# Introduction to the Special Section on Rich Transcription

THE term Rich Transcription spans multiple areas in audio processing, and its study marks a broadening of the concerns of automatic speech recognition (ASR) to cover the affiliated areas necessary for maximally useful applications. Whereas classical speech recognition focuses purely on converting a sequence of audio words to a sequence of textual words—without regard for capitalization, punctuation, speaker identity, pragmatic intent, and other high-level information—rich transcription attempts to produce a more highly annotated and informative output.

The study of rich transcription received a great impetus in 2002 when the Defense Advanced Research Projects Agency (DARPA) started the Effective Affordable Reusable Speech-to-Text (EARS) program. This program extended the previous HUB-4 and HUB-5 programs by adding an emphasis on metatdata extraction, in addition to traditional word recognition. The particular metadata tasks that were studied (http://nist.gov/speech/tests/rt/rt2004/fall/docs/rt04f-eval-plan-v14.doc) are as follows.

- Speaker diarization: the problem of segmenting speech into regions where only one person is talking, and then linking together speech (possibly from disjoint regions of time) from the same speaker.
- Identification of sentence-like units (SUs): the task of segmenting speech into units expressing separate thoughts or ideas, similar to sentences in written language, but taking into account that spoken language might not exhibit complete grammatical sentences.
- Disfluency detection: the dual problems of detecting the speech locations where a fluent word stream is interrupted (interruption point detection), and identifying those words that need to be removed in order to obtain the fluent word sequence of the intended utterance. This involves the labeling of pause fillers (e.g., "uh"), edit words (e.g., "I mean"), and the words that the speaker meant to replace in a self-repair.

Clearly, other forms and definitions of metadata are possible, and above tasks are offered only for illustrative purposes.

While rich transcription adds a new emphasis on various forms of metadata annotation, it also maintains a strong focus on improving automatic speech recognition from a core word-error-rate point of view. This is reflected in the composition of the special issue, with about half the papers addressing ASR. Here, there is a great deal of current interest in topics such as discriminative training, the use of large amounts of training data, unsupervised and semisupervised training, and system architecture and combination. Several of the papers presented here describe the design, implementation, and results of state-of-the-art large-vocabulary ASR systems, and the reader will get a good feeling for the algorithms and data processing involved.

The production of rich transcriptions is important in enabling machines to do a better job at extracting, indexing, summarizing, and presenting important information. Diarization, for example, when linked with speaker-identification, allows for indexing and querying on specific speaker names. The identification of sentence-like units is closely linked to the production of proper punctuation and capitalization—which makes reading the written text much easier. Similarly, the detection and removal of disfluencies is relevant to producing more readable transcripts. Both sentence-like segmentation and disfluency detection are also a prerequisite for downstream natural language processing, such as parsing, translation, or summarization.

For this special issue, we solicited articles on original theoretical and applied research broadly related to rich transcription. In particular, we encouraged submissions in the following areas: speech recognition algorithms and methods, natural language processing for rich transcription, speaker recognition algorithms and methods, algorithms for exploiting large amounts of training data, novel approaches to feature extraction and ASR, unsupervised and semisupervised training, and performance analysis and evaluation. We received 19 submissions spanning a broad range of topics. The accepted papers include two on novel approaches to speech recognition, two on diarization, three on other aspects of metadata annotation, and four on speech transcription systems *per se*. Taken together, this collection is an excellent representation of the state-of-the-art in rich transcription today.

As a final word, we would like to thank the Editor-in-Chief under which this project started, Isabel Trancoso, and the Editor-in-Chief under which it was concluded, Mari Ostendorf. Both have provided outstanding encouragement and guidance. We would also like to thank the nearly 60 reviewers who did an excellent job in reviewing the papers, and Kathy Jackson from the IEEE Signal Processing Society for her administrative assistance throughout the process.

GEOFFREY ZWEIG, *Guest Editor*
IBM T. J. Watson Research Center
Yorktown Heights, NY 10598 USA

JOHN MAKHOUL, *Guest Editor*
BBN Technologies
Cambridge, MA 02138 USA

ANDREAS STOLKE, *Guest Editor*
SRI International
Menlo Park, CA 94025 USA

Digital Object Identifier 10.1109/TASL.2006.882751

**Geoffrey Zweig** (M'02) received the B.A. degree in physics (highest honors) in 1985 and the Ph.D. degree in computer science from the University of California, Berkeley, in 1998. His thesis was on the application of Bayesian networks to automatic speech recognition.

He joined the IBM T. J. Watson Research Center, Yorktown Heights, NY, in 1998 and is the Manager of Advanced LVCSR Research. At IBM, he has pursued research ranging from the use of boosting in speech recognition to automated call-center quality monitoring and directory assistance applications. He participated in the 2001 DARPA-sponsored HUB-5 evaluations, and led IBM's efforts in the 2003 and 2004 "Effective Affordable Reusable Speech-to-Text" (EARS) evaluations. Currently, Geoffrey leads the IBM speech recognition effort for the DARPA "Global Autonomous Language Exploitation" (GALE) program.

Dr Zweig is an Associate Editor of the IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING.

**John Makhoul** (S'64–M'70–SM'78–F'80) received the B.E. degree from the American University of Beirut, Beirut, Lebanon, the M.Sc. degree from The Ohio State University, Columbus, and the Ph.D. degree from the Massachusetts Institute of Technology, Cambridge, all in electrical engineering.

Since 1970 he has been with BBN Technologies, Cambridge, where he is a Chief Scientist working on various aspects of speech and language processing, including speech recognition, optical character recognition, language understanding, speech-to-speech translation, and human–machine interaction using voice. He is also an Adjunct Professor at Northeastern University, Boston, MA.

Dr. Makhoul has received several IEEE awards, including the IEEE Third Millennium Medal. He is a Fellow of the Acoustical Society of America.

**Andreas Stolcke** (M'95–SM'05) received the Ph.D. degree in computer science from the University of California, Berkeley, in 1994.

He is currently a Senior Research Engineer at SRI International, Menlo Park, CA, and the International Computer Science Institute, Berkeley. His research interests are in applying novel modeling and learning techniques to speech recognition, speaker identification, and natural language processing. He authored and coauthored over 120 research papers, as well as a widely used toolkit for statistical language modeling.