Spoken Content Retrieval:
Searching Spontaneous Conversational Speech

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

Martha Larson
University of Amsterdam, Netherlands
m.a.larson@uva.nl

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

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

Roeland Ordelman
University of Twente, Netherlands
ordelman@ewi.utwente.nl

Abstract

The second workshop on *Searching Spontaneous Conversational Speech* (SSCS 2008) was held in Singapore on July 24, 2008 in conjunction with the 31st Annual International ACM SIGIR Conference. The goal of the workshop was to bring the speech community and the information retrieval community together. The forum was designed to be conducive to the close interaction and the intense discussion necessary to promote fusion of these fields into a single discipline with a concerted vision of spoken content retrieval. At the workshop, talks and posters were presented covering a wide range of topics including vocabulary independent search, spoken term detection, combination of models/indexes, use of speech recognition lattices for search, segmentation, temporal analysis, benchmarking, exploitation of prosody, speech surrogates for user interfaces and multi-language collections. Demonstrations of speech-based retrieval systems from a variety of application domains introduced a strong practical emphasis into the workshop program. The workshop concluded with a panel discussion, whose goal it was to identify future research directions for speech retrieval. Among the important challenges identified during the panel discussions were: dealing with large scale multimedia collections, representing audio/video content effectively in the user interface, focusing on perfecting the component technologies on which speech retrieval systems are based, and developing systems and approaches that will enable users (both content seekers and content providers) to actively create their own speech search applications or contribute to the indexability of their content.

1 Introduction

Recent years have witnessed a speech retrieval renaissance, a renewed dedication of research and development effort to tackling the problem of intelligent access to spoken content collections. The resurgent interest in speech retrieval can be attributed to two trends. First, the need for intelligent indexing and retrieval of spoken content is becoming increasingly pressing as collections of digital audio and video content continue to grow. Second, speech recognition technology, together with audio event detection and other audio processing techniques, have undergone an extended period of development during which they have matured sufficiently to provide adequate basis for robust spoken content search systems.

The renewed impetus has carried the speech search discipline beyond the conventional domain of broadcast news content into areas in which spoken content is not scripted, but rather spontaneously produced by speakers interacting in natural communication settings. These domains include interviews (cultural heritage), lectures (education), meetings (business), debates (public life) and consumer/professional internet media, especially podcasts (education, entertainment). Multimedia content seekers have been primed to expect that robust speech-based search should, and will, be possible for any multimedia collection containing spoken content. These expectations encompass both very large collections, but also the myriad of smaller collections available world wide. Such smaller collections represent the challenging long-tail cases for spoken content access. Examples of such cases include collections using languages traditionally neglected in speech recognition research, collections that are challenging with respect to channel conditions or speaker style and collections in specialized subject domains.

In order to meet the scientific and technological challenges inherent in providing access to collections of unscripted spoken audio, tight cooperation is necessary between the information retrieval community and the speech recognition and audio processing community. The first workshop on *Searching Spontaneous Conversational Speech* (SSCS 2007)[1] held in conjunction with SIGIR 2007, brought together retrieval experts with speech and audio researchers. SSCS 2008 was motivated by the vision that, if well nurtured, this incipient forum can develop into a platform for the sustained interaction and intense discussion necessary to fuse speech research and retrieval research into a concerted discipline.

2 Context

The growing number of initiatives, projects and applications that set their central focus on speech- and audio-based search reflects the momentum that has been continuously rising in the field of spoken content retrieval. In order to provide a context for the SSCS 2008, we mention a selection of these initiatives here.

2.1 Projects

Table 1 lists a selection of projects devoted either in part or in whole to carrying our research and development in the area of spoken content retrieval. The heavy emphasis on unscripted content is reflected by the column containing a note on the project's focus. An important

[1] `http://hmi.ewi.utwente.nl/sscs`

contribution of projects is the development and dissemination of test sets. An example is the AMI/AMIDA meeting corpus.[2]

