Test of an automatic syllable peak detector

M. Fleck and M. Y. Liberman

Citation: The Journal of the Acoustical Society of America **72**, S78 (1982); doi: 10.1121/1.2020070
View online: https://doi.org/10.1121/1.2020070
View Table of Contents: https://asa.scitation.org/toc/jas/72/S1
Published by the Acoustical Society of America

4:05

QQ8. Estimates of power coupling two beams. P. W. Smith, Jr. (Bolt Beranek and Newman Inc., 10 Moulton St., Cambridge, MA 02138)

Davies and Wahab have recently reported extensive precise calculations of the average power coupling the two parts of a beam held by three simple supports, averages being taken over frequency bands and an interval of the ratio of lengths of the two parts [J. Sound Vib. **77**, 311 (1981)]. Those results are discussed and compared with prior studies based on one or another statistical hypothesis, including Statistical Energy Analysis. Principal results are (1) for small loss factor η ($\eta \lesssim 1/N\pi$, where N is average resonant mode number), the new and old results agree very closely; this is the tightly coupled regime where half the power input to each subsystem is dissipated in the other; (2) agreement is less good for larger η; in this regime reverberation is slight since round-trip attenuation is large; coupling power varies with the source location within each subsystem; (3) for the particular structure, where the wave energy transmission coefficient of the coupling is large (0.5), there is no regime that is both loosely coupled and highly reverberant.

THURSDAY AFTERNOON, 11 NOVEMBER 1982 ORANGE ROOM, 1:30 TO 5:05 P.M.

Session RR. Speech Communication VI: Analysis and Synthesis

Edward P. Neuburg, Chairman

National Security Agency, Ft. Meade, Maryland 20755

Chairman's Introduction—1:30

Contributed Papers

1:35

RR1. Pitch extraction by adaptive window time averaging (AWTA). Joan E. Miller (Bell Laboratories, Murray Hill, NJ 07974)

A running time average applied to a periodic signal will produce a constant, nonvarying output if the averaging interval is equal to the period length or a multiple thereof. Hence, pitch detection may be accomplished by finding the averaging interval which minimizes the variation in the output and taking the reciprocal of this time interval as the fundamental frequency. This notion has been implemented for the extraction of pitch from speech signals. Some important features of the techniques are (1) time integration, an inherent aspect of this method, contributes to the robustness of its performance, (2) the variance of the averaged signal may be normalized, which facilitates a voiced–unvoiced decision, and (3) crucial parameters of the analysis can be easily adapted to the current pitch estimate for better results in sections of changing period length. Details of the procedure will be described, and results obtained for a variety of speech samples will be shown.

1:49

RR2. On reducing pitch sensitivity of LPC parameters. Sharad Singhal and Bishnu S. Atal (Acoustics Research Department, Bell Laboratories, Murray Hill, NJ 07974)

Mean-squared prediction error criterion used in linear predictive coding of speech has a number of inherent shortcomings. For low pitch frequencies, the LPC analysis largely ignores the pitch-related fine structure of the speech spectrum. However, the sensitivity of the LPC parameters increases rapidly as a function of the pitch frequency. This sensitivity can be directly traced to the mean-squared prediction error criterion used in LPC analysis. Moreover, in the presence of noise and errors resulting from the assumption of an all-pole model, the minimization of prediction error leads to a perceptually suboptimal solution. In this paper, the LPC analysis is considered as a short-time spectral envelope matching problem. Using the LPC-derived parameters as initial values, a search procedure is used to refine the LPC parameter estimates. The new procedure minimizes a perceptual distance metric between the spectrum based on LPC parameters and the samples of the speech spectrum at spectral peaks.

