Multimedia with a Speech Track:
Searching Spontaneous Conversational Speech

Martha Larson
Delft University of Technology, Netherlands
m.a.larson@tudelft.nl

Roeland Ordelman
Netherlands Institute for Sound & Vision
University of Twente, Netherlands
ordelman@ewi.utwente.nl

Franciska de Jong
University of Twente, Netherlands
f.m.g.dejong@ewi.utwente.nl

Joachim Kohler
Fraunhofer IAIS, Germany
joachim.koehler@iais.fraunhofer.de

Wessel Kraaij
Radboud University Nijmegen
TNO ICT, Netherlands
wessel.kraaij@tno.nl

Abstract

After two successful years at SIGIR in 2007 and 2008, the third workshop on *Searching Spontaneous Conversational Speech* (SSCS 2009) was held conjunction with the ACM Multimedia 2009. The goal of the SSCS series is to serve as a forum that brings together the disciplines that collaborate on spoken content retrieval, including information retrieval, speech recognition and multimedia analysis. Multimedia collections often contain a speech track, but in many cases it is ignored or not fully exploited for information retrieval. Currently, spoken content retrieval research is expanding beyond highly-conventionalized domains such as broadcast news in to domains involving speech that is produced spontaneously and in conversational settings. Such speech is characterized by wide variability of speaking styles, subject matter and recording conditions. The work presented at SSCS 2009 included techniques for searching meetings, interviews, telephone conversations, podcasts and spoken annotations. The work encompassed a large range of approaches including using subword units, exploiting dialogue structure, fusing retrieval models, modeling topics and integrating visual features. Taken in sum, the workshop demonstrated the high potential of new ideas emerging in the area of speech search and also reinforced the need for concentrated research devoted to the classic challenges of spoken content retrieval, many of which remain yet unsolved.

1 Introduction

In recent years, renewed effort has been dedicated to research and development in the area of providing access to spoken content. The growing amounts of multimedia material, improved speech recognition technology, increasing awareness on the part of the users and also increasing availability of computational power have all come together to create the appropriate conditions for a break through in the use of speech recognition technology to provide intelligent access to multimedia. The SSCS workshop series strives to support progress in the area of information retrieval techniques that make use of speech recognition transcripts. The first two workshops on *Searching Spontaneous Conversational Speech*, SSCS 2007 [1] and SSCS 2008 [4] were held in conjunction with SIGIR and brought together information retrieval experts with speech and audio researchers. In 2009, SSCS reached out the multimedia community by holding SSCS at ACM Multimedia 2009. The SSCS 2009 workshop took place on October 23, 2009 in Beijing, China.

Speech search is currently moving beyond the conventional domain of broadcast news, into areas in which speech is not pre-scripted, but instead is produced spontaneously and often in the context of a conversation or a natural communication setting. Such domains include: interviews (cultural heritage), lectures (education), meetings (business), debates (public life), consumer/professional internet media, especially podcasts (education, entertainment), telephone conversations and voice mail (enterprise) and spoken annotations, for example, for photo archives (personal media). These domains offer multiple challenges that spoken content search must address in order to offer users effective solutions. Spontaneous conversational speech is well known to be highly unpredictable. The variability arises from a range of sources including, individual speaker style, speaker accent, articulation, topic and also differences in channel conditions. Searching spontaneous conversational speech is made even more challenging by the fact that humans produce speech in an unstructured matter, meaning that the decision of where to place the boundaries of a document or a result is critical if a retrieval system is to be truly effective. Creating appropriate surrogates for time-continuous media like audio or video is also important. Good surrogates allow users to review results and chose items for further listening or viewing in a time-efficient manner. Finally, spoken audio contains a bounty of information that is encoded in structure, prosody and non-lexical audio. New methods must be developed in order to fully exploit these additional information sources.

Recent trends indicate that speech-based indexing is coming into widespread use. In addition to the projects mentioned in the report of SSCS 2008 [4], new developments in speech indexing have been observed on the Web. These include the introduction of web-based speech recognition, i.e., webASR[1] and also the automatic captioning service on YouTube[2]. The relatively recent rise in general-public awareness and use of systems such as Voxalead[3] suggests that speech-based search of news may have finally come into its own. The interest of the research community in spoken audio is evidenced by the success of systems like the Multimedia Grand Challenge 2009 winner Joke-o-mat, which uses audio analysis to enable punch-line based browsing in sitcoms [2]. Further, we would like to mention the strength of and persistence of benchmarks involving speech transcripts, not only TRECVID[4] evaluation,

[1] `http://www.webasr.com`

[2] `http://googleblog.blogspot.com/2009/11/automatic-captions-in-youtube.html`

[3] `http://voxaleadnews.labs.exalead.com`

[4] `http://www-nlpir.nist.gov/projects/trecvid`

but also the newer MediaEval[5] benchmark, which places explicit focus on the use of speech and language.