Project	Focus	Funding	URL
AMI/AMIDA	meetings	EU	http://www.amiproject.org
Boemie	sports video	EU	http://www.boemie.org
CHIL	meetings	EU	http://chil.server.de
CHoral	cultural heritage	Dutch	http://hmi.ewi.utwente.nl/choral
CHORUS	coordination action	EU	http://www.ist-chorus.org
DIVAS	video	EU	http://www.ist-divas.eu
IM2	meetings	Swiss	http://www.im2.ch
LIVE	sports video	EU	http://www.ist-live.org
MALACH	oral history	US	http://malach.umiacs.umd.edu
MESH	news	EU	http://www.mesh-ip.eu
MultiMatch	cultural heritage	EU	http://www.multimatch.eu
MultimediaN	multimedia	Dutch	http://www.multimedian.nl
Theseus	web	German	http://theseus-programm.de
SpeechFind	cultural heritage	US	http://speechfind.utdallas.edu
VIDI-Video	video	EU	http://www.vidi-video.it
VITALAS	video	EU	http://vitalas.ercim.org
Quaero	multimedia	EU	http://www.quaero.org

Table 1: A selection of projects concerned with spoken content retrieval

2.2 Benchmark Evaluations

Benchmark evaluation campaigns play a central role in driving the state of the art forward, since they provide data sets and standardized formulations of fundamental challenges that make it possible for research groups around the globe to synchronize their efforts. NIST evaluations such as Rich Transcription evaluation,[3] Spoken Term Detection evaluation,[4] and the Classification of Events Activities and Relationships (CLEAR) evaluation,[5] co-sponsored with CHIL make a substantial contribution to driving forward the state of the art. The TRECVid[6] benchmark focuses on retrieving visual content, but the audio/speech channel is increasingly exploited as a secondary source of features. At CLEF,[7] the Cross-Language Speech Retrieval track was succeeded this year by a pilot track, VideoCLEF,[8] devoted to processing dual language video content.

[2] http://corpus.amiproject.org
[3] http://www.nist.gov/speech/tests/rt
[4] http://www.nist.gov/speech/tests/std
[5] http://www.clear-evaluation.org
[6] http://www-nlpir.nist.gov/projects/trecvid
[7] http://www.clef-campaign.org
[8] http://ilps.science.uva.nl/Vid2RSS

2.3 Online Applications

The growing momentum of speech retrieval research and development is witnessed by the increasing number of applications using speech-based search techniques that are available on line. English language examples include PodScope,[9] EveryZing,[10] Blinkx,[11] and Pluggd.[12] Ten days before the SSCS 2008 workshop convened in Singapore, Google launched the Google Elections Video Search gadget,[13] which uses speech recognition to make it possible to search for words spoken by politicians in the speeches and other video material they upload to YouTube. During the 2004 US-elections, similar functionality was offered by StreamSage, now a division of Comcast, on the website campaignsearch.com. The current Comcast speech-based video search system[14] was presented at SSCS 2008.

2.4 Conferences and Workshops

Conferences and workshops are a key locus of dissemination and exchange of research results for the spoken content retrieval field. Conferences that have featured spoken content retrieval papers include Interspeech and SIGIR. Workshops that have been dedicated to the subject of speech content retrieval include the MLMI workshop.[15]

2.5 The SSCS Workshops

As mentioned above, the immediate precursor of SSCS 2008 was the first workshop on *Searching Spontaneous Conversational Speech* SSCS 2007. This workshop identified research issues projected to have future impact on the field of spoken content retrieval. These issues were expanded in a topic list that was used to solicit contributions for SSCS 2008.

- Representation of spoken content for optimal search (e.g., LVCSR, word lattice search, STD on phone lattice)
- Exploitation of evidence beyond word-level (e.g., emotional state, speaker characteristics, topic shifts, audio events)
- Application of text IR techniques to the speech domain
- Speech mining in multimedia data
- Multi-modality (integrating features from associated non-speech content)
- Search effectiveness (e.g., evidence combination, expansion)
- Access to large scale collections
- Evaluation resources and benchmarking activities
- Multi-/cross-lingual retrieval
- Cross-media mining (e.g., coupling images or text fragments to speech)

[9] http://www.podscope.com
[10] http://search.everyzing.com
[11] http://www.blinkx.com
[12] http://www.pluggd.tv
[13] http://googleblog.blogspot.com/2008/07/in-their-own-words-political-videos.html
[14] http://www.comcast.net
[15] http://www.mlmi.info

- Interaction design and system development (e.g., query formulation, result presentation, search strategies)
- Spoken audio visualization (e.g. results lists, individual results)
- Spoken query search

