Searching Multimedia Content with a Spontaneous Conversational Speech Track

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

Roeland Ordelman
University of Twente
Enschede, Netherlands
& Sound and Vision
Hilversum, Netherlands
rordelman@beeldengeluid.nl

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

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

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

Categories and Subject Descriptors: H.3.1 [Information Storage and Retrieval]:Content Analysis and Indexing

General Terms: Algorithms, Design, Languages

Keywords: spoken content, speech, speech recognition, audio-visual retrieval, multimedia access

1. INTRODUCTION

For multimedia with a speech track, the spoken word is a key carrier of meaning. The potential of spoken audio to support multimedia access is undisputed, yet speech remains under-exploited by most audio-visual retrieval systems. Spoken document retrieval research effort invested into developing broadcast news retrieval systems has yielded impressive results [4]. This success has, however, yet to be extended to spoken content domains involving unplanned, less-highly conventionalized, conversational speech. Such domains include podcasts, video diaries, lifelogs, meetings, call center recordings, social video networks, Web TV, conversational broadcast, lectures, discussions, debates, interviews and cultural heritage archives.

2. WHAT IS SPONTANEOUS CONVERSATIONAL SPEECH?

We produce spontaneous conversational speech when we speak to each other without previously selecting words or composing sentences. No crisp distinction exists between planned and unplanned speech. However, a characterization of what does *not* constitute spontaneous conversational speech proves helpful in characterizing the difference. Speech that is produced by reading a script or making use of other supporting material, such as tele-prompter text or detailed notes is not spontaneous. When the speaker speaks without the intention of engaging in dialogue, implicit or explicit, the speech produced is not conversational.

Speech produced in a spontaneous conversational context has a number of distinctive characteristics. Particularly challenging to speech recognition technology is the variability in speaking rate, care of articulation and pronunciation. Variability also leads to a markedly looser adherence to syntactic and semantic production rules, noted in [5]. In conversational settings, human speakers stray off topic, or make use of a rich reperatoire of figurative language – the resulting lexical variability exacerbates the out-of-vocabulary problem faced by the recognizer. Within conversational exchange, sentence fragments, emotional speech dynamics and speaker overlap contribute to communication flow and appear as the rule rather than the exception. A typical inventory of spontaneous effects and disfluencies is provided by [1], which lists filled pauses, stressed or lengthened function words, false-starts, self edits, word fragments, breaths, long pauses and extraneous noise. Style of speech, spontaneous vs. read, is an important factor in determining word error rate [7].

Planned speech and spontaneous conversational speech also differ with respect to their overall structure. Planned speech, such as broadcast news, is typically organized into topical segments, for example, news stories, that can be treated as documents within the retrieval system. Spontaneous, conversational speech is less well structured and topic can change spontaneously [2]. Within the speech stream topic boundaries are challenging to identify and may not be well defined. A system that provides access to spontaneous conversational content faces the challenge of automatically determining the unit of retrieval that will be returned in response to the user's query.

3. SSCS WORKSHOP SERIES

Two converging trends have led to a renewed interest in developing algorithms and systems that provide access to multimedia collections containing spontaneous conversational content. First, speech recognition technology has reached a level of maturity that automatically generated speech transcripts have sufficient quality to be useful for spoken audio indexing in conversational domains. The speech recognition community has invested focused research effort in the area of spontaneous speech since the early nineties [5]. Second, the amount of content accumulating both on the World Wide Web and in private collections has become over-

MM'09, October 19–24, 2009, Beijing, China.
ACM 978-1-60558-608-3/09/10.

whelming and the demand for access solutions to multimedia with a spoken component is steadily increasing.

The development of robust, scalable, affordable approaches for accessing multimedia collections with a spoken component requires the sustained collaboration of researchers in the areas of speech recognition, audio processing, multimedia analysis and information retrieval. The SSCS workshop is dedicated to supporting and promoting work that integrates these diverse disciplines to confront the challenges of spontaneous conversational spoken content. SSCS applies sustained effort to the goal of creating a forum that brings together otherwise isolated speech retrieval researchers. Previous SSCS workshops [3, 6], held in conjunction with ACM SIGIR, resulted in a list of key issues that will be addressed by presentations, demos and discussion at SSCS 2009.

