Sadaoki Furui, Li Deng, Mark Gales,
Hermann Ney, and Keiichi Tokuda

Fundamental Technologies in Modern Speech Recognition

Various modern techniques that form the basis of fundamental technologies underlying today's automatic speech recognition (ASR) research and applications have recently attracted new attention. These techniques have significantly contributed to the progress in ASR as a field of research. This special issue of *IEEE Signal Processing Magazine (SPM)* intends to bring together leading experts from various disciplines to explore the impact of new approaches to ASR—fortunately, we have been very successful in this regard. This issue of the magazine provides a comprehensive overview of not only recent developments but also open problems.

The first article, "Large-Vocabulary Continuous Speech Recognition" by Saon et al., focuses on major large-vocabulary continuous speech recognition (LVCSR) techniques that have been applied to many languages: front-end processing, acoustic modeling, language modeling, hypothesis search, and system combination. Despite the commercial success and widespread adoption, the problem of LVCSR is far from being solved; background noise, channel distortion, foreign accent, and casual and disfluent speech or unexpected topic change can cause automated systems to make egregious recognition errors. This is because current LVCSR systems are not robust to mismatched training and test conditions and cannot handle context as well as human listeners, despite being trained on thousands of hours of speech and billions of words of text. This article also covers recent techniques of model adaptation and discriminative training that have been investigated to solve this problem.

Many feature extraction methods that have been used for ASR have either been inspired by analogy to biological mechanisms, or at least have similar functional properties to biological or psychological properties for humans or other mammals. These methods have, in many cases, provided significant reductions in errors, particularly for degraded signals, and are currently experiencing a resurgence in community interest. Stern et al. summarize these biologically inspired methods and emphasizes the importance of this research trend to increase robustness of ASR systems against acoustic disturbances such as additive noise and reverberation in their article, "Hearing Is Believing."

Large-vocabulary recognizers represent each word in terms of subword units to cover all possible acoustic phenomena with a limited size of training corpus. Typically, the subword unit is the phone, one type of basic speech sound element such as a single consonant or a vowel. Each word is then represented as a sequence, or several alternative sequences, of phones specified in a pronunciation dictionary. Other choices of subword units have been studied as well. The choice of subword units and the way in which the recognizer represents words in terms of combinations of those units is the problem of subword modeling, sometimes also called pronunciation modeling. The article, "Subword Modeling for Automatic Speech Recognition" by Livescu et al., reviews past, present, and emerging approaches to subword modeling, based on the popular computational framework of graphical modeling developed from machine-learning research.

Discriminative training or learning techniques have been shown to consistently outperform the maximum likelihood paradigm for model training in ASR. Consequently, discriminative training methods of today are fundamental components of state-of-the-art systems. The article, "Discriminative Training for Automatic Speech Recognition," by Heigold et al., gives a comprehensive overview of discriminative training methods for acoustic model training, covering various related aspects of discriminative training, i.e., specific training criteria and their relation, statistical modeling, different parameter optimization approaches, efficient implementation of discriminative training, and a performance overview. This article focuses mainly on parametric learning, or training the parameters of the ASR models given fixed model structures.

On the other hand, the next article, "Structured Discriminative Models for Speech Recognition," by Gales et al., shifts the focus to learning "structures." Discriminative models, in which the posterior probability of the classes (sentences) given the observations is directly modeled, can increase performance while allowing a wide range of features from the observation and word sequences to be used for inference. The number of labels (in the word sequence) and the number of observations (frames) differ vastly for typical ASR problems, and it is the information at the global sequence level rather than at the local frame level that ASR decisions are based on. Discriminative models that handle this type of "structured" data for both input and output are referred to as structural discriminative models. This article describes various structural discriminative models for ASR, which differ

Digital Object Identifier 10.1109/MSP.2012.2209906
Date of publication: 15 October 2012

from each other in terms of the observation features considered, training criterion, and how the latent variables are handled. In addition to the models that directly map from an observation sequence to a word sequence, probability distributions over word sequences can also be represented in the same structured form, yielding discriminative language models that is also covered in this article.