Fifteen experts from industry and academia were invited to serve on the program committee; papers were accepted for the workshop on the basis of their reviews. The papers accepted at the workshop reflected the ability of the Singapore locale to bring together researchers who are otherwise geographically widely flung. The presentations represented a mixture of groups traditionally working in the speech area moving towards retrieval issues and groups traditionally concentrated on text or web retrieval moving towards working with spoken content. On the program, both industrial and academic groups were represented. About 30 participants took part in the workshop, also drawn from a broad variety of backgrounds. This report summarizes the invited presentations, the oral presentations, the poster presentation and finally the panel session. The complete proceedings of SSCS 2008 [2] is available online.[16]

3 Invited Presentations

SSCS 2008 opened with a keynote entitle *Query-by-Example Spoken Document Retrieval* by Haizhou Li of Institute for Infocomm Research (I^2R) Singapore. The talk gave an overview of the work being carried out at the Human Language Technology of (I^2R), particularly in the areas of rich transcription and of techniques to characterize spoken documents. Li also presented the Star Challenge multimedia search engine competition,[17] which has attracted intense international attention. Not only did the keynote make a strong connection between the topic of the workshop and the workshop's Singapore venue, it also provided an appropriate kick-off by raising a key theme: the ongoing shift of research focus from spoken term detection to spoken content retrieval. The talk provided examples of the solutions necessary to face the challenges of voluminous real time speech streams and multilingual data. Anticipating the future path of speech search, Haizhou Li challenged the workshop participants with the injunction, "Don't rely on the words!"

SSCS 2008 also featured an invited talk given by Tony Davis of StreamSage, a division of Comcast Cable. The talk presented StreamSage's experience with speech-based search systems. Davis' talk made an important contribution to the workshop since it brought in discussion of the challenges accompanying commercial deployment of speech search, including scale and behavior trends in large groups of users. Davis discussed exploitation of linguistic approaches and integration of information from multiple sources including speech transcripts, closed captions and metadata. He reminded workshop participants that in order to solve the problem of large-scale access to multimedia content, "Speech is not enough!" Davis also stressed the importance of relevance intervals, segments of multimedia streams that are pertinent to a user's information need. Automatic determination of relevance intervals makes it possible for the system to present the user with useful points at which to listen in to the multimedia stream. Davis talk included a demonstration of Comcast video search technology, already mentioned in the introduction.

[16] `http://ilps.science.uva.nl/SSCS2008`
[17] `http://www.thestarchallenge.sg/`

4 Oral Presentations

The workshop included eight oral presentations, which are described here in turn. The oral presentations covered a range of issues, with particular attention being devoted to vocabulary independent search, spoken term detection, combination of models/indexes and exploitation of speech recognition lattices.

Cluster-based Model Fusion for Spontaneous Speech Retrieval *Muath Alzghool, Diana Inkpen (University of Ottawa, Canada):* Training topics (queries) in the collection are clustered. The best weighting scheme for combination of retrieval models is determined for each cluster. Test topics are classified into the topic cluster and retrieval is performed using the corresponding weighting scheme.

Using Term Clouds to Represent Segment-Level Semantic Content of Podcasts *Marguerite Fuller, Eamonn Newman, Gareth Jones (Dublin City University, Ireland) Manos Tsagias, Jana Besser, Martha Larson, Maarten de Rijke (University of Amsterdam, The Netherlands):* Structured surrogates, which prove useful for the semantic representation of spoken audio in the user interface, are created automatically. TextTiling techniques applied to speech transcripts divide the audio into topical segments and each segment is represented by a mini-term-cloud derived from the speech transcript.

Combination of Multiple Speech Transcription Methods for Vocabulary Independent Search *Jonathan Mamou, Yosi Mass, Benjamin Sznajder (IBM Haifa Research Lab, Israel) Bhuvana Ramabhadran (IBM T.J. Watson Research Lab, USA):* Two algorithms are presented that combine speech transcripts generated using different word and sub-word speech recognition methods. The approach tackles the challenge that out-of-vocabulary terms present in a spoken term detection task.

Fast Approximate Spoken Term Detection from Sequence of Phonemes *Joel Pinto, Hynek Hermansky (IDIAP Research Institute, Switzerland) Igor Szoke (Brno University of Technology, Czech Republic) S.r.m. Prasanna (IIT-Guwahati, India):* A phoneme recognition approach to spoken term detection is used to achieve a smaller index size and faster detection speed. Recognizer error is compensated with a probabilistic model based on word pronunciation and the recognizer's phoneme confusion matrix.