The SSCS series is motivated by an awareness that advances in the area of speech-based retrieval are made possible by sustained research effort requiring the interaction between communities of researchers in information retrieval, speech recognition, audio analysis, and multimedia. SSCS aims to provide a reliable yearly forum for such interaction and by doing so support the move forward into the next generation of technology that provides access to multimedia by exploiting the speech track.

2 Workshop Summary

SSCS 2009 was held as an afternoon workshop and consisted of a series of presentations, each followed by questions, and a demo session in which a range of speech search technologies were demonstrated. The participants (ca. 17 in total) were a highly international group, drawn mainly from academia but also from industry. The SSCS 2009 proceedings were published as part of the ACM Multimedia proceedings, cf. [5].

2.1 Keynote

The workshop keynote, entitled "Multimedia Retrieval through Indexing Speech" was delivered by Frank Seide of Microsoft Research Asia, and treated the topic of using speech transcripts for multimedia retrieval in the enterprise setting. Institutional memory is often encoded in the form of the spoken word, including audio or video recordings of presentations, lectures meetings, conference calls and voicemail. The basic issues are further detailed by the extended abstract of the keynote [7] included in the proceedings. In particular, speech indexing technology for the enterprise needs to be both cost-effective and easy to deploy, requiring not specialized skills. A single platform solution is more appropriate that individual solutions. Speech indexing should be used alongside of other indexing techniques and the goal should be to have the resulting search functionality well integrated into the existing information retrieval infrastructure of the enterprise. The keynote covered the use of word lattices and word lattice approximations for probabilistic retrieval capable of addressing spoken content for which speech recognition accuracy is limited and also techniques for tackling the out-of-vocabulary problem, including phonetic search and vocabulary adaptation using parallel information sources.

2.2 Presentations

The workshop is designed with a strong practical and user orientation, and the main body of the workshop began with two presentations of real world systems who submitted demonstration papers to the workshop. Then, a series of four oral presentations treating different aspects of searching spontaneous conversational speech followed. This section provides a short summary of all presentations.

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

PodCastle: A Spoken Document Retrieval System for Podcasts and its Performance Improvement by Anonymous User Contributions (Jun Ogata and Masataka Goto, National Institute of Advanced Industrial Science and Technology, Japan) The PodCastle system[6] uses automatic speech recognition to index Japanese language podcasts and offer users full text search in the transcripts. One of the most interesting aspects of PodCastle is its use of collaborative training for speech recognition. This process involves anonymous users correcting recognition errors in speech transcripts that have been disclosed online by the system. The corrected transcripts are used to train podcast-dependent acoustic models, which improve the speech recognition accuracy. The system has been online since 2006 and users have been found to correct speech transcripts voluntarily. Online text from online news sources is used to keep the language model up to date.

A Latent Semantic Retrieval and Clustering System for Personal Photos with Sparse Speech Annotation (Yi-Sheng Fu, Winston H. Hsu, Lin-Shan Lee, National Taiwan University) This system provides a user-friendly approach to retrieval and clustering of photos with speech annotations. It implements a probabilistic latent semantic indexing technique that is suitable for sparse annotations. The technique fuses speech features with low-level visual features.

Topic Modeling for Spoken Document Retrieval using Word- and Syllable-level Information (Shih-Hsiang Lin and Berlin Chen, National Taiwan Normal University) This paper presents an investigation of the use of subword indexing units with different types of topic models for retrieval of spoken Mandarin. Topic modeling approaches are shown to outperform unigram language modeling approaches. Another, more surprising, result of the experiments is that topic modeling approaches that use combinations outperform those that only use only word-level information.

Exploring Fusion in a Spontaneous Speech Retrieval Task (Muath Alzghool and Diana Inkpen, University of Ottawa, Canada) Late fusion of search results returned by a range of retrieval models (weighting schemes) is explored. In particular, this paper introduces a technique that reduces the number of weighting schemes that must be combined in order to exploit the benefits of late fusion.

Locating Case Discussion Segments in Recorded Medical Team Meetings (Saturnino Luz, Trinity College Dublin, Ireland) Real world data collected from weekly meetings of health care professionals is analyzed by exploiting the patterns of conversational exchange. The approach presented in this paper is particularly interesting because it is "content free", it does not make use of the words spoken. Rather, the duration and sequence of vocalizations is exploited in order to segment meetings into meaningful sub-units, namely, individual patient case discussions.

