Speech Information Processing: Theory and Applications

By DOUGLAS O'SHAUGHNESSY
Guest Editor

LI DENG
Guest Editor

HAIZHOU LI
Guest Editor

This special issue of the PROCEEDINGS OF THE IEEE is dedicated to the processing of information found in speech. Viewing the vocal utterances of humans, whether natural or produced synthetically, as signals allows analysis of their information content at various levels and for different applications. The papers in this issue examine both theory and applications of speech information processing. It has been 13 years since the last special issue of the PROCEEDINGS focusing on signal aspects of speech processing; in this issue, we highlight recent progress focusing on higher level, information processing aspects in contrast with the lower level, signal processing aspects highlighted in the 2000 special issue.

The most natural form of communication for people is speech. Even in modern society, where computers and personal devices are mostly text and image based, people talk to each other (often using phones) more than doing almost any other activity. Speech communicates information from a speaker to one or more listeners. To assist that transfer of information, and to facilitate interaction with computers and databases via speech, we can employ machines to help the process.

This special issue highlights recent progress in speech processing, in both speech theory and applications.

In the past decade, we have seen intensive progress of speech technology that has helped people gain access to information (voice-automated call centers and voice search) and access huge volumes of speech information (e.g., spoken document retrieval, speech understanding, and speech translation). This research has been spurred by initiatives such as international benchmarking and standardization, and by increasingly fast and affordable computing facilities. Researchers have combined speech processing, natural language processing, and information retrieval technologies to address operational needs in real-world applications. As a result, speech information processing increasingly has been multidisciplinary.

Speech information processing primarily comprises automatic speech recognition (ASR), text-to-speech (TTS) synthesis, recognition of other useful aspects of speech (e.g., language, dialect, stress, emotional state, and speaker identity), and most importantly, all aspects of subsequent processing of the outputs of these systems for symbolic-information-focused applications such as translation, semantic understanding, and ranked information retrieval. One important potential application is for dialog with structured or even unstructured databases, where people can interact readily with computers over various communication links or with devices directly. For example, for phone calls, most companies have replaced the earlier interface of a human operator with that of a menu-based series of questions and answers, sometimes via voice but still often using a keypad. Users are often displeased with the inconvenience of

Digital Object Identifier: 10.1109/JPROC.2013.2252718

such cumbersome interaction and prefer a simpler voice interface, where they could easily state their needs and receive efficient service via speech understanding and dialogs.

Such ASR is very functional today in many systems, but applications remain very limited. Systems often search for simple keywords in a user's expression among a very small set of expected possibilities, which varies with the application. For example, callers may be prompted to state a telephone or credit card number, and many ASR systems often do quite well in deciphering well-spoken digit strings. Similarly, companies often ask callers to state the name of the person they seek; ASR accuracy is often facilitated by a small list of names. When callers instead may say whatever they wish, ASR is often far less accurate, especially in noisy environments or with cellphones subject to fading channels or other data network limitations.

Both acoustic modeling and language modeling are important parts of modern statistically based speech recognition algorithms. Hidden Markov models (HMMs) are widely used in many systems. Such HMMs are mathematical models that associate an output sequence of phonemes with an input sequence of short frames (e.g., every 10 ms) of speech. HMMs are popular as they can be trained automatically and are simple to use. Spectral analysis of each frame yields a vector consisting of mel-frequency cepstral coefficients (MFCC; the inverse Fourier transform (FT) of a short time window of speech, after use of an auditory frequency scale, a logarithmic-amplitude scaling (decibels), and an FT discarding the phase). The first paper in this Proceedings issue, by O'Shaughnessy, examines how ASR analyzes an input speech signal, from the point of view of data compression. As in any pattern recognition application, one must extract relevant features from the input signal in an efficient way.

Modern ASR uses context-dependent phoneme-based HMM models, with cepstral normalization to account for different speakers and recording conditions. For further speaker normalization one can use maximum-likelihood linear regression (MLLR). Many systems employ discriminative training techniques that optimize some classification-related measure of the training data, e.g., maximum mutual information (MMI), minimum classification error (MCE), and minimum phone error (MPE). Support vector machines (SVMs) have found use in ASR as discriminative classifiers, even better than artificial neural networks. Another important example of recent work here is conditional random fields (CRFs), which are probabilistic sequence models applied to many applications in audio, speech, and language processing. Our second paper, by Fosler-Lussier *et al.*, examines how CRFs are used in ASR and in related applications. This article provides an overview of CRF, describing the common linear-chain model and extensions. It gives a description of the mathematical techniques used in training and evaluating these models, as well as a discussion of the relationships with other probabilistic models.

