Searching Spontaneous Conversational Speech

Franciska de Jong
University of Twente, The Netherlands
fdejong@ewi.utwente.nl

Douglas W. Oard
University of Maryland, USA
oard@umd.edu

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

Stephan Raaijmakers
TNO ICT, The Netherlands
stephan.raaijmakers@tno.nl

July 27, 2007

Abstract

The SIGIR Workshop on Searching Spontaneous Conversational Speech was held as part of the 2007 ACM SIGIR Conference in Amsterdam. The workshop program was a mix of elements, including a keynote speech, paper presentations and panel discussions. This brief report describes the organization of this workshop and summarizes the discussions.

1 Background

Nearly a decade ago, we learned from the Text Retrieval Conference's Spoken Document Retrieval 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 means that search effectiveness held up well over a reasonable range of transcription accuracy.
- Sufficiently accurate Large-Vocabulary Continuous Speech Recognition (LVCSR) systems had been 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 communities 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).

Test collections are being developed in individual projects around the world, including AMI/AMIDA and CHIL (recorded meeting projects funded by the EU under the 6th Framework Program), IM2 (a Swiss recorded meeting project), MALACH (a NSF-funded project in the USA working with oral history), CHoral (a cultural heritage project in the Dutch NWO-funded programme CATCH), and GALE (a DARPA-funded project in the USA working with broadcast conversations). Some comparative evaluation activities for speech search technology are ongoing, including the Spoken Term Detection (STD) evaluation run by the National Institute for Standards and Technology (NIST) in the USA and the Cross-Language Evaluation Forum's Cross-Language Speech Retrieval track in Europe.

Each of the research communities involved in the initiatives mentioned above has established venues for agenda setting and for comparison of research results. For recorded meetings, this has included the MLMI workshops, and the NIST Rich Transcription evaluation, and the CLEAR evaluation sponsored by NIST and CHIL. Research on cultural heritage materials has recently been reported at workshops at the 2007 conference of the Association for Computational Linguistics in Prague and at the 2007 User Modeling conference in Corfu, Greece. For broadcast conversations, the DARPA GALE program (which includes research teams in North America, Europe and Asia) has to date been a principal research venue. Cross-cutting workshops have been held before at SIGIR (in 2001) and at the Human Language Technologies conference (in 2004), and a EU/NSF working group on spoken word archives recently identified several research issues related to the accessibility of recorded speech [3]. The time therefore seemed right to look more broadly across these research communities for potential synergies that can help to shape the information retrieval research agenda.

2 Before the Workshop

In the call for participation, contributions on a range of cross-cutting issues were solicited, including segmentation, content characterization, classification, exploiting multimodality, search effectiveness, interaction design, evaluation, and broader issues (e.g., applications, intellectual property, privacy). We invited fifteen experts from industry and academia to serve on the workshop's program committee. On the basis of their recommendations, seven papers that together spanned the identified topics were accepted.

On July 16, Technology Review published an interview with Peter Norvig (head of Google Research) in which he remarked on the key role of speech retrieval technology for providing access to large collections of multimedia materials [4]. Eleven days later, we met in Amsterdam to take up that challenge.

3 During the Workshop

Thirty researchers with a broad range of experience and expertise participated in the workshop. The program included a mix of elements designed to maximize interaction among participants from diverse backgrounds.

Authors	**Title**
Cuendet et al.	*An Analysis of Sentence Segmentation Features for Broadcast News, Broadcast Conversations, and Meetings*
Fiscus et al.	*Results of the 2006 Spoken Term Detection Evaluation*
Jones et al.	*Examining the Contributions of Automatic Speech Transcriptions and Metadata Sources for Searching Spontaneous Conversational Speech*
Kim et al.	*Advances in SpeechFind: CRSS-UTD Spoken Document Retrieval System*
Larson et al.	*Supporting Radio Archive Workflows with Vocabulary Independent Spoken Keyword Search*
Olsson	*Improved Measures for Predicting the Usefulness of Recognition Lattices in Ranked Utterance Retrieval*
van der Werff et al.	*Evaluating ASR Output for Information Retrieval*

Table 1: Papers presented at the workshop.

3.1 Keynote

Mark Maybury, Executive Director of MITRE's Information Technology Division (USA), led off the workshop with a keynote address. He began by summarizing the challenges posed by searching spontaneous conversational speech. Two MITRE efforts were then presented to illustrate some of those challenges: Audio Hot-Spotting and Cross-Language Automatic Speech Recognition. Some promising opportunities for future research were outlined as well. The keynote session was followed by a discussant, Gareth Jones (Dublin City University, Ireland).

