Proceedings of The Institute of Acoustics

SPEECH PROSODICS AND SPEECH UNDERSTANDING SYSTEMS

RUSSELL R. FEARS

TEESSIDE POLYTECHNIC, COMPUTER SCIENCE DEPARTMENT

Introduction

The term 'prosodic feature' refers to those phonological features of English which have an essentially variable relationship to the words to which they apply. Prosodic features include pitch, loudness, duration, silence, etc.

Speech understanding systems are defined in Newell (1973, ref 18), and tend to consist of several modules which apply knowledge of a particular subset of linguistics or phonetics to the speech signal. Systems based on the 'Newell definition' tend to have ignored prosodics in the past, this paper reviews some of the current trends in prosodic research with applications to S.U.S.s.

Prosodic features

Crystal (1969, ref 2) refers to prosodic features as those non-segmental characteristics of speech referable to variations in pitch, loudness, duration and silence. This paper concentrates primarily on pitch, since this is one of the most easily identifiable of the prosodic features. Pitch has also been shown to be a valuable indicator of the utterance's syntactic structure. It is worth citing work such as Lieberman (1967, ref 13) which showed that the interval between vowel centres is a reliable cue for estimating disjuncture positions (as in "light housekeeper" and "lighthouse keeper").

Pitch detection

Trager and Smith (1951, ref 23) state that monitoring pitch allows one to establish procedures for the recognition of parts of speech syntax. The work of Lea, Medress and Skinner (1972, 1973a, 1973b, refs 9, 10, 11) has helped to formalise these procedures. There are many pitch extraction algorithms, which fall into two broad categories. Frequency-domain algorithms are inherently accurate but tend to be slow as the computation of the necessary spectra is a time-consuming business (unless special-purpose hardware is available). Time domain procedures can run in real-time, but tend to be inaccurate unless the segment under consideration is heavily voiced.

time domain algorithms: Gold (1962, ref 4), Reddy (1967, ref 21) and Miller (1974, ref 17) have all published algorithms based on time-domain methods. Miller's algorithm, for instance, runs in real-time on a medium speed computer at a sampling rate of 20 kHz.

frequency domain algorithms: are based on the identification of peaks in frequency spectra. Fast Fourier Transform spectra are used by Aschkenasy, Weiss and Parsons (1974, ref 1), and Harris and Weiss (1963, ref 5).

Markel has published a number of papers on algorithms based on linear prediction (Markel 1972a, ref 14; Markel 1972b, ref 15; Markel and Gray 1976, ref 16). Noll (1964, ref 19, and 1968, ref 20) uses the cepstrum representation. Autocorrelation analysis is used by Gillman (1975, ref 3), and Lea, Medress and Skinner (1973a, ref 10).

Hess (1974, ref 6) proposes a hardware device based on an inverse filter formulation.

Use of prosodics - the prosodic knowledge component of a S.U.S.

Ever since 1972 Lea, Medress and Skinner at Sperry Univac have been investigating prosodic aids to speech recognition, under a number of ARPA contracts. In the fifth report (Lea 1974, ref 7) gives a summary of results to date in which he suggests that

> "vital assumptions of a prosodically guided approach to speech understanding have been verified from a variety of experiments. In particular, stressed syllables have been shown to be of prime importance in speech recognition, because of
>
> a) the occurrence of stressed syllables in semantically important words,
> b) the close correspondence between detected phonetic structure and underlying phonemic structures in stressed syllables,
> c) the much higher reliability of phonetic classification possible in stressed syllables, and
> d) the vital cues to syntactic structure that stressed syllables provide.

From the study of a variety of speech texts, the Univac team have demonstrated that over 90% of all intuitively-predicted syntactic boundaries are detected from substantial fall-rise valleys in pitch contours. They have shown that over 85% of all syllables perceived as stressed are located by a particular combination of energy duration and pitch contour cues. Lea, Medress and Skinner (1975, ref 12) propose a 'prosodically guided speech understanding strategy' which would govern the construction of a S.U.S. They argue that successful understanding of spoken sentences involves early use of linguistic structure in combination with the most reliable acoustic information, and base their system on the early isolation of syntactic structure using prosodic information. The system is an analysis - by-synthesis hypothesis-driven speech understander where the hypothesisation process is driven by prosodic cues as well as the more usual phonetic cues. Indeed preliminary syntactic hypotheses are generated directly from prosodic cues derived from an analysis of energy functions, fundamental frequency and voicing functions. The preliminary hypothesis is then 'filled' by a lexical hypothesiser which uses information both from the prosodic structure analysis and the phonetic parametrisation. Once a total hypothesis is built up, generative phonological rules are applied to yield an acoustic pattern which is then matched to the input acoustic phonetic pattern.

In 1976 Lea (1976, ref 8) presented the results of various experiments which attempted to apply the Univac hypothesisation process to the 'Hear What I Mean' speech understander under development at Bolt, Beranek and Newman (Woods, et.al.

1976, ref 24). Prosodic information was encoded into the BBN Augmented Transition Network grammar by specially marking state-transition arcs in the ATN if they are expected to be immediately preceded by intonationally-detected phrase boundaries. These boundaries are detected in the utterance by applying fundamental frequency fall-rise detection algorithms. Scores are increased in expected boundaries are found, decreased if they are not. The BBN Final Report (Woods et.al. 1976, ref 24) notes that, with relation to prosodics

> "time did not permit us to carry the work on the prosodics component as far as testing its effectiveness in aiding the speech understanding process."

Work at Teesside Polytechnic

A suite of programs for data acquisition and handling, and a pitch tracker based on Gillmann's (1975, ref 3) autocorrelation algorithm are being implemented at Teesside Polytechnic on Data General NOVA computers. The system is still being tested with simulated data and with the sentences of the IEEE Speech Data Base.