2:03

RR3. Performance evaluation of a real-time LPC coding technique. N. Rao Vemula and Phil T. McLaughlin (General Instrument, Microelectronics, 600 W. John Street, Hicksville, NY 11802)

There are a number of LPC-based speech synthesis integrated circuits available on the market today. For many practical applications of these chips, it is advantageous to utilize a real-time code generation system. Traditional LPC analysis techniques, such as covariance, autocorrelation, and PARCOR, process blocks of speech samples yielding one set of LPC parameters for each block. Autocorrelation and covariance techniques require matrix inversion and are too complex for hardware implementation. PARCOR-type lattice techniques require a large amount of data storage at each stage of the lattice. In this study we have evaluated a fast and efficient LPC analysis technique suitable for hardware implementation for real-time operation. Unlike the traditional LPC analysis techniques, the real-time technique processes individual samples of speech. The LPC parameters are first initialized and then updated as each sample is processed. Computer simulation has shown that the real-time technique yields reasonable spectra. Computer generated synthetic sounds from the real-time technique have almost the same quality as those generated from the block-data techniques. Sounds synthesized using the real-time technique on our SP0256 chip are indistinguishable from those synthesized using traditional techniques.

2:17

RR4. Test of an automatic syllable peak detector. M. Fleck (Yale University, New Haven, CT and Bell Laboratories, Murray Hill, NJ 07974) and M. Y. Liberman (Bell Laboratories, Murray Hill, NJ 07974)

It is a venerable and commonplace intuition that syllables are "peaks of sonority." Previous work by Mermelstein [J. Acoust. Soc. Am. **58**, 880–883 (1975)] suggests that this intuition is nearly true on the acoustic surface of speech, and that relatively crude algorithms based on amplitude variation can work fairly well at counting syllables and locating their approximate extent. We have devised a simple speaker-independent algorithm for detecting sonority peaks in continuous speech, and tested its accuracy as a syllable finder on several paragraphs of casual reading by four speakers. Performance was encouraging: less than 10% misses, and less than 1% false triggers. The misses were about evenly divided among (1) vowel–vowel transition, where no sonority minimum is expected, (2) intervocalic sonorants with insufficient amplitude dip, and (3) devoiced or deleted schwas. Methods for dealing with these cases will be discussed.

2:31

RR5. Articulatory constraints on vocal tract area functions and their acoustic implications. S. J. Butler and H. Wakita (Speech Technology Laboratory, 3888 State St., Santa Barbara, CA 93105)

A mathematical model for generating vocal tract area functions for vowel-like sounds has been developed. Using computer simulation of a transmission line analog of the vocal tract, the relationship between acoustic features (in this case formants) and parameters of the model (i.e., articulatory parameters) has been investigated in order to develop improved algorithms for acoustic–articulatory and articulatory–acoustic transformations. The model has been found to be a useful tool for study of acoustic phonetics, examination of the applicability of speaker normalization techniques, and for the refinement of high-quality speech synthesis. Extensions of the model to include consonantal vocal tract configurations are also presented.

2:45

RR6. The dialogue terminal. Ron Sorace (Acoustics Research Department, Bell Laboratories, Murray Hill, NJ 07974)

The dialogue terminal is a direct application of research in speech recognition, synthesis, and coding at Bell Laboratories. It represents a rather novel approach to state-of-the-art interactions between man and machine and has potential for a variety of applications. Current speech coding techniques permit integration of voice and data for digital tranmission and, coupled with high density memory circuits, enable storage of large amounts of speech for playback, synthesis, or recognition. Speech storage and editing and rate modification are necessary for applications such as voice-mail. Speech synthesis capability such as formant or speech-from-text techniques are available and have sufficient quality to provide a degree of intelligibility which is adequate for output of data and issuing spoken prompts. The terminal is designed to perform isolated word recognition since this method represents the state-of-the-art. However, for the more experienced user the capability to parse a continuous sequence of commands will give the terminal the ability to adapt to the experience level of the user. The resulting mode of operation consists of synthetic voice prompts and voice commands in a dialogue format.

