Li Deng, Kuansan Wang,
and Wu Chou

# Speech Technology and Systems in Human-Machine Communication

Speech technology and systems in human-machine communication have witnessed a steady and remarkable advancement over the last two decades. Fundamental changes have taken place from theoretical foundations to practical systems, from laboratory prototypes to commercial products, and from proprietary softwares to industrial standards. As the information age continues, research in speech technology is further accelerated by the advent of powerful computing devices, the data-driven pattern recognition methods, and the need to generate machine understandable metadata for Web contents and other information sources.

Although various systems are built and applied to numerous applications, the full potential of speech technology still remains to be uncovered. This special section fills the need of a comprehensive review of new approaches and advances of speech technology under a broad perspective of intelligent human-machine communication. Speech technology and systems touch upon many essential signal processing techniques and are in the core of multimodal/multimedia communication research. We hope that such a systematic and up-to-date overview of the field can bring the awareness and applications of speech technology closer to the general signal processing community.

New research trends and directions in the field of speech technology and human-machine communication systems have been evolving rapidly in recent years due to the changing business environment and technology advances. With the Internet and the Web, an increasingly large amount of voice and speech data is made available. This, together with fast computing devices, leads to a new wave of advances on speech "document" understanding and multimedia/multimodal content search. New algorithms are being studied, many of which may not have been computationally feasible in the old days. Also, the large deployment of voice over IP (VoIP) has revitalized the research on noise-robust speech processing and recognition over the IP network, which is a very different environment than in the past. While scientific rigor remains the paramount selection criterion, an attempt is made to provide a balanced coverage of new research trends among articles selected for this special section.

About one year ago, we put out the call for papers for articles about speech technology and human-machine communication. We received a large number of submissions, and we would like to thank all authors for their submission. After a peer-review process, nine articles were selected that provide a comprehensive overview of the landscape in speech technology and human-machine communication. A brief overview of the selected articles is provided below in the context of the general theme of this special issue—human-machine communication.

## IN THIS ISSUE

Although speech is the most natural modality to convey meanings and intentions among humans, in human-machine communication, tremendous work has to be done to make sense of the recognized words that are often error prone. This is called speech understanding or spoken language understanding. The article by Wang et al. presents an introduction to the statistical approach to speech understanding based on a noisy source-channel modeling framework that guides most of the modern research in this area. In addition to outlining the common model components shared between speech recognition understanding technologies, the article addresses specific issues in speech understanding, including how the "meanings" are defined and represented in terms of the semantic frame and how problems of word alignment and multiple segments arise due to the greater flexibility in the temporal ordering of word segments (phrases) in speech understanding than the counterpart problem of pronunciation variation in speech recognition.

A practical human-machine dialogue and understanding system is described in the article by Gilbert et al. A system level overview is provided of a human-machine communication system that uses speech recognition, spoken language understanding, natural language generation and dialogue management for contact center automation. The system is capable of rudimental "understanding" of the intention of the user's speech (e.g., a telephone call) by classifying or routing it into one of several categories predefined according to business needs. This system serves as an example of the spoken language understanding technology introduced in the earlier article.

With the availability of overwhelmingly large multimedia contents including speech information, the under- standing and organization of such content is becoming increasingly important. Without the understanding in some form, efficient use of such information would not be possible. The article by Lee and Chen addresses this special "understanding" issue in depth, categorizing the problems into various well-defined subareas: spoken document indexing, segmentation, summarization, title generation, topic analysis, and named-entity extraction.

Basic principles and techniques developed for accomplishing these rudimentary "document understanding" tasks are reviewed using simple yet precise formulation. A prototype system realizing the various tasks is also presented including performance evaluation to support the concepts discussed in the article.

Yet another example of spoken signal "understanding" is provided in the article by Koumpis and Renals. It focuses on a somewhat different perspective for "understanding" the content of the spoken audio signal. This article provides a system-level overview of various approaches that can be applied to make content search of spoken audio as transparent as searching a text document but with much richer annotations derived from acoustic and semantic cues which are not available in a text document.

The article by Bellegarda presents a novel technique that uses a data-driven framework for modeling global relationships implicit in large volumes of data, including speech data. The technique, called latent semantic mapping, is motivated by a similar technique in information retrieval. It exploits the hidden "meaning" (i.e., semantics) embedded in the surface form of the observations (words in text or speech as a signal) and can be considered as an implicit way of "understanding" the text or speech observations. Applications of this semantic-oriented technique to various speech and language processing problems, such as language modeling, document classification, junk e-mail filtering, pronunciation modeling, and speech synthesis, are described.

The next two articles, one by Morgan et al. and another by Bilmes and Bartels, are devoted to novel techniques and systems for automatic speech recognition. The first article advocates the use of acoustic features that are beyond the currently dominant ones based on power spectral envelops. The main theme of the article is to introduce ways of incorporating long-time (as well as short-time) functions of the time-frequency plan in the feature extraction process and to express these functions as posterior probabilities of speech categories. Systematic experiments and results are presented and analyzed. On the other hand, the second article focuses mainly on the acoustic modeling, instead of feature extraction, aspect of the speech recognizer architectures. It provides a review of statistical graphical models, which include the commonly used hidden Markov models as a special member. Representative examples are given to demonstrate the ability of the graphical models in expressing speech recognition systems.

The final two articles, one by Raj and Stern and anther by Wolfel and McDonough, cover the area of novel signal processing techniques for robust processing of speech in noise. The first article concerns the missing-feature methods closely linked with computational auditory scene analysis. Various techniques based on the identification of "missing" spectral features are reviewed that have proved to be effective in dealing with strong noise corruption of speech including impulsive noise or background music. The second article provides an overview of the mathematical framework and algorithmic realization of minimum variance distortionless response (MVDR) spectral estimation. This approach has received increasing attention in robust processing of speech signals.

## THANK YOU

We thank all the authors for the hard work they put in for preparing these articles that made this special section possible. We are also grateful to Area Editor Prof. Petar Djuric and Editor-in-Chief Prof. Ray Liu for the support and guidance during our preparation of this issue. Finally, we want to thank all the reviewers for their time and effort that provided valuable recommendations and suggestions helping us select the articles and improve their quality. [SP]