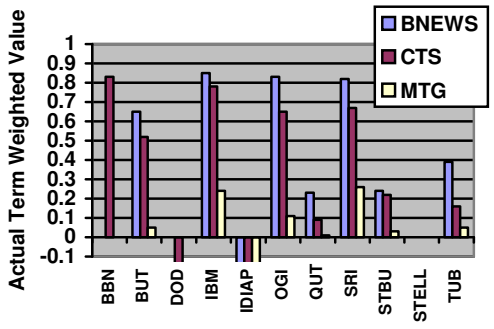

Figure 2: English Actual Term Weighted Values

The high ATWVs indicate developers strove to build accurate systems. For this initial evaluation though, most developers did not have the resources to build fast systems. Instead, developers used existing language technologies to build their STD systems. Figure 3 shows the performance of systems as a function of Indexing Speed measured in processing hours per indexed speech hours. On this graph, scores that appear in the upper left quadrant are better because they indicate accurate and fast STD systems. With the wide range of indexing speeds, it would be difficult to quantify the tradeoff with a single measurement that combines accuracy and speed into a single measure. Instead, next year's evaluation will likely require specific processing speed thresholds (e.g., 0.01, 0.1, and 1.0 Indexing Speeds) so that processing speed can be controlled while accuracy is measured.

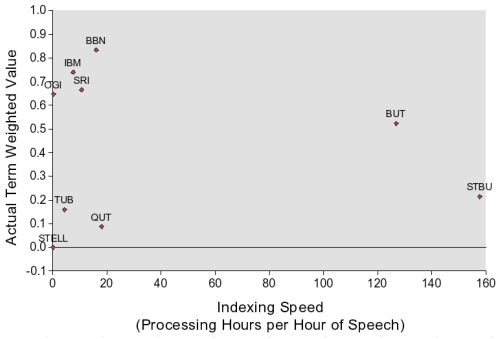

Figure 3: ATWV as a Function of Indexing Speed for the CTS data

2.5. Mandarin Results

BBN and DOD participated in the Mandarin tests and achieved scores of 0.38 and –1.02 ATWV respectively on the CTS data set. Neither participant processed the BNEWS data.

The evaluation infrastructure relied on human segmentation for both term selection and reference term location. This was acceptable for term selection because a human was in the loop. However, we are studying whether or not word segmentation negatively affected the scoring.

3. CONCLUSIONS

NIST conducted a pilot evaluation of Spoken Term Detection systems in December 2006. The evaluation was successful in that: it drew a significant number of participants (10) for a first such evaluation; the evaluation proved the feasibility of the STD technology measurement approach; it provided a useful baseline for future work; it touched on challenges with regard to technology robustness including speed, scalability, multilinguality, and domain independence. While the challenges of scalability and domain independence were not fully explored in the pilot, the evaluation set the stage for future efforts which explore these important dimensions in more depth.

The evaluation resulted in all ten of the participants having developed systems to process the English Conversational Telephone Speech subset of the test data. The highest ATWV for these systems was 0.83. The indexing speeds for these systems were extremely variable -- ranging from 0.168 to 157.6 processing-hours-per-hour-of-speech in the test corpus.

The most important advance to measurement science from this effort was the adaptation of the detection evaluation methodology to STD. In the course of creating the metric for this task, we developed a new approach which permitted us to measure detection accuracy when the events to be detected are not discrete trials.

Furthermore, the evaluation components developed to map system-to-reference term occurrences and build partial DET curves will be useful for a variety of other detection-oriented evaluations.

We intend to expand the scope of future STD evaluations to address the scalability and domain diversity issues and we will continue to study and refine the evaluation protocol with regard to: a term selection process that exercises the depth and breadth of the application domain in the most effective and informative manner, an assessment of the impact of transcription accuracy on performance measurements, develop metrics that combine accuracy and speed in informative and intuitive ways, improve the consistency of Arabic term diacritization, and assess the impact of Mandarin word segmentation. Toward this end, we expect to run a second STD evaluation in 2008 using a much larger and more diverse test set. The evaluation will challenge the technology in two dimensions: data robustness and processing speeds. The evaluation data will include a wider variety of data and processing speed will play a major role in the evaluation of systems.

4. DISCLAIMER

These tests are designed for local implementation by each participant. The reported results are not to be construed, or represented, as endorsements of any participant's system, or as official findings on the part of NIST or the U. S. Government. Certain commercial equipment, instruments, or materials are identified in this paper in order to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the materials or equipment identified are necessarily the best available for the purpose.

5. REFERENCES

[1] "2006 STD Website and Evaluation Plan", http://www.nist.gov/speech/tests/std/std2006/.

[2] Przybocki, M., Martin, A., "NIST Speaker Recognition Evaluation Chronicles", Proceedings of Odyssey 2004.

[3] Martin, A., Doddington, G., Kamm, T., Ordowski, M., Przybocki, M., "The DET Curve in Assessment of Detection Task Performance", Eurospeech 1997, pp 1895-1898.

[4] Fiscus et. al., "Results of the Fall 2004 STT and MDE Evaluation", RT-04F Evaluation Workshop Proceedings, November 7-10, 2004.