Towards Large Scale Vocabulary Independent Spoken Term Detection: Advances in the Fraunhofer IAIS Audiomining System *Daniel Schneider, Jochen Schon, Stefan Eickeler (Fraunhofer IAIS, Germany):* Performance improvements in vocabulary-independent spoken term detection are achieved by improved acoustic models and a hybrid word and syllable search system. In the improved system, it is no longer necessary to accommodate recognition error by allowing an inexact match between query word and transcript.

Hybrid Word-Subword Decoding for Spoken Term Detection *Igor Szoke, Michal Fapso, Lukas Burget, Jan Cernocky (Brno University of Technology, Czech Republic):* A hybrid recognition system directly produces lattices containing both words and subwords. Using multigram models and searching for in-vocabulary and out-of-vocabulary terms in separate steps makes possible performance gains on a spoken term detection task.

Audio stream segmentation and classification based on Jump Function Kolmogorov analysis in wavelet domain *Huy Dat Tran, Haizhou Li (I2R Singapore):* A novel method for analysis of audio content in noisy conditions is presented. The method is based on Jump Function Kolmogorov (JFK), which makes possible an advantageous separability of signal and noise.

Word-Lattice Based Spoken-Document Indexing with Standard Text Indexers *Roger Peng Yu, Kit Thambiratnam, and Frank Seide (Microsoft Research Asia):* The reliability scores output by the speech recognizer are an important source of information for speech search. A set of approximations makes it possible to transform the problem of matching lattices with numeric values into a symbolic text-based problem that can be handled by a commercial text indexer.

5 Poster Presentations

During the poster session, six posters were presented. The posters dealt with a variety of techniques, including segmentation, temporal analysis and exploitation of prosodic features. A variety of domains of application of speech retrieval were covered, including call center recordings, lectures, cultural heritage and dual language video.

From Audio Content Analysis to Conversational Speech Detection and Characterization *Benjamin Bigot, Isabelle Ferrane (IRIT Universite Paul Sabatier, France):* Zones in spoken audio containing conversational speech are automatically identified using temporal analysis to identify activity. The method can be used to determine the relative role of a particular speaker within the recording, and as such constitutes a bridge between low level audio features and higher level semantics.

Enhanced Search and Navigation on Conversational Speech *Frederik Cailliau (University of Paris, France; Sinequa Labs, France) Aude Giraudel (Sinequa Labs, France):* A search engine for full text search of speech recognition transcripts of call center calls is improved by incorporating text analysis in the form of a language model specially developed to handle disfluencies as well as an improved user interface.

Classification of Dual Language Audio-Visual Content: Introduction to the Video- CLEF 2008 Pilot Benchmark Evaluation Task *Martha Larson (University of Amsterdam, The Netherlands) Eamonn Newman, Gareth Jones (Dublin City University, Ireland):* The pilot task of the VideoCLEF track is introduced. The task involves classification, translation and semantic keyframe extraction performed on a collection of dual language video.

A Korean Spoken Document Retrieval System for Lecture Search *Donghyeon Lee, Gary Geunbae Lee (Pohang University of Science and Technology, Korea):* A search system for Korean lectures is introduced that exploits speech transcripts as well as additional information such as text from texts books and notes from slides.

Using Prosodic Features to Prioritize Voice Messages *Tim Polzehl, Florian Metze (Deutsche Telekom Laboratories, Germany):* A variety of features reflecting speech prosody, such as F_0 and loudness is used to classify voice messages left by callers at a call center into semantic categories.

Subword-based Indexing for a Minimal False Positive Rate *Laurens van der Werff, Willemijn Heeren (University of Twente, The Netherlands):* A method for subword-based speech retrieval that would exploit, in an intelligent manner, in-vocabulary words added to the query by the user.

6 Demonstrations

The demonstration session was included in the workshop to emphasize the importance of practical applications of speech search research in real world settings. Researchers from the Sinequa Labs presented a retrieval system for conversations recorded by call centers. Fraunhofer IAIS presented its search engine for radio content. Finally, the University of Twente presented a number of demos including work arising from the project MESH.[18] Examples of (spoken Dutch) demos available online are Radio Oranje[19] and Buchenwald.[20]

7 Panel Discussion and Future Research Directions

The workshop concluded with a panel discussion of future research directions in the field of spoken content retrieval. Five panelists were chosen, representing an international mix of industry and academic research groups: Tony Davis (Senior Researcher at StreamSage based in Washington D.C), Gareth Jones (Senior Lecturer in the School of Computing at Dublin City University), Franciska de Jong (Professor of language technology at the University of Twente in the Netherlands), Haizhou Li (Deparment Head of Human Language Technology at the Institute for Infocomm Research (I^2R) in Singapore) and Frank Seide (Lead Researcher Speech Technology at Microsoft Research Asia). During the course of the panel discussion, panel members contributed their opinions about the direction that speech retrieval research and development is moving and also formulated recommendations for future research priorities. The discussion crystalized around a number of key areas; the remainder of this section is devoted to discussing these areas in turn.