One area of recent major focus has been the robustness of ASR accuracy, against noise and other mismatches between training and testing speech. A widely accepted process for ASR has been to use: pree-mphasis, short-time windowing, MFCCs, and HMMs, with various search and training methods. Search often involves the Viterbi method to convert a full examination of all HMM paths to the simpler task of finding the single best path. Much research has occurred recently on various ways to do discriminative and rapid training, to help overcome serious mismatch problems, where available speech training data often do not correspond well with speech from new speakers and transmission channels.

Our third paper, by Hermansky, looks at ways to handle unexpected acoustic elements in speech. This is motivated by observations of how humans recognize speech, which suggest the existence of multiple parallel processing streams in the human cognitive system. The paper examines earlier and recent work on multi-stream ASR, to treat noisy speech and handle unexpected out-of-vocabulary words. This focuses on feedback-based techniques that fuse information from different processing streams. The paper notes that the difference between ASR behavior on its own training data versus during normal operation can be viewed as a substitute for the human ability of "knowing when knowing."

Our fourth paper, by Lee and Siniscalchi, discusses ASR from the point of view of retrieving information. State-of-the-art ASR achieves good accuracy for well-formed utterances in many languages, by decoding speech into the most likely sequence of words among all possible sentences in a finite-state network (FSN) approximation of the knowledge sources for a given ASR task. However, not all information available can be directly integrated into the FSN. Humans recognize speech based on evidence that exists at various levels of the speech knowledge, ranging from acoustic phonetics to syntax and semantics. This suggests a bottom-up attribute detection and knowledge integration framework that could link speech processing with information extraction. This paper examines how to spot speech cues with a bank of attribute detectors, combining acoustic evidence to form cognitive hypotheses and verify theories. It describes an automatic speech attribute transcription (ASAT) framework, in an attempt to mimic some HSR capabilities with asynchronous speech event detection followed by bottom-up knowledge integration and verification.

Our fifth paper, by He and Deng, provides an optimization-oriented approach to the design of speech-centric information processing. Speech recognition has increasingly become a popular machine interface in consumer applications, such as internet search, smartphone, in-car telematics, call center services, and translation

services, where the speech recognition engine works with other technology components, which are subsystems themselves, in a pipeline process to deliver end-to-end solutions. In practice, we build most of the subsystems independently and optimize them under their respective performance criteria, ignoring the interactions between them. As a result, there is a mismatch between how the subsystems are trained and how the trained subsystems are used in the operation environment. This paper presents an optimization-oriented statistical framework for system design where the interactions between the subsystems are fully incorporated, establishing design consistency between the optimization objectives and the end-to-end system performance metrics. Such a unified perspective at the system level allows for effective system design and optimization of overall performance.

Our sixth paper, by Li *et al.*, provides an overview of state-of-the-art spoken language recognition technology. Automatic recognition of spoken language is a key enabling technology in many speech processing applications, such as multilingual speech recognition, spoken document retrieval, call center automation, and speech intelligence and surveillance. Humans are born with the ability to discriminate between spoken languages as part of human intelligence. We know that humans recognize languages through a perceptual or psychoacoustic process that is inherent in the auditory system. Therefore, the type of perceptual cues that human listeners use is always the source of inspiration. The history of spoken language recognition research generally started in 1990s, which benefited from previous findings in computational phonetics and linguistics, human perceptual experiments, and technological breakthroughs in related areas, such as signal processing, pattern recognition, and machine learning. In this paper, Li *et al.* provide an introductory tutorial on the fundamentals of the theory and the state-of-the-art solutions, from both phonological and computational perspectives. It is a must-read article that provides a comprehensive review of current trends and future research directions.