The Effect of Language Models on Phonetic Decoding for Spoken Term Detection (Roy Wallace, Brendan Baker, Robbie Vogt and Sridha Sridharan, Queensland University of Technology, Australia) The paper explores the relationship between phone recognition and spoken term detection. The experimental results that are presented support the conclusion that language models may improve phone recognition accuracy without making a positive contribution to spoken term detection. These findings suggest that it is important to make a wise choice of the metric used to measure the quality of an indexing lattice and in particular to make sure that it is suited for the particular task.

[6] `http://podcastle.jp`

2.3 Demo session

As in past years, SSCS was designed to promote interaction and active debate between participants. In order to provide concrete issues to seed discussion, a demo session was held that gave participants an opportunity to interact with the systems. The demo system included "official" workshop demos that had undergone the review process and had been introduced with presentations at the beginning of the workshop, namely PodCastle (cf. *Ogata at al.* summarized in Section 2.2) and the personal photo system (cf. *Fu et al.* summarized in Section 2.2). Further, participants were encouraged to bring their own demos in order to stimulate comparison and feedback. Two demos were shown in this category, the popular Radio Oranje [3], which uses alignment to give speech-based access to a collection of historical speeches, and also a system for multimodal reranking of speech retrieval results [6].

3 Conclusions

During the panel discussion that was held the previous year (i.e., at SSCS 2008), panelists formulated statements concerning key issues in spoken content retrieval. These issues found reflection in the papers presented at SSCS 2009. In particular, the SSCS 2008 panel observed that for spoken content retrieval, *the problems remain the problems*. Here, we would particularly like to point to the out-of-vocabulary (OOV) problem, i.e., the mismatch between the finite vocabulary of the speech recognizer and the large number of highly unpredictable lexical forms produced in spontaneous conversational speech. The OOV problem can only be solved by concerted effort dedicated to vocabulary adaptation (cf. the keynote [7] summarized in Section 2.1) and also spoken term detection (cf. *Wallace et al.* summarized in Section 2.2). Another key statement from the 2008 panel, was that *users give rise to the research problems*. Here, we would like to point out the papers concerning the segmentation of meeting data (cf. the paper of Saturnino Luz), which was developed to solve a well-defined real-world problem. The demos that treated podcast retrieval (cf. *Ogata at al.*) and retrieval using spoken annotation of photos (cf. *Fu et al.*) also directly addressed real-world use scenarios.

In closing, we would like to highlight a point that demonstrates how far speech retrieval has yet to progress until it lives up to the expectations of researchers and of users: The SSCS 2009 call for papers for the workshop included several forward-looking topics formulated on the basis of the SSCS 2008 panel discussion that failed to attract submissions. These topics included cross-modal concept detection, user interfaces, intelligent players and cross-media linking. Currently, a fourth workshop on *Searching Spontaneous Conversational Speech* is in planning and we hope that by bringing repeated attention to the existence and importance of such topics we can stimulate researchers to tackle these topics and present their work to the community.

References

[1] F.M.G. de Jong, D. Oard, R. Ordelman, and S. Raaijmakers. Searching spontaneous conversational speech. *SIGIR Forum*, 41(2):104–108, 2007.

[2] G. Friedland, L. Gottlieb, and A. Janin. Joke-o-mat: Browsing sitcoms punchline by punchline. In *MM '09: Proceedings of the 17th ACM international conference on Multimedia*, pages 1115–1116, New York, NY, USA, 2009. ACM.

[3] W. Heeren, L. van der Werff, R. Ordelman, A. van Hessen, and F. de Jong. Radio Oranje: Searching the Queen's speech(es). In *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 903–903, New York, NY, USA, 2007. ACM.

[4] J. Kohler, M. Larson, F.M.G. de Jong, W. Kraaij, and R.J.F. Ordelman. Spoken content retrieval: Searching spontaneous conversational speech. *SIGIR Forum*, 42(2):66–75, 2008.

[5] M. Larson, R.J.F. Ordelman, F.M.G. de Jong, W. Kraaij, and J. Kohler. Searching multimedia content with a spontaneous conversational speech track. In *MM '09: Proceedings of the 17th ACM international conference on Multimedia*, pages 1159–1160, New York, NY, USA, 2009. ACM.

[6] S. Rudinac, M. Larson, and A. Hanjalic. Exploiting visual reranking to improve pseudo-relevance feedback for spoken-content-based video retrieval. In *Image Analysis for Multimedia Interactive Services, 2009. WIAMIS '09. 10th Workshop on*, pages 17–20, 2009.

[7] F. Seide, K. Thambiratnam, L. Lu, and R. P. Yu. Multimedia retrieval through indexing speech: An enterprise perspective. In *SSCS '09: Proceedings of the third workshop on Searching spontaneous conversational speech*, pages 1–2, New York, NY, USA, 2009. ACM.

4 Acknowledgements

Sources of support for the SSCS 2009 workshop and the authors of this report included: PetaMedia (IST-FP7-216444) and AMIDA (IST-FP6-033812).