Linear predictive residual analysis compared to bandpass filtering for automatic speech recognition

G. M. White

Citation: The Journal of the Acoustical Society of America **58**, S106 (1975); doi: 10.1121/1.2001826
View online: https://doi.org/10.1121/1.2001826
View Table of Contents: https://asa.scitation.org/toc/jas/58/S1
Published by the Acoustical Society of America

tion algorithm, similar to the one used in the isolated digit recognition work described by Sambur and Rabiner, is used to classify the individual digits in the utterance. Experiments with the system using ten speakers (five male, five female) in a fairly low noise environment yielded a 91% correct digit recognition score. Similar experiments using ten new speakers (five male, five female) in a noisy computer room yielded an 87% correct digit recognition score.

10:00

ZZ6. Linear predictive residual analysis compared to bandpass filtering for automatic speech recognition. G. M. White (Xerox Palo Alto Research Center, 3180 Porter Drive, Palo Alto, CA 94304)

It has been recently proposed by Itakura [F. Itakura, "Minimum Predictive Residual Principal Applied to Speech Recognition," IEEE Symp. Speech Recog. CMU (1974)] that the linear predictive residual can be used as a measure of speech waveform similarity. To measure the similarity between two waveforms, Itakura proposed to construct a linear predictive filter for one waveform and measure the residual (predictive error) for the other waveform. Itakura used this technique to achieve some remarkably good speech recognition scores. We constructed a speech recognition system using both bandpass filtering and linear prediction in order to compare the two techniques. The classifier used dynamic programming. A 36-word vocabulary was used consisting of the alphabet plus digits spoken five times by the same speaker. A single word list was used for training and the other four were used for testing. Speech input was through a noise cancelling microphone. For the digital linear predictive, inverse filtering, analysis, speech was low pass filtered at about 5 kHz and digitized at 10 kHz. For the bandpass filtering experiment, 21 filter channels each 1/3 octave wide were used covering the audio spectrum from about 100 Hz to 10 kHz. The recognition scores in both cases were 98% correct showing that the linear predictive residual technique is essentially equivalent to bandpass filtering as a means of measuring speech waveform similarity.

10:12

ZZ7. Partial word boundary detection from stress contours. D. C. Sargent (Rt. 4, Box 133 B, Pittsboro, NC 27312)

A machine algorithm was developed for partial word boundary detection in continuous speech. Word boundary detection was achieved by comparing a computer-extracted stress contour for the 700 syllable test passage with that contour which would have been predicted from rigid adherence to the Alternating Stress Rule. Since this rule functioned only at the word level and below, it was more likely to be violated when crossing word boundaries within a word. The position of any Alternating Stress Rule violation in the extracted stress contour was therefore marked as a probable word boundary location. Utilizing this concept, 44% of the word boundaries in the test passages were correctly positioned with a false alarm rate of less than 10%. Most of the false alarms were caused by the presence of adjacent reduced syllables within the same word. Research is presently being conducted to incorporate additional regularities in the stress patterns of English to further improve the algorithm's performance.

10:24

ZZ8. Algorithm to detect the beginning and end points of a speech utterance. K. Ganesan and W. C. Lin (Department of Computing and Information Sciences, Case Western Reserve University, Cleveland, OH 44106)

There is a great need to detect the beginning and end points of a speech utterance in applications like speech recognition and speaker identification. In this paper, we present a method for beginning and end-point detection which makes use of the maximum likelihood principle. The features that are used by the algorithm are (1) total per-unit energy, (2) zero-crossing rate, and (3) absolute amplitude of the speech samples. Conditional probability densities are estimated for these three features using a database of 60 phonetically balanced words and ten phonetically balanced sentences spoken by four male speakers with General American accents. A set of optimum thresholds are obtained for each feature such that the probability of classification error is minimized. The algorithm was tested for both isolated words and sentences over a population of six speakers and an error rate of nearly 0% was observed.

10:36

ZZ9. On-line, adaptive speaker-independent word recognition system based on phonetic recognition techniques. W. C. Lin and K. Ganesan (Department of Computing and Information Sciences, Case Western Reserve University, Cleveland, OH 44106)

The research reported in this paper deals with a new method of phonemic analysis of speech by statistical pattern recognition techniques and its application to the problem of Automatic Speech Recognition (ASR). An on-line, adaptive, trainable speaker-independent system is implemented using this approach. The details of the system follow: first, the beginning and end points of the speech utterance are detected. The utterance is then sent for automatic segmentation where it is segmented into the following classes: (1) voiced, (2) unvoiced, (3) transition, and (4) silence. An 11-dimensional feature vector consisting of 10 linear predictor coefficients and zero-crossing rate is extracted from these regions. For voiced and transition region, the feature extraction is done pitch synchronously and for unvoiced regions, a constant frame of 6.4 msec is used. A new phonetic unit called *phoneme-pair* is defined for the transition regions, while the unvoiced and voiced regions are represented using the phonemes of the IPA. Conditional probability densities for each of the phonemes and phoneme-pairs are estimated using non-parametric methods as a single polynomial in the 11-dimensional space. The classifier makes Bayes' minimum risk decision based on these probability densities. The recognition results of the ASR system are Training Set: 98.4%, Test Set: 96.0% (for speakers in the training set) and 91.0% (for speakers not in the training set). The present vocabulary of the system is 60 words and any new word can be added by entering its corresponding phonetic transcription. The adaptive and trainable characteristics of the system will also be demonstrated.

10:48

ZZ10. On the similarity of noisy phonetic strings produced by different words. James K. Baker (IBM Thomas J. Watson Research Center, P. O. Box 218, Yorktown Heights, NY 10598)

In a speech recognition system with an acoustic processor which attempts to automatically estimate a phonetic transcription, it is necessary to know the similarity of the probability distributions of phonetic strings when different words are spoken and input to the acoustic processor. Let $a \epsilon A, a = a_1 a_2 a_3 \ldots a_n$ represent an arbitrary phonetic string. Define the similarity between the words W_1 and W_2 by

$$S = \sum_a \Pr(a \mid W_1) \Pr(a \mid W_2).$$

The number of terms in the sum defining S grows exponentially with the length of the words W_1 and W_2. However, if the nodes of the phonological graphs for W_1 and W_2 are properly ordered, S can be calculated inductively by a generalization of the computations used in modeling a probabilistic function of a Markov process. The number of computations is approximately the product of the number of arcs in W_1 times the number of arcs in W_2.