Our seventh paper, by Young *et al.*, discusses statistical dialog systems. ASR requires language models (LMs) to help interpret the acoustic patterns of input speech. Early attempts in the 1970s to decipher speech using only acoustical analysis failed, as much of what humans use to interpret speech lies not only in the acoustic input to the ear, but also in the listener's expectations of what they are hearing. As most applications of ASR involve humans interacting with machines, we need ways to manage the human–computer dialog by modeling the coherent sequences of sentences, propositions, speech acts, or turns-at-talk. Statistical dialog systems have been studied to address the need for reducing the cost of laboriously handcrafting complex dialog managers and for providing robustness against speech recognition errors. The partially observable Markov decision process (POMDP) is a practical solution under a data-driven framework. In this paper, Young *et al.* provide an overview of the current state of the art in the development of POMDP-based spoken dialog systems.

Our eighth paper, by Zhou, provides a unique perspective to statistical machine translation. The history of machine translation generally started in the 1950s. However, the real progress was much slower; early systems used large bilingual dictionaries and manually coded rules to map a sentence from source to target language, which was called rule-based machine translation (RBMT). Since the late 1980s, as computational power increased and became less expensive, statistical models for machine translation have become popular. While statistical machine translation (SMT) generally offers better translation quality and shorter development cycle than RBMT, RBMT is more predictable and grammatically superior than SMT. There has been a renewed interest combining syntactic and linguistic knowledge into statistical systems. This paper takes a unified perspective to reveal both the connections and contrasts between ASR and SMT that are based on finite-state transducer and synchronous context-free grammar, providing insights into possible tighter integration of ASR and SMT for improved speech translation performance.

Our ninth paper, by Narayanan and Georgiou, deals with handling behavioral aspects of speech, e.g., ways to measure and model human behavior via one's speech. Many applications from commerce to healthcare need efficient ways to assess people's behavior, and speech processing may allow a simple and convenient way to do so. This paper describes emerging methodologies and applications to quantitatively understand and model both typical and atypical human behavior via their speech. It develops models that offer both predictive and decision-making support, using examples from real-world applications, e.g., literacy assessment, autism diagnostics, and psychotherapy.

For many years, TTS was handled by expert system methodology, using a pole-zero model based on manipulating several vocal-tract resonances (formants) to simulate movements of the speech articulators while uttering different phoneme sounds. TTS was implemented with parallel or serial formant models (e.g., the Klatt model), due to its ease of programming. However, formant models suffered from the use of artificial excitation of the vocal-tract models, leading to a buzzy quality. Such methods were largely replaced in the 1990s by concatenation methods using unit selection, which chooses appropriate subword units from a large speech database to generate high-quality speech. While unit selection was highly popular, it was limited in synthesizing varied voice characteristics, such as voices of different styles, speakers, and emotions. In the last

decade, a statistical parametric method based on HMMs was invented to overcome these problems. The HMM-based method not only allows for a compact synthesis system, but also offers flexibility in voice training and adaptation.

Our final paper, by Tokuda *et al.*, gives a general overview of HMM-based speech synthesis. A decade or so ago, much TTS was generated using approaches based on vocal-tract models, varying second-order resonances that modeled how human formants varied dynamically in natural speech, i.e., an expert-system method based on much spectral analysis of how humans spoke. Many versions of such (based on the Klatt model of the 1980s) still occur in the media, e.g., Stephen Hawking's TTS. More recently, a unit-selection approach employing segments of real human speech lent a more natural sound to TTS. This paper's statistically based TTS has dominated in recent years. Its main advantage is its flexibility in changing speaker identities, emotions, and speaking styles. Segment concatenation TTS could only simulate the voice of the speaker who recorded the database.

Recent work on alternatives to HMMs, e.g., missing-feature analysis and example-based ASR, seem to still be in preliminary stages, not yet ready to compete seriously with HMMs. Deep belief networks are even newer approaches in machine learning that may make a significant impact on ASR technology. The interested reader can find more detail on these matters in a recent issue of the IEEE Signal Processing Magazine (November 2012). ■

ABOUT THE GUEST EDITORS

Douglas O'Shaughnessy (Fellow, IEEE) received the Ph.D. degree from the Massachusetts Institute of Technology (MIT), Cambridge, MA, USA, in 1976.

He has been a Professor at the Institut national de la recherche scientifique (INRS), Montreal, QC, Canada and an Adjunct Professor at McGill University, Montreal, QC, Canada, since 1977. He is the author of the textbook *Speech Communications: Human and Machine* (Piscataway, NJ, USA: IEEE Press, 2000). His research interests include diverse aspects of speech processing, focusing recently on automatic speech recognition.