4. ISSUES IN SEARCHING SPEECH

The development of real-world systems for accessing spoken content – multimedia with a speech track – faces a series of challenges, some of which are the subject of on-going investigation and others which remain to be addressed. These challenges naturally fall into four categories.

4.1 Integration

Multimedia retrieval systems must become highly effective in exploiting the voluminous quantities and diverse forms of information made available by speech recognition and audio analysis. Researchers continue to develop and refine methods for applying information retrieval techniques to speech recognizer output. This output takes many forms, including lattices of word-sequence hypotheses generated by recognizers and representations involving subword units. Audio analysis is able to produce information concerning speaker identity, other speaker characteristics such as gender, speaker emotional state and speaking style. This information has yet to be fully exploited by spoken content retrieval algorithms. Multimedia systems become “self improving” when they are able to improve their performance by integrating their own byproducts. For example, speech recognition transcripts can be used for unsupervised adaptation or recursive metadata refinement. Finally, it is important not to neglect the “multi” in multimedia. Indexing features derived from video or accompanying metadata can be exploited. Multimedia retrieval approaches based on concept detection can be expanded to include cross-modal concepts.

4.2 Interface/Interaction

Retrieval algorithms achieve efficacy only in combination with intuitive, comfortable interfaces that offer users an appropriate range of interaction functionality. Further research is needed in the area of generating visual surrogates for representing spoken content in results lists and in browsing interfaces. Cross-media linking and link visualization has an important role to play. Audio and video players should offer users the best possible “intelligent playback” functionality to support the review of retrieval results. It is clear that users must be offered listen-in points relevant to their information needs, but players should also make it possible for users to fast forward by speaker turn or topic boundary.

4.3 Scale/Scope

Multimedia collections are no longer restricted in size by the availability of storage or bandwidth. As a result, spoken content retrieval is confronted with the need for large-scale indexing approaches that allow high-speed search on large collections. Systems providing cross-lingual access or containing multilingual content face the additional challenge of managing multiple languages in parallel. Finally, affordable, light-weight speech indexing solutions are necessary for small collections with specific appeal, such as users with narrowly defined topical interests or belonging to a specific linguistics community.

4.4 Community

Multimedia access systems benefit from the participation of stakeholders, both content providers and users, in system design, realization and optimization. These groups are positioned to provide not only specifications and examples of typical user information needs, but also to supply or create parallel resources, including professional metadata or user contributed annotations, which can improve the system by optimizing both speech recognition and retrieval algorithms. In particular, communities that emerge around digital collections made available online are a source of valuable information and speech-based multimedia systems need to take advantage of its full range (i.e, tags, ratings, comments, corrections, usage information, friendship relationships).

5. ACKNOWLEDGMENTS

We would like to thank the AMI Consortium and the Netherlands Institute for Sound and Vision for their sponsorship of SSCS 2009. Projects that support(ed) the authors include the Dutch-funded research program MultimediaN[1] (Twente) and the EU NoE PetaMedia[2] (Delft).

6. REFERENCES

[1] J. Butzberger, H. Murveit, E. Shriberg, and P. Price. Spontaneous speech effects in large vocabulary speech recognition applications. In *HLT ’91: Workshop on Speech and Natural Language*, pages 339–343, 1992.

[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. *Speech and Audio Processing, IEEE Transactions on*, 12(4):420–435, 2004.

[3] F. de Jong, D. Oard, R. Ordelman, and S. Raaijmakers. Spoken content retrieval: Searching spontaneous conversational speech. *ACM SIGIR Forum*, 41(2):104–108, 2007.

[4] J. Garofolo, G. Auzanne, and E. Voorhees. The TREC Spoken Document Retrieval track: A success story. In *RIAO*, pages 1–20, 2000.

[5] B. Juang and L. R. Rabiner. Automatic speech recognition – A brief history of the technology. *Elsevier Encyclopedia of Language and Linguistics*, 2005.

[6] J. Kohler, M. Larson, F. de Jong, W. Kraaij, and R. Ordelman. Spoken content retrieval: Searching spontaneous conversational speech. *ACM SIGIR Forum*, 42(2):67–77, 2008.

[7] M. Weintraub, K. Taussig, K. Hunicke-Smith, and A. Snodgrass. Effect of speaking style on LVCSR performance. In *ICSLP*, pages 16–19, 1996.

[1] http://www.multimedian.nl

[2] http://www.petamedia.eu