2:59

RR7. Synthesis of continuous speech by concatenation of isolated words. Mark A. Randolph and Victor W. Zue (Room 36-541, Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, MA 02139)

This paper reports a feasibility study of synthesizing continuous speech by concatentation of modified isolated word templates. Continuous speech obtained simply by word template concatenation has several inherent problems. It cannot account for the prosodic features normally found in continuous speech, nor can it account for coarticulation; the natural transitions that occur at word boundaries. In an effort to alleviate the first problems, the synthesizer is provided with information describing the duration, the fundamental frequency ($f0$) contour, and the energy contour of each sentence. The synthesizer draws from a dictionary of word templates each stored as a sequence of LPC parameters. Before concatenation, the time scales of the synthesis templates are nonlinearly warped using the alignment path obtained from a level building connected speech recognition algorithm. In addition, the $f0$ and energy contours of the word templates are modified. To account for coarticulation, the various parameters are smoothed at word boundaries. The goal of this synthesis system is an output information rate of 200 bits per second of speech. A demonstration tape will be played.

3:13

RR8. Implementation of a prosody scheme in a constructive synthesis environment. Kathleen M. Goudie, Kun-Shan Lin, and Gene A. Frantz (Consumer Products Group, Texas Instruments, P.O. Box 10508, M.S. 5893, Lubbock, TX 79408)

In a constructive synthesis environment, lack of natural prosody detracts both from the naturalness and from the intelligibility of synthetic speech. It is time consuming and frustrating for the linguistically untrained user of a constructive-synthesis system to assign pitch patterns and unit durations by mere guesswork. A semi-automatic prosody assignment scheme has been designed here which takes the majority of the prosody-assignment burden off the user and which is capable of assigning fairly natural-sounding pitch patterns to a constructive synthesis string of allophones. The user is required only to mark the primary and secondary stress locations in the text or in the allophone string, plus an indication of whether the constructed phrase should rise or fall in pitch at the end. These stress marks serve as anchor points for the system to compute both an intonation contour and a timing-adjustment contour for the whole phrase. Thus, with a minimum of effort and linguistic knowledge, the user may achieve a relatively natural-sounding constructive speech phrase.

3:27

RR9. Speech synthesis using allophones. Janet G. May and Eugene H. Lee (General Instrument Corporation, Microelectronics Division, 600 W. John Street, Hicksville, NY 11802)

Synthesizing speech by concatenating allophones is a new application of linear predictive coding (LPC) that provides an unlimited vocabulary. Since the stored units are individual speech sounds, any English word can be synthesized by concatenating the appropriate sounds. Our inventory contains 59 speech sounds and five pauses. Each allophone was created by extracting it from the digital waveform of a word, and then synthesizing it. The sounds are called allophones because some phonemes have two or three versions for different environments. For example, /k/ and /g/ each have three allophones: one to be used before front vowels, one before back vowels, and one in final position. In contrast, research revealed that the same /p/ can be used in all environments. Isolating final /r/ from preceding vowels was impossible because of important transitional information. Therefore there are five phonemically unique V + /r/ combinations. Due to the omission of many transitions, allophone synthesis suffers from reduced intelligibility when compared with the technique of synthesizing words as units. This compromise seems minor, however, when its flexibility and low cost are considered.

3:41

RR10. Formant synthesis: Technique to account for source/tract interaction. J. J. Yea and D. G. Childers (Department of Electrical Engineering, University of Florida, Gainesville, FL 32611)

Research in speech synthesis is presently concerned with methods for improving intelligibility and naturalness. But most work to date has dealt with intelligibility, i.e., the synthesized speech is intelligible but monotone and machinelike in quality. Although this type of speech is useful in some applications, the users of speech synthesizers are demanding a more natural sounding synthesized speech. We are addressing this issue in our research. In particular we are studying the influence of the glottal source function on the production of synthesized speech. Our data base consists of inverse filtered speech, ultra-high-speed laryngeal films, and the electroglottograph waveforms, all temporally synchronized in our experiments. We use the experimentally derived source waveforms obtained from these various method as the excitation for a serial/parallel Klatt formant synthesizer to synthesize sentences. We then evaluate the naturalness of this synthetic speech by listening tests. Our goal is to rank order the contribution of different source parameters to the naturalness of synthetic speech. More specifically, we report on our results to date of the effect of source–tract interaction upon the production of natural sounding speech.