Prof. O'Shaughnessy is a Fellow of the Acoustical Society of America. He served 12 years as an Associate Editor for the *Journal of the Acoustical Society of America*, and is the founding Editor-in-Chief of the *EURASIP Journal on Audio, Speech, and Music Processing*. He is now the Vice-President of the International Speech Communication Association (ISCA) and Chair of the IEEE Signal Processing Society's (SPS's) Speech and Language Technical Committee. He has served also as an Associate Editor for the IEEE Transactions on Speech and Audio Processing, on the SPS Board of Governors, and now on the IEEE Technical Activities Board (TAB) Periodicals Committee. He has presented tutorials on speech recognition at the 1996, 2001, and 2009 International Conference on Acoustics, Speech, and Signal Processing (ICASSP) and the 2003 International Conference on Communications (ICC).

Li Deng (Fellow, IEEE) received the Ph.D. degree from the University of Wisconsin—Madison, Madison, WI, USA.

He joined the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON, Canada, in 1989, as an Assistant Professor, where he became a tenured full Professor in 1996. In 1999, he joined Microsoft Research, Redmond, WA, USA, as a Senior Researcher, where he is currently a Principal Researcher. Since 2000, he has also been an Affiliate Full Professor and graduate committee member in the Department of Electrical Engineering, University of Washington, Seattle, WA, USA. Prior to MSR, he also worked or taught at the Massachusetts Institute of Technology, Cambridge, MA, USA; ATR Interpreting Telecom. Research Laboratory, Kyoto, Japan; and the Hong Kong University of Science and Technology (HKUST), Hong Kong. In the general areas of speech/language technology, machine learning, and signal processing, he has published over 300 refereed papers in leading journals and conferences and three books, and has given keynotes, tutorials, and distinguished lectures worldwide. His recent technical work (since 2009) on industry-scale deep learning with colleagues and collaborators has created significant impact on speech recognition, signal processing, and related applications.

Prof. Deng is a Fellow of the Acoustical Society of America and the International Speech Communication Association (ISCA). He served on the Board of Governors of the IEEE Signal Processing Society (SPS; 2008–2010). More recently, he served as the Editor-in-Chief for the IEEE Signal Processing Magazine (2009–2011), which earned the highest impact factor among all IEEE publications and for which he received the 2011 IEEE SPS Meritorious Service Award. He currently serves as the Editor-in-Chief for the IEEE Transactions on Audio, Speech, and Language Processing.

Haizhou Li (Senior Member, IEEE) received the B.Sc., M.Sc., and Ph.D. degrees in electrical and electronic engineering from South China University of Technology, Guangzhou, China, in 1984, 1987, and 1990, respectively.

Currently, he is the Principal Scientist and Department Head of Human Language Technology, Co-Director of Baidu-I2R Research Centre, Institute for Infocomm Research, Agency for Science, Technology, and Research (A*STAR), Singapore. He is also a conjoint Professor at the School of Electrical Engineering and Telecommunications, University of New South Wales, Kensington, N.S.W., Australia. He has worked on speech and language technology in academia and industry since 1988. He taught at the University of Hong Kong (1988–1990), South China University of Technology (1990–1994), and Nanyang Technological University (2006–present). He was a Visiting Professor at CRIN in France (1994–1995). He was appointed as Research Manager at the Apple-ISS Research Centre (1996–1998), Research Director in Lernout & Hauspie Asia Pacific (1999–2001), and Vice President in InfoTalk Corp., Ltd., (2001–2003). His current research interests include automatic speech recognition, speaker and language recognition, and natural language processing. He has published over 200 technical papers in international journals and conferences.

Dr Li has served as an Associate Editor of the IEEE Transactions on Audio, Speech and Language Processing, *ACM Transactions on Speech and Language Processing*, and *Computer Speech and Language*, and a Guest Editor of the Proceedings of the IEEE. He is an elected Board Member of the International Speech Communication Association (ISCA, 2009–2013), President-Elect of Asia Pacific Signal and Information Processing Association (APSIPA, 2013–2014), and Executive Committee Member of the Asian Federation of Natural Language Processing (AFNLP, 2010– 2012). He was a recipient of the National Infocomm Award of Singapore in 2001. He was named one of the two Nokia Visiting Professors 2009 by the Nokia Foundation in recognition of his contribution to speaker and language recognition technologies.