PROGRAM OF

The 113th Meeting of the Acoustical Society of America

Hyatt Regency Hotel • Indianapolis, Indiana • 11–15 May 1987

MONDAY EVENING, 11 MAY 1987 CELEBRATION HALL, 7:00 TO 9:00 P.M.

Tutorial Lecture

Digital signal processing. Lawrence R. Rabiner (AT&T Bell Laboratories, Speech Research Department, Room 2D-533, Murray Hill, NJ 07974)

In all fields of acoustics, the response of systems of interest involves the analysis and the processing of a signal. This signal can be the radiated acoustic pressure from a hydrophone, the acceleration of a mass due to local acoustic excitation, the speech of a spoken word, or the melodious pattern of a music instrument. In each of these cases the signal has to be processed and analyzed. The field of digital signal processing is one area that has been attracting more interest over the last few years. In this tutorial the field of digital signal processing (DSP) will be explained together with the theory of linear systems, sampling, and spectrum analysis. Comparisons between digital and analog processing will be discussed. Other topics that will be presented are discrete Fourier series, sampling, aliasing, FIR and IIR filters, spectrum analysis, and fast Fourier transform. The session will conclude with detailed examples of applications of digital systems in different areas of interest such as communication in general and speech processing.

TUESDAY MORNING, 12 MAY 1987 REGENCY BALLROOM A & B, 8:15 TO 11:54 A.M.

Session A. Speech Communication I: Speech Perception

Robert A. Fox, Chairman

Department of Speech and Hearing Science, Ohio State University, Columbus, Ohio 43210

Chairman's Introduction—8:15

Contributed Papers

8:20

A1. The problem of serial order in auditory word recognition. Howard C. Nusbaum (Department of Behavioral Sciences, The University of Chicago, 5848 S. University Avenue, Chicago, IL 60637), Steven L. Greenspan (AT&T Bell Laboratories, Naperville, IL 60566), and Mathew Jensen (Department of Behavioral Sciences, The University of Chicago, Chicago, IL 60637)

Recently, there has been a resurgence of interest in parallel distributed processing models of human perception. When speech perception is modeled in this type of spatially distributed network, a problem arises in coding the temporal order of perceptual units such as phonemes or words. In general, three solutions to this problem have been proposed: First, perceptual units may be context coded such as in context-sensitive allophones. The order of units presented at different points in time can be determined by matching the context "edges" of each activated unit. Second, different portions of the network may represent different time frames. By this approach, the recognition of each successive perceptual unit activates representations in successive segments of the network. Finally, temporal order may be represented in the computational dynamics of the network. In this case, expectations about serial order are used to shift the focus of processing attention within the network. Thus, while the first two approaches recode temporal order into a spatial or spatial-like representation, the third uses a temporal representation. Each of these approaches has positive and negative attributes, the implications of which will be discussed for a neuromorphic theory of speech perception. [Work supported, in part, by NIH.]

8:32

A2. The neighborhood activation model of auditory word recognition. Paul A. Luce (Speech Research Laboratory, Department of Psychology, Indiana University, Bloomington, IN 47405)

The neighborhood activation model (NAM) of auditory word recognition describes the processes by which a stimulus word is identified in the context of phonetically similar words activated in memory. Stimulus input activates a set of acoustic–phonetic patterns in memory that must be discriminated and chosen among. These acoustic–phonetic patterns receive activation levels proportional to their similarities to the stimulus input. The activation levels may then be adjusted by biases arising from higher-level information, such as word frequency. The interaction of the bottom-up sensory input and top-down biasing information is assumed to take place within individual processing units called word decision units. These units monitor the activation levels of their acoustic–phonetic patterns, any higher-level information that may optimize decisions among the competing patterns, and the activity of all other word decision units.

1:17

II2. Vocal jitter and shimmer measurements with a personal computer. David G. Drumright (Department of Mechanical Engineering, 22 Hammond Building, Pennsylvania State University, University Park, PA 16802), J. Anthony Seikel, and Kim A. Wilcox (Department of Speech–Language–Hearing, University of Kansas, 2101 Haworth, Lawrence, KS 66045)

Vocal fundamental frequency is one of the most widely analyzed components in speech acoustics. As such, there is a need for an efficient and accurate means of producing fundamental frequency data. This paper describes a series of machine language and basic routines that yield average fundamental frequency, vocal jitter, and vocal shimmer measures with a resolution sufficient for most research purposes. The algorithm permits the user to determine major period divisions through an efficient interactive graphic process. The microcomputer is then responsible for computing accurate period and intensity data for each cycle. In the present installation, the accuracy of these procedures is limited only by the 20 000-Hz maximum sampling rate of the hardware. In addition to the $f0$ analysis routines, programs for high-speed digital sampling, waveform editing, and waveform manipulation have been developed for use with PC-compatible computers. Documentation of the installation, use, and accuracy of the routines will be provided. [Work supported by NINCDS and U. of Kansas General Research Fund.]