Over the last few years, major advances in both machine-learning algorithms and computer hardware have led to efficient methods for training deep neural networks (DNNs) that contain many layers of nonlinear hidden units, and it has been shown that, with the new learning architectures and methods, these DNNs can outperform Gaussian mixture models (GMMs) by a large margin on a variety of ASR benchmarks from small tasks such as TIMIT to large tasks such as switchboard consistently. Such drastic progress has not been seen in the ASR history for a long time. It has also been shown that DNNs can make better use of input representations, such as filterbank outputs, that retain more information about the sound wave than MFCCs or PLPs. The article, "Deep Neural Networks for Acoustic Modeling in Speech Recognition" by Hinton et al., provides an overview of this progress including the pioneering work in this area and specifically in large vocabulary ASR in which the direct use of a large set of tied context-dependent phone units as the DNN output vector plays a crucial role. It also describes some technical detail on how DNNs may be effectively trained, including several different variations of the method, and a number of successful uses of DNNs and related deep architectures for acoustic modeling in ASR. This article draws key parts of content from several articles published in the January 2012 special issue of *IEEE Transactions on Audio, Speech, and Language Processing*.

For the last 30 years, ASR has been dominated by techniques using hidden Markov models to model the time-varying aspects of the acoustics. GMMs are typically used to represent the observation densities in the Markov chain. This method allows for a generalization of the observed data as long as the distribution estimated by the model is a reasonable description of the unseen data. In many cases, such a description must be simplified to allow reliable estimates of all free parameters in the model, and as a result, fine details in the model are lost. Exemplar-based models have the potential to address this deficiency by building an instance of the model based only on the relevant and informative exemplars selected, exploiting sparse coding or compressive sensing techniques popular in both signal processing and machine learning, for that instance of the test data. The article "Exemplar-Based Processing for Speech Recognition" by Sainath et al., reviews exemplar-based processing and underscore its value in improving performance across a variety of speech recognition tasks. In addition to DNNs described in the preceding article, this article represents another emerging trend in ASR, thus, for this reason, it was selected to be included in this special issue.

Most of the current speech recognition applications require a close-talking microphone or a microphone placed near the speaker. Almost all of the applications would benefit from distant-talking speech capturing, where speakers are able to speak at some distance from the microphones without the encumbrance of hand-held or body-worn equipment. The major problem in distant(-talking) speech recognition (DSR) is the corruption of speech signals by both interfering sounds and reverberation. In recent years, research on reverberant speech processing has achieved significant progress in both the fields of audio processing and speech recognition, mainly driven by multidisciplinary approaches combining ideas from room acoustics, optimal filtering, machine learning, speech modeling, and speech enhancement. The article, "Making Machines Understand Us in Reverberant Rooms" by Yoshioka et al., reviews the state of the art of reverberant speech processing.

A major issue in DSR is to suppress noise and reverberation effects without distorting speech features. Microphone array techniques have received a great deal of attention for DSR due to their potential to achieve the highest quality in capturing speech signals emanating from a distance. The article, "Microphone Array Processing for Distant Speech Recognition" by Kumatani et al., provides an overview on the recent work on acoustic beamforming for DSR using microphone arrays, along with experimental results verifying the effectiveness of various algorithms. This article also describes an emerging technology in the area of far-field audio and speech processing based on spherical microphone arrays.

We hope you enjoy reading the articles in this special issue as much as we enjoyed creating them. We hope this special issue will be of use for you as an experienced researcher already conducting research on ASR, or for you as a student or a young researcher interested in ASR research.

THANK YOU

We thank all the authors who submitted proposals to this special issue, particularly those with accepted articles for the hard work they put in that made this special issue possible. We also thank all the reviewers for providing valuable feedback and recommendations that have significantly improved the articles. For many of those whose proposals did not get accepted, we would like to thank you for your effort and time but also would like to comfort you in that the decision was often not so much based on the quality of the proposal as the sole criterion but on our perceived theme as well as the allocations of the submitted proposals that would accomplish the theme. Finally, we are grateful to Abdelhak Zoubir, *SPM's* editor-in-chief, as well as to Dan Schonfeld, former area editor, special issues, for their support and guidance throughout the process of preparing and finalizing these articles. [SP]