First, a significant portion of the discussion revolved around issues of operationalization of large scale, real world systems. For today's speech-based search applications, being able to deal with large amounts of audio data is important. Scale issues include both efficient indexing of content, fast filtering and acceptable response times. Also, systems must be able to exploit their own content in order to be self training or self improving. The importance of these issues is reflected in the fact that they were also central discussion themes at SSCS 2007 [1]. It is not possible to depend on content being neatly structured, as is the case

[18] `http://www.mesh-ip.eu`
[19] `http://hmi.ewi.utwente.nl/choral/radiooranje.html`
[20] `http://www.buchenwald.nl`

of broadcast news. Instead, a search system must perform necessary topical structuring automatically and also be able to offer users listen-in points, points to which to jump within the multimedia stream, which are relevant to their information needs. Speech retrieval research must continue to push beyond the broadcast news domain and dedicate serious and sustained energy to domains involving large scale collections of unscripted, ad hoc speech.

Second, panelists agreed on the *raison d'être* of the discipline of speech retrieval: "Users give rise to the research problems." Mobilizing users to take active part in designing, developing and training speech search systems is an important step in improving spoken content access, especially in the case of long-tail collections: collections that are set apart by their uniqueness, their small size, or the fact that they are interesting for a specialized niche of content seekers. The counter point to the increased involvement of users in defining the agenda of speech search research should, however, not be neglected. "Users can't imagine the technology before you show it to them." It is a considerable challenge to find the balance between solving the problems that content seekers need solved and developing technologies that will allow content seekers to discover creative new needs. The suggestion was made to bring the SSCS workshop closer to other related research communities in the future, especially those placing explicit emphasis on user information needs and human computer interaction.

Third, the panel observed that *the problems remain the problems.* In other words, the classic problems of speech retrieval have not yet been solved. In particular, out-of-vocabulary (OOV) words remains a challenge to systems indexing large amounts of heterogeneous audio. *The OOV problem has not yet been solved.* Also, the challenge of how to present audio in the user interface remains open, facing this challenge involves developing methods to present time-continuous media in a results list or an intelligent player. Another enduring challenge is acoustic model adaptation. As collection sizes continue to grow and increasingly more data are available, acoustic model adaptation methods must exploit this bounty. Finally, continuous effort must be devoted to developing methods for evaluating speech search systems, so that progress achieved can also be measured.

Ultimately, there was a call for a paradigm shift. The goal of speech retrieval should move from being system focused to being service focused. Instead of bringing the spoken content collection into the speech search lab, technologies should be made available to build speech search systems as near as possible to the source of the content and the people who use it. Research and development should focus on putting component services necessary for building spoken content retrieval systems at the disposal of content seekers and content providers. These groups are in a unique position to conceive and design systems that address specific use cases for specific collections. Additionally, systems built by groups who are close to the content are best positioned to exploit human effort, in the form of direct or implicit annotation or semantic enrichment. Information generated by human interaction with collections is invaluable in refining the performance of speech content retrieval systems.

As the workshop closed, there was an atmosphere of sober awareness of the magnitude and number of challenges yet facing the field of spoken content retrieval, but also a strong sense of optimism for the future. The future by necessity, will be one in which speech technology and information retrieval are more tightly combined to develop the algorithms, techniques and technologies necessary to address concrete user needs in real-world application scenarios.

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] J. Kohler, M. Larson, F.M.G. de Jong, W. Kraaij, and R.J.F. Ordelman, editors. *Proceedings of the ACM SIGIR Workshop Searching Spontaneous Conversational Speech*, Singapore, July 2008. Centre for Telematics and Information Technology, Enschede, The Netherlands.

8 Acknowledgements

Sources of support for the SSCS 2008 workshop and the authors of this report include: CHORUS (IST-FP6-045480), MESH (IST-FP6-027685), AMIDA (IST-FP6-033812) and MultiMATCH (IST-FP6-033104).