[5] Fiscus, J., et al., "The Rich Transcription 2005 Spring Recognition Meeting Evaluation", 2nd International Workshop on Machine Learning for Multimodal Interaction, LNCS 3869.

[6] Fiscus, J., Ajot, J., Michel, M., Garofolo, J., "The Rich Transcription 2006 Spring Meeting Recognition Evaluation", 3rd International Workshop on Machine Learning for Multimodal Interaction, LNCS 4299.

[7] Garofolo, J., Auzanne, C., Vorhees, E., "The TREC Spoken Document Retrieval Track : A Success Story", Proceedings of the Recherche d'Informations Assiste par Ordinateur: Content Based Multimedia Information Access Conference, April 12-14, 2000

[8] Allan, J., "Topic Detection and Tracking: Event-based Information Organization", ISBN 978-0792376644

[9] Harold W. Kuhn, "The Hungarian Method for the assignment problem", *Naval Research Logistic Quarterly*, **2**:83-97, 1955.

[10] Rose, R. C., Paul, D. B., "A Hidden Markov Model Based Keyword Recognition System", 1990 International Conference on Acoustics, Speech, and Signal Processing, 1990, pp. 129-132 vol.1.

[11] http://www.nist.gov/speech/tools/index.htm

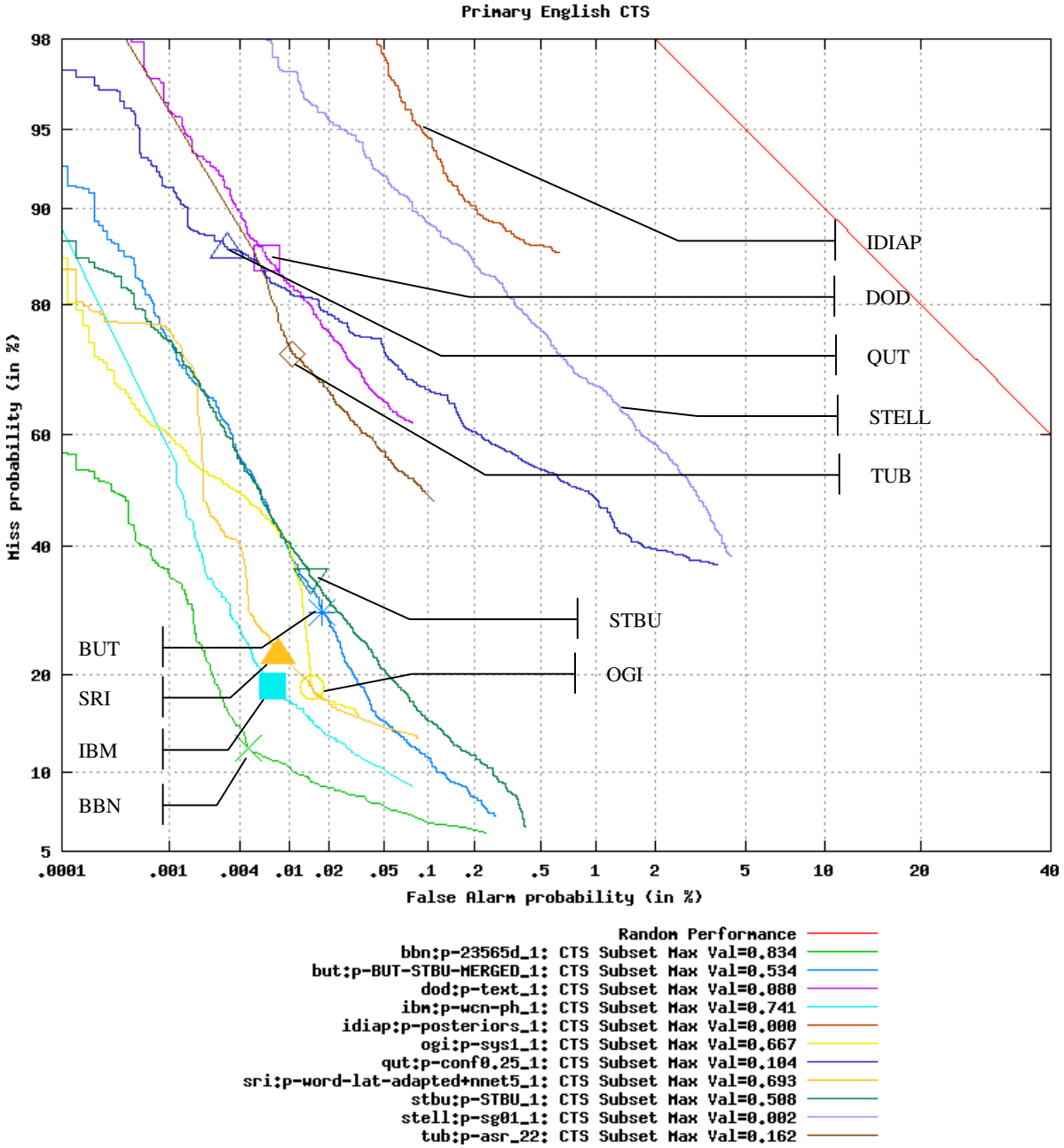

Figure 4: DET Curve for English, CTS Primary Systems. The symbols on the chart is the point of Maximum ATWV