A mixture modeling approach to text-independent speaker ID

Douglas A. Reynolds, Richard C. Rose and Mark J. T. Smith

JASA
THE JOURNAL OF THE ACOUSTICAL SOCIETY OF AMERICA
Special Issue:
Additive Manufacturing and Acoustics
Read Now!

RR31. On the role of amplitude and phase in the synthesis of male and female voices. W. D. E. Verhelst and J. H. Eggen (Inst. for Perception Res./IPO, P.O. Box 513, NL 5600 MB Eindhoven, The Netherlands)

A pitch-synchronous segmentation, which was shown to be perceptually close to a deconvolution [C. Hamon *et al.*, Proc. IEEE ICASSP'89, 238–241 (1989)], was used to obtain a short-time Fourier representation of the LPC residual. After selected amplitude and phase manipulations of voiced segments, a residue was reconstructed, which was used to drive the LPC synthesis filter. Twenty utterances (ten male, ten female) were investigated under two amplitude (original/flat) and two phase conditions (original/zero), yielding four versions for each utterance. The quality of these versions was judged by 12 subjects in a paired-comparison experiment. Original amplitude information was consistently preferred over original phase information. For female voices, there were significant quality differences between any of the four versions. However, for male voices the original amplitude information alone proved to be sufficient to make the synthetic speech almost indistinguishable from natural speech. [Work was supported in part by the Dutch SPIN–ASSP program.]

RR32. A mixture modeling approach to text-independent speaker ID. Douglas A. Reynolds (Digital Signal Processing Lab., Elec. Eng. Dept., Georgia Inst. Technol., Atlanta, GA 30332), Richard C. Rose (Lincoln Lab., MIT, Lexington, MA 02173), and Mark J. T. Smith (Georgia Inst. Technol., Atlanta, GA 30332)

Automatic speaker identification (ASI) systems generally fall into two classes: text dependent and text independent. High recognition rates for short utterances (<1 s) are more common for text-dependent systems since the process benefits from the *a priori* knowledge of the underlying acoustic–phonetic stream. Without this added information, text-independent ASI usually requires long test utterances for averaging out the unknown phonetic variations. In this paper a new technique is introduced to bridge the gap between text-dependent and text-independent ASI, which allows for high recognition rates using short utterances in a text-independent ASI system. Speakers are parametrically represented by a Gaussian mixture probability density function, where the parameters are maximum likelihood estimates obtained from a form of the iterative estimate–maximize (EM) algorithm [G. J. McLachlan, *Mixture Models* (Dekker, New York, 1988)]. The components in the mixture model can be considered to represent "hidden" acoustic classes so that during testing an utterance is automatically segmented into subword units for comparison to a speaker's subword models, as in text-dependent ASI, allowing for reliable scoring over short utterances. An application of this technique is for speaker spotting in conversational utterances over long distance telephone connections, where speaker identity may change frequently in time. A preliminary study [R. C. Rose and D. A. Reynolds, "Text independent speaker identification using automatic acoustic segmentation," ICASSP-90] on a clean conversational database has shown great promise, producing an 89% recognition rate for 1-s test utterances. This new ASI technique and various channel compensation methods will be applied to a telephone database and results will be presented at the conference.

RR33. Degradation of speaker identification for LPC formant-coded speech. J. H. Eggen and L. L. M. Vogten (Inst. for Perception Res./IPO, P.O. Box 513, NL 5600 MB Eindhoven, The Netherlands)

Degradation of speaker characteristics for LPC formant-coded speech was measured with a speaker identification experiment. Spontaneous speech spoken by 14 speakers was analyzed and synthesized with two LPC speech-coding schemes using 2.4 and 12 kbit/s, respectively. PCM-coded speech (120 kbit/s) served as reference speech. Here 20 subjects had to identify the speakers by listening to the utterances. Listeners' familiarity with speakers was scaled before the experiment. The percentage that correctly identified speakers differs signficantly for the three speech types and depends on familiarity. The degradation of speaker identification can also be expressed by the index of loss in voice recognition. This index is defined as the ratio of the percentage correctly identified LPC voices and the percentage correctly identified PCM voices. The index is a function of familiarity only for 2.4 kbit/s LPC speech. If familiarity is controlled, the speaker identification test provides a suitable method for evaluating speech-coding techniques with respect to voice recognition. [Work supported by the Dutch government as part of the national SPIN program "Analysis and synthesis of speech."]

RR34. Speech preprocessing using analog VLSI. Weimin Liu, Andreas G. Andreou, and Moise H. Goldstein, Jr. (Speech Processing Lab., Elec. and Comput. Eng. Dept., Johns Hopkins Univ., Baltimore, MD 21218)

The aim of this project is analog CMOS VLSI implementation of a low-power, real-time front-end processor for speech recognition and for multichannel aids for the deaf. Physiological studies of the mammalian auditory periphery indicate that synchrony coding of speech signals by the temporal pattern of activity in the auditory nerve is noise resistant and performs amplitude compression without distortion over a wide range [M. B. Sachs, H. F. Voigt, and E. D. Young, J. Neurophysiol. **50**, 27–45 (1983)]. In a number of studies using digital computation to model the auditory periphery, the properties of synchrony coding of speech signals have been explored [K. L. Payton, J. Acoust. Soc. Am. **83**, 145–162 (1988), and articles in J. Phonet. **16**, 1 (1988)]. Here VLSI realization of the middle ear, the basilar membrane, and the hair cell and synapse are presented along with design features of the analog VLSI approach. The multichannel outputs correspond to the time-varying firing rates of neurons from discrete places on the cochlear partition.

THURSDAY MORNING, 24 MAY 1990 402–403 KELLER, 8:30 TO 11:45 A.M.

Session SS. Underwater Acoustics VII: Scattering and Ambient Noise

Suzanne T. McDaniel, Chairman
Applied Research Laboratory, Penn State University, P.O. Box 30, University Park, Pennsylvania 16804

Chairman's Introduction—8:30

Contributed Papers

8:35

SS1. Fourth moments of the acoustic wave forwardly scattered by a rough ocean surface. C. C. Yang (CSSL, Dept. of Elec. Eng., Penn State Univ., University Park, PA 16802) and S. T. McDaniel (Penn State Univ., University Park, PA 16804)

Various fourth moments of the acoustic wave field forwardly scattered by a rough ocean surface are evaluated. The fourth moment at the same position is related to the scintillation index. The fourth moments at two positions can be used for extracting the statistical information of the phase difference between two receiving points. The Kirchhoff approximation is utilized to formulate the scattered wave field. Also, the statistical fluctu-