3.2 Presentations and Panels

Table 1 briefly summarizes the seven research papers that were presented; full titles, author lists and abstracts are available on the workshop's Web page[1], and the full text of each paper is available in the workshop proceedings [2]. In addition to the paper presentations, one invited presentation (by Doug Oard, entitled *Who needs this?*) was included to stimulate discussion of interactions between user needs and technical capabilities. Two panels discussion we interleaved with the more formal presentations. The first, on "What new technologies do we need?" included Pavel Ircing, Marijn Huijbregts, Martha Larson, and Jonathan Mamou as panelists, with Stephan Raaijmakers as moderator. The second, on "Research directions" included Ken Church, Jon Fiscus, Franciska de Jong and Mark Maybury as panelists, with Doug Oard as moderator.

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

3.3 Discussion Themes

Sessions were structured to maximize opportunities for discussion, and a wide range of both high-level and detailed issues were addressed. The summary below is an effort to draw together some of the broader themes that emerged.

- **Leveraging Existing Capabilities.** Word error rates (WER) for planned speech (e.g., by news announcers) in studio conditions are nowadays around 10%, whereas for conversational speech, error rates are still often as high as 30 or 40%. Variations across recordings are, however, often far greater than variations across words: it is therefore often more reasonable from an IR perspective to ask what fraction of the content can be processed well enough to support specific tasks. Supervised machine learning techniques for topic segmentation, for example, place a greater premium on consistency than on raw accuracy, and "bag of words" retrieval techniques are robust in the presence of occasional errors. Extractive summarization, by contrast, requires that consecutive words be correctly recognized (so higher error rates may yield shorter and less informative snippets), and more sophisticated analysis (e.g., the entity tagging used in question answering systems) may be even more sensitive to recognition errors. As one of our panelists observed many years ago (in a machine translation context [1]), we already have some "good applications for crummy speech recognition." Those opportunities deserve our attention, even as speech researchers work to further improve their techniques.
- **Getting Beyond the Laboratory.** As is often the case early in the technology life cycle, leading-edge speech technology has relied on carefully controlled benchmark evaluations to stimulate and evaluate progress. One consequence of this is that robustness to training-test mismatch is well understood as an important issue, but it remains an under-researched problem. Scalability is recognized as another important challenge, but present speech processing techniques are in general quite resource-intensive. Information retrieval research, by contrast, often emphasizes both robustness and scalability. There is therefore significant potential for synergy, with speech research bringing us new capabilities that we can productively use, and our experience bringing new application contexts that can help to drive speech research in important directions.
- **Operational Employment.** Questions about what technologies we can build are an important first step, but our long experience with users of our technology allow us to bring another important set of questions to the table. Indexing workflows often contain specialized resources (e.g., topic inventories for use with text classification systems), and the "digital library" researchers with whom we work often pay particular attention to how those resources will be created. Selecting and preparing domain-specific training data for speech recognition would be one example of a similar task in the context of speech processing. Can we foster the development of a new generation of tools that leverage the participation of domain experts in such tasks? The collections people work with in the real world are often quite diverse; can we provide ways for managers of such collections to use some of their materials (e.g., e-text) to improve access to others (e.g., by allowing large scale adaptation of language models in the field rather than in the laboratory)? And do we have anything to say to the people who are initially creating spoken word materials; for example, are there simple techniques (e.g., speaker enrollment for talk show hosts) that might dramatically improve access in

some applications if the search technology could be designed to optimally leverage the resulting improvements?

Ultimately, information retrieval research brings two things to the table: real collections, and real users. The recent progress on processing spontaneous conversational speech serves a complementary role, bringing us new types of collections, and hence new types of research questions. Together, it seems that we're a good match!

4 Acknowledgments

The organizers are grateful to the IST project AMIDA (http://www.amidaproject.org/) and CTIT (http://www.ctit.utwente.nl/research/sro/nice) for supporting the travel of the keynote speaker and some student participants. The work of Douglas Oard was supported in part by NSF award IIS-0122466. The work of Franciska de Jong and Roeland Ordelman was partly supported by the Dutch bsik-programme MultimediaN (http://www.multimedian.nl) and the EU projects MESH (http://www.mesh-ip.eu/), and MediaCampaign (http://www.media-campaign.eu/).

References

[1] K.W Church and E.H. Hovy. Good applications for crummy machine translation. *Machine Translation*, 8(4):239–258, 1993.

[2] F.M.G. de Jong, D.W. Oard, R. Ordelman, and S. Raaijmakers (eds.), editors. *Proceedings of ACM SIGIR Workshop on Searching Spontaneous Conversational Speech.* 2007. ISBN=978-90-365-2542-8.

[3] J. Goldmann and S. Renals et al. Accessing the spoken word. *International Journal on Digital Libraries*, 5(4):287–298, 2005. ISSN=1432-5012.

[4] K. Greene. The future of search: The head of google research talks about his group's projects. *Technology Review*, 2007. http://www.technologyreview.com/Biztech/19050/.