Special Issue on Searching Speech

The special issue of the *ACM Transactions on Information Systems* is devoted to algorithms and systems that use speech recognition and speech processing techniques to make spoken audio searchable. The research area of spoken content indexing and retrieval has a long history dating back to the development of the first broadcast news retrieval systems in the 1990s. Recent years, however, have brought a new wave of research interest in the topic of "searching speech", which can be attributed to the convergence of two trends. First, the recent growth of digital collections containing spoken audio or video with a speech track has been unprecedented. This growth is most readily evident in the tremendous volume of multimedia content now available on the Internet. Second, after having undergone an extended period of development, speech recognition technology and other audio processing techniques have reached a new level of maturity, offering new technical possibilities.

The renewed attention devoted to searching speech has carried the focus of research efforts beyond broadcast news into more challenging domains. Work on searching speech is steadily moving towards spoken audio that is produced outside of the studio and under challenging conditions. The speech is spontaneous, conversational, and cannot be assumed to be limited to English or other well-researched languages. In these settings, spoken audio is characterized by high variability in subject matter, recording conditions, as well as speaker pronunciation and speaking style. Often, systems must be built with limited training resources. This special issue presents four articles that address a range of topics related to spoken content indexing and retrieval that open perspectives on the new challenges facing the research area.

Dong Wang et al. address spoken term detection (STD), the task of locating occurrences of query words or groups of words in spoken audio. STD is challenging since the identity of the words that must be detected is not known to the system at indexing time. This work focuses on compensating for weaknesses in system models that lead to errors in cases in which the system must detect a term that was not only unknown at indexing time, but also failed to ever occur in the training data.

Saturnino Luz explores using the structure of discourse and spontaneous interactions to automatically segment meetings. The study is carried out in the domain of interdisciplinary medical teams meetings. Structure information including duration of speech and sequence of speakers is used in order to divide meetings into their components, which are discussions of individual patient cases. The approach is compared to more expensive alternatives involving more complex additional resources.

Javier Tejedor et al. present a comparison of methods for detecting spoken terms that are presented to the system in spoken form. This task, referred to as query-by-example spoken term detection (QbE STD), presents challenges beyond those of detecting spoken terms expressed as text queries. The focus of the work is on QbE STD systems that are query-independent in the sense that they identify occurrences of query terms without needing resources from the target language. Different methods of feature extraction and different detection frameworks are compared.

Pere R. Comas et al. describe a question-answering system that returns sections of spoken audio containing answers to factoid questions. The system incorporates named-entity detection, syntactic parsing, and coreference resolution. Their investigation explores techniques that work well for finding answers in both text and spoken audio

DOI 10.1145/2328967.2328968 http://doi.acm.org/10.1145/2328967.2328968

and isolates the challenges that arise when developing a question-answering system specifically for spoken audio documents.

This special issue grew out of a series of workshops on Searching Spontaneous Conversational Speech held at ACM SIGIR 2007 (Amsterdam), ACM SIGIR 2008 (Singapore), ACM Multimedia 2009 (Beijing) and ACM Multimedia 2010 (Firenze). The series pursued the goal of fostering the development of robust, scalable, affordable approaches for accessing spoken content collections and encouraging cross-pollination among research activities in the areas of speech recognition, audio processing, multimedia analysis, and information retrieval. It aimed to promote investigation in new areas of spoken content including lectures, meetings, interviews, debates, conversational broadcast (e.g., talk-shows), podcasts, call center recordings, cultural heritage archives, social video on the Internet, spoken natural language queries, and the Spoken Web. We hope that this special issue will make a further contribution to this goal, providing groundwork that will support future productive research in the area of searching speech.

Martha Larson
Delft University of Technology, Netherlands
Franciska de Jong
University of Twente, Netherlands
Wessel Kraaij
Radboud University Nijmegen & TNO, Netherlands
Steve Renals
University of Edinburgh, UK
Guest Editors