1:29

II3. Some results of speech analysis using a parallel image processor. John F. Hemdal (Department of Electrical Engineering, The University of Toledo, Toledo, OH 43606)

Special purpose parallel pipelined image processors are not only faster for images than general purpose computers but also allow one to investigate the applications of the latest image processing algorithms and techniques to continuous speech. This is particularly valuable when trying to automate the spectrogram reading capabilities of Victor Zue [Proc. IEEE **73**, 1602–1615 (1985)]. Several sentences of continuous speech were sampled and converted to images using the FFT. These images were analyzed on a parallel cellular image processor called the Cytocomputer™. Suitable structuring elements were selected to smooth the inherently bumpy FFT spectra and to remove noise. Several of the distinctive features of speech are described in mathematical morphology terms and these features were tracked. Specifically, stop consonants, fricatives, vowels, and nasals were located and labeled. Examples of intermediate image transformations and resultant labeling were recorded by photographic process. [Work supported by Ohio Board of Regents.]

1:41

II4. ARMA parameter estimation of speech. John J. Wygonski and Guy Sohie (Department of Electrical and Computer Engineering, Arizona State University, Tempe, AZ 85287)

Autoregressive moving average (ARMA) or pole-zero modeling techniques are known to provide better spectral estimation capabilities than autoregressive (AR) or all-pole methods such as linear prediction for nasalized speech and speech corrupted by noise. This paper describes the results of applying two ARMA modeling techniques to synthesized and natural segments of noisy and clean speech. The overdetermined normal equation (ODNE) and the extended-order singular value decomposition (SVD) methods of ARMA parameter estimation are compared to the covariance method of linear prediction and are evaluated using parametric distance measures in the cepstral domain. Use of the AR part of the ARMA model to estimate the envelope of noisy speech is demonstrated. Coefficient transformations among ARMA, AR, and cepstral parameters are also discussed.

1:53

II5. A codebook of articulatory shapes. M. M. Sondhi, J. Schroeter, and J. N. Larar (AT&T Bell Laboratories Murray Hill, NJ 07974)

The acoustical properties of a vocal tract depend on its shape which, at frequencies below about 4 kHz, is adequately described by the area function (i.e., cross-sectional area as a function of position from glottis to lips). Obtaining even this simplified description of the tract is a difficult problem. The solution that is proposed is to construct a linked codebook of area functions and corresponding LPC vectors. Then, a given LPC vector is transformed to an area function by finding the closest LPC vector in the codebook and accessing the linked area function. The authors have constructed just such a codebook (to our knowledge, the first). At present, it has about 1500 shapes clustered from a training set of about 10 000 shapes generated by a vocal tract model [P. Mermelstein, J. Acoust. Soc. Am. **53**, 1070–1082 (1973)]. Distance between area functions is defined as the Itakura distance between the corresponding LPC vectors. Handling the unique problems arising from this definition of distance and from the large size of the training set will be described. The plan for refining and enlarging the codebook will also be discussed.

2:05

II6. Statistical representation of word-initial obstruents: General considerations. Gary Weismer, Karen Forrest, and Paul Milenkovic (Speech Motor Control Laboratories, University of Wisconsin—Madison, Madison, WI 53705-2280)

The description of obstruent spectra has typically been done in categorical terms [e.g., S. E. Blumstein and K. N. Stevens, J. Acoust. Soc. Am. **66**, 1001–1017 (1979)]. Whereas this may be satisfactory for some purposes, the categorical approach is not helpful in cases where obstruent sounds are produced that are not easily described according to the normal set of segmental contrasts. In particular, many dysarthric and apraxic subjects produce stops and fricatives that give the perceptual impression of being "between" the normal segmental categories. The purpose of this paper is to describe some of the general considerations that led us to develop measures capable of indexing spectra associated with such indeterminate obstruent productions. A statistical representation of obstruent spectra will be described that is similar to certain approaches to describing a more general class of impulse noises [J. Erdreich, J. Acoust. Soc. Am. **79**, 990–998 (1986)]. [Work supported by NIH Awards NS13274 and NS20976.]

2:17

II7. Statistical representation of word-initial obstruents: Adult data. Karen Forrest, Gary Weismer, and Paul Milenkovic (Speech Motor Control Laboratories, University of Wisconsin—Madison, Madison, WI 53705-2280)

The utility of a statistical representation of word-initial obstruents (Weismer *et al.*, previous paper) was assessed for the speech of normal, young adults. Ten subjects repeated monosyllabic words, where the initial consonant was a voiced or voiceless stop or fricative, in a carrier sentence. Each place of obstruent articulation was paired with at least two vowels and each target word was repeated six times. The target words were digitized and the first four moments (mean, variance, skewness, and kurtosis) were calculated for successive 10-ms spectral slices, beginning with the onset of the obstruent and continuing through the third cycle of the vowel. Moments were calculated from linear and bark transformed spectra. Three of the moments (mean, skewness, and kurtosis) provide a good description of the shape and center of the spectral slices. These moments were plotted in three-dimensional space to determine whether a unique, three-dimensional pattern could be defined for each place of articulation. Discussion will center on the use of these moments, as they define a three-dimensional space, to quantify acoustic events associated with obstruent production. [Work supported by NIH Awards NS13274 and NS20976.]

2:29

II8. Graphics editor for speech synthesis experiments. Andrew Helck, Ian Coville, and Noriko Umeda (Institute for Speech and Language Sciences, 10 Washington Place, New York University, New York, NY 10003)