Speech Communication and Signal Processing

FOREWORD

Communicating with a machine in a natural mode such as speech brings out not only several technological challenges, but also limitations in our understanding of how people communicate so effortlessly. The key is to understand the distinction between speech processing (as is done in human communication) and speech signal processing (as is done in a machine). When people listen to speech, they apply their accumulated knowledge of speech in relation to a language to capture the message. In this process, it is interesting to note that the input speech is processed selectively using the knowledge sources acquired over a period of time such as sound units, acoustic-phonetics, prosody, lexicon, syntax, semantics and pragmatics. This processing varies from person to person, and it is difficult for any individual to articulate the mechanism he/she is using in processing the input speech. This makes it difficult to write a program to perform the task of extracting message in speech by a machine. It should be noted that, for a machine, only the speech signal is available in the form of a sequence of samples, the rest of the mechanism involving identification of knowledge sources and invoking them on the input signal is a scientific challenge. Thus speech signal processing is one of the most interesting challenges that arouses curiosity among different scientific groups, such as linguists, phoneticians, (psycho)acousticians, electrical engineers, computer scientists and application engineers. The editorial board of SĀDHANĀ has rightly identified this topic to be addressed in a special issue. They have asked me to take the initiative to collect views of leading scientific groups, in the form of articles for this special issue. I am indeed fortunate to have been able to persuade several highly accomplished scientists in their field to contribute papers to this special issue. Here I present a brief overview of this special issue.

The paper by Saraha Hawkins rightly questions the informativeness of the acoustic speech signal, and explains the strengths and limitations of the standard representation of the speech signal in the phonological features and phonemes. The author proposes an alternative approach, called Firthian prosodic analysis, which places more emphasis on the formation of an utterance. This approach suggests formalism that forces us to recognize that every perceptual decision is context- and task-dependent, and hence there cannot be any predetermined or rigid sequence that is a result of the assumption that speech processing proceeds in a strictly serial order. In the next paper, Peri Bhaskara Rao highlights another important aspect of speech production, namely the false impression of similarity of text-to-sound rule sets across Indian languages. Using several illustrations, the author shows the divergence in the phonetic realizations of a given letter across Indian languages, and the importance of this for developing speech systems in these languages.

The next three papers focus on the voice source analysis, and in particular, the significance of the glottal closure instants (GCI) or epochs. Christophe and Nicolas propose time-scale Lines of Maximum Amplitude (LoMA) for the detection of GCI. The time-scale analysis is implemented using wavelet transforms. Using the LoMA the authors estimate the voice source parameters such as open quotient, amplitude of voicing and strength of excitation. Paavo Alku provides a critical review of the methods of glottal inverse filtering (GIF) for estimating the glottal volume

velocity waveform. The paper also discusses the parametrization method developed for quantification of the estimated glottal excitations, and also potential applications of the GIF method. The third paper in the voice source analysis category is on 'Epoch-based analysis of speech signals' in which the authors Yegnanarayana and Suryakanth review different epoch extraction methods, and describe how epoch locations can help in the estimation of instantaneous fundamental frequency, analysis of Lombard effect speech, etc. The authors also discuss several possible applications of the epoch-based analysis such as in speech enhancement and prosody manipulation.

Representation of speech is an important issue in many speech applications. In their paper on 'Auditory-like filter bank: An optimal speech processor for efficient human speech communication', Ghosh *et al* argue that the auditory filter bank in human ear is a near-optimal speech processor for efficient speech communication between human beings. They use mutual information criterion to design the optimal filter bank that provides maximum information on talkers' articulatory gestures derived from X-ray microbeam speech production database. In the next paper, Kawahara and Morise comprehensively discuss the technical foundations of the highly successful speech modification tools STRAIGHT and TANDEM-STRAIGHT. Another effective representation of speech information is through modulation spectrum that describes the temporal dynamics of the spectral envelope of short-time speech spectrum. Hynek Hermansky in his paper on 'Speech recognition from spectral dynamics' reviews the efforts to exploit the features of modulation spectrum in automatic speech recognition (ASR).

Fourier transform (FT) phase is normally ignored in the spectral representation of speech information. Hema Murthy and Yegnanarayana provide a comprehensive review emphasizing the importance of phase in speech processing using group delay functions. They discuss the effectiveness of group-delay functions in capturing the spectral information, and show several applications of the group delay functions in speech systems. The effectiveness of nonlinear models in capturing information in speech is demonstrated for various applications by Sreenivasa Rao in the paper on 'Role of neural network models for developing speech systems'. The author explores the possibility of capturing the prosody features using neural network models, which in turn could be used for text to speech synthesis, speech recognition, speaker recognition, language identification and characterization of emotion.

Statistical parametric speech synthesis based on hidden Markov model has become popular over the past few years, and seems to provide a good alternative to the well-established concatenative techniques. Simon King provides a tutorial introduction to this practical approach to speech synthesis in his paper on 'An introduction to statistical parametric speech synthesis'.

The last three papers deal with issues on realizing practical ASR systems. Umesh addresses the issue of interspeaker variability in speech, and reviews the studies made on this topic in the context of ASR. The paper provides an overview of vowel normalization studies as well as universal warping approach to speaker normalization. The next paper by Herve Bourlard *et al*, addresses the trends in multilingual speech processing. They emphasize that the prime mover behind the current trends has been the rise of statistical machine translation. The last paper of this issue is by Steve Renals on 'Automatic analysis of multiparty meetings', in which the author discusses the issues and challenges in the recognition and interpretation of multiparty meetings captured as audio, video and other signals. It is not only a multimodal, multiparty and multispeaker problem, but the challenge is in dealing with spontaneous and conversational interaction among a number of participants.

The special issue thus deals with many challenging issues in speech processing. As guest editor, I am very fortunate to have been able to gather this information from my friends who have readily accepted my invitation to contribute papers for this special issue. I am indeed

grateful to all the authors for their efforts. I am also grateful to the following reviewers who have provided a timely review of the papers, which helped the authors to improve the presentation of the material: G V Anand, S Chandrasekhar, C Chandra Sekhar, Rohit Sinha, Rajesh Hegde, K Sri Rama Murty, S R Mahadeva Prasanna, K Samudravijay, S P Kishore, Peri Bhaskararao, Douglas N Honorof, Louis Ten Bosch, Paavo Alku and Mark Hasegawa Johnson. Finally, I would like to thank Prof. R Narayana Iyengar, former Editor and Prof. G V Anand, Associate Editor of SĀDHANĀ for conceiving the idea of the special issue on the topic of 'Speech Communication and Signal Processing'.

October 2011

B YEGNANARAYANA
Guest Editor

International Institute of Information Technology,
Gachibowli, Hyderabad 500 032, India
email: yegna@iiit.ac.in