Gillmann's algorithm works on down-sampled and filtered speech using a centre-clipping autocorrelation method. Sondhi (1968, ref 22) showed that centre-clipping enhances peaks in the autocorrelation spectrum, and peaks found in the autocorrelation spectrum computed from the end-off formulation

$$A(j) = \left(\sum_{i=1}^{N-j} C_i * C_{i+j}\right)/(N-j)$$

are tested as possible pitch peaks, to produce a string of raw pitch values for each 10 msec segment of the utterance. A heuristic editing routine using a three-point median smoother attempts to correct any dubious values.

Sampling is at 2 kHz, and the signal is low-pass filtered before sampling. Display routines allow the display of pitch data, as well as the original signal, and the signal can be replayed through an amplifier and speaker.

Much of the work done by Lea and his colleagues is based on read text. I want to test the difference in pitch contours between a spoken utterance and a reading of a transcript of that same utterance. I also want to investigate the relationship, if any, between pitch contour and the kinds of interjections such as 'er' and 'ah' which frequently occur in normal conversational speech, with a view to providing, if possible, a set of rules for the elimination, in a speech understanding system, of possible confusions caused by these 'words'.

References

(1) E.ASCHKENASY, M.R. WEISS and T.W. PARSONS 1974 Proc. IEE Symp. on Speech Recognition, 208-209. "Determining Pitch from Fine-Resolution Spectrograms"

(2) D. CRYSTAL 1969 Cambridge University Press "Prosodic Systems and Intonation in English"
(3) R.A. GILLMANN 1975 J.A.S.A 58(1) presented at the ASA meeting, San Fransisco, Nov 1975. "A Fast Frequency-Domain Pitch Algorithm"
(4) B. GOLD 1962 J.A.S.A 34, 916-921 "A Computer Program for Pitch Extraction"
(5) C.M. HARRIS and M.R. WEISS 1963 J.A.S.A. 35, 339-343. "Pitch Extraction by Computer Processing of High-Resolution Fourier Analysis Data"
(6) W.J. HESS 1974. Proc Intnl. Zurich Seminar of Dig. Comm. paper A5. "On-Line Digital Pitch Period Extractor for Speech Signals"
(7) W.A. LEA 1974. Univac Report PX 11087, 31st Oct 1974. "Prosodic Aids to Speech Recognition V: A Summary of Results to Date"
(8) W.A. LEA 1976. Presented at the 92nd meeting, ASA, California 1976 "Use of Intonational Phrase Boundaries to Select Syntactic Hypotheses in a Speech Understanding System"
(9) W.A. LEA, M.F. MEDRESS, and T.E. SKINNER 1972. Univac Report PX 7496, "Prosodic Aids to Speech Recognition I: Basic Algorithms and Stress Studies"
(10) W.A. LEA, M.F. MEDRESS and T.E. SKINNER 1973a. Univac Report PX 10232, "Prosodic Aids to Speech Recognition II: Syntactic Segmentation and Stressed Syllable Location"
(11) W.A. LEA, M.F. MEDRESS and T.E. SKINNER 1973b. Univac Report PX 10430, "Prosodic Aids to Speech Recognition III: Relationships between Stress and Phonemic Recognition Results"
(12) W.A. LEA, M.F. MEDRESS and T.E. SKINNER 1975. IEEE Trans. ASSP-23, 1. "A Prosodically Guided Speech Understanding Strategy"
(13) P. LIEBERMAN 1967 M.I.T. Press "Intonation, Perception and Language"
(14) J.D. MARKEL 1972a IEEE Trans. AU-20,5. "The SIFT algorithm for Fundamental Frequency Estimation"
(15) J.D. MARKEL 1972b Conf. Rec., 1972 Intl. Conf. Speech Comm. & Proc. paper B9 81-84. "Automatic Formant and Fundamental Frequency Extraction from a Digital Inverse Filter Formulation"
(16) J.D. MARKEL and A.H. GRAY 1976 Springer-Verlag "Linear Prediction of Speech"
(17) N.J. MILLER 1974. Proc. IEEE Symp. on Speech Recognition 122-130, "Pitch Detection by Data Reduction"
(18) A. NEWELL, J. BARNETT, J.W. FORGIE, C. GREEN, D. KLATT, J.C.R. LICKLIDER, J. MUNSON, D.R. REDDY, W.A. WOODS 1973 North Holland/Elsevier "Speech Understanding Systems - Final Report of a Study Group"
(19) A.M. NOLL 1964 J.A.S.A. 36, 296-302. "Short-Time Spectrum and Cepstrum Techniques for Vocal-Pitch Detection"
(20) A.M. NOLL 1968 J.A.S.A. 43, 293-302. "Cepstrum Pitch Detection"
(21) D.R. REDDY 1967 Comm ACM 10, 343-348. "Pitch Period Determination of Speech Sounds"
(22) M.M. SONDHI 1968 IEEE Trans AU-16 262-266 "New Methods of Pitch Extraction"
(23) G.L. TRAGER and H.L. SMITH 1951. Battenburg Press, Studies in Linguistics occasional papers 3 "An Outline of English Structure"
(24) W.A. WOODS, M. BATES, G.BROWN, B. BRUCE, C. COOK, J. KLOVSTAD, J. MAKHOUL, B. NASH-WEBBER, R. SCHWARTZ, J. WOLF, V. ZUE 1976. Bolt Beranek and Newmann Report 3438 "Speech Understanding Systems Final Technical Progress Report 30th Oct 1974 to 29th Oct 1976".