On the performance of isolated word speech recognizers using vector quantization and temporal energy contours

L. R. Rabiner, K. C. Pan and F. K. Soong

Citation: The Journal of the Acoustical Society of America **75**, S93 (1984); doi: 10.1121/1.2021689

View online: https://doi.org/10.1121/1.2021689

View Table of Contents: https://asa.scitation.org/toc/jas/75/S1

Published by the Acoustical Society of America

THURSDAY AFTERNOON, 10 MAY 1984 YORK HALL, 2:00 TO 3:20 P.M.

Session WW. Speech Communication X: Speech Recognition

J. G. Wilpon, Chairman

Acoustics Research Department, AT&T Bell Laboratories, Murray Hill, New Jersey 07974

Chairman's Introduction—2:00

Contributed Papers

2:05

WW1. On the performance of isolated word speech recognizers using vector quantization and temporal energy contours. L. R. Rabiner, K. C. Pan, and F. K. Soong (AT&T Bell Laboratories, Acoustics Research Department, Murray Hill, NJ 07974)

The technique of vector quantization has been widely applied in the area of speech coding and has recently been introduced into the area of speech recognition. For the conventional statistical pattern recognition word recognizer using LPC feature sets as the analysis frames, the use of vector quantization leads to a large reduction in computation for the dynamic time warping pattern matching, and a concomittant small increase in average word error rate. A second technique that has been recommended for improving the performance of isolated word recognizers is the addition of temporal energy information into the distance metric for comparing frames of speech. It has been shown that the information in the prosodic energy contour complements the segmental information of the LPC spectrum, thereby providing small but consistent improvements in performance for small word vocabularies. In this talk we present results of a series of speaker independent, isolated word recognition tests using a 10-word digits vocabulary and a 129-word airlines vocabulary. We show the effects, on recognition accuracy, of adding both vector quantization and temporal energy in various combinations, to the recognition paradigm.

2:20

WW2. Global spectrum vowel recognition and human performance. Maxine Eskanazi (LIMSI-CNRS, BP30, 91406 Orsay, Cedex, France)

We have shown [Eskenazi and Lienard, J. Acoust. Soc. Am. Suppl. 1 **73**, S87 (1983)] that global characterizations of the French oral and nasal vowels in a speaker-independent automatic recognition task give generally better recognition results than formant-based methods. In particular, a very rough representation in the frequency domain, characterizing the curvature of the spectrum gave good results using very little reference information for each vowel. By now, changing the analysis that the curvature characterization is based on from an FFT to an LPC has significantly improved global results. This is very close to human intelligibility of the same databases used, in terms of distance between confusion matrices. We shall compare results of human and automatic recognition in order to better evaluate machine performance. There follows a comparison between the FFT and LPC results in order to estimate the pertinence of the information furnished by each in view of vowel recognition and in the light of the problems inherent in a speaker-independent task as well as in specific vowel properties.

2:35

WW3. A modified *K*-means clustering algorithm for use in speaker-independent isolated word recognition. J. G. Wilpon and L. R. Rabiner (Acoustics Research Department, AT&T Bell Laboratories, Murray Hill, NJ 07974)

Recent studies of isolated word recognition systems have shown that a set of carefully chosen templates can be used to bring the performance of speaker-independent systems up to that of systems trained to the individual speaker. The earliest work in this area used a sophisticated set of pattern recognition algorithms in a human-interactive mode to create the set of templates (multiple patterns) for each word in the vocabulary. Not only was this procedure time consuming but it was impossible to reproduce exactly, because it was highly dependent on decisions made by the experimenter. Subsequent work led to an automatic clustering procedure which, given only a set of clustering parameters, clustered tokens with the same performance as the previously developed supervised algorithms. The one drawback of the automatic procedure was that the specification of the input parameter set was found to be somewhat dependent on the vocabulary type and size of population to be clustered. Since the user of such a statistical clustering algorithm could not be expected, in general, to know how to choose the word clustering parameters, even this automatic clustering algorithm was not appropriate for a completely general word recognition system. It is the purpose of this paper to present a new clustering algorithm based on a *K*-means approach which requires no user parameter specification. Experimental data show that this new algorithm performs as well or better than the previously used clustering techniques when tested as part of a speaker independent isolated word recognition system.

2:50

WW4. Use of synthetic speech parameters to estimate success of word recognition. John J. Ohala, Mariscela Amador, Lynn Araujo, Steve Pearson, and Margot Peet (Phonology Laboratory, Department of Linguistics, University of California, Berkeley, CA 94720)

In an automatic speech recognition task, it would be good to estimate beforehand how recognizable the target vocabulary will be. Tests which involve numbers of human speakers using the ASR device are expensive and time consuming. Using stimuli synthesized by rule, although not without some drawbacks, would be cheaper and quicker. As a first step towards this latter goal we attempted to find out whether measures derived from rule-synthesized words would predict *human* listeners' performance in recognizing target words in continuous speech embedded in noise. Even a very simple measure of word detectability (essentially the length of the word's trajectory through the space whose dimensions are normalized $F1$, $F2$, $F3$, and rms amplitude) correlated significantly with listener's performance ($r = 0.53$). The results of refined measures of detectability and the results of word confusability will be reported.

3:05

WW5. A speaker-independent isolated word recognition board. S. Kabasawa (Matsushita Electric Industrial Co., Ltd. Central Research Lab., Moriguchi Osaka 570, Japan), M. S. Hsieh, S. M. Chang, and C. H. Lin (Matsushita Electric Institute of Technology (Taipei) Co., Ltd., Republic of China)

A one-boarded, out-performance speech recognition system for speaker-independent isolated words has been developed. This board consists mainly of the 4-bit, 1-chip microcomputer and newly developed speech recognition LSI [Ohga *et al.*, IEEE Trans. Consumer Electron. **CE-28**, 263–270 (1982)]. The board takes multiple templates and the KNN rules to cover various tokens of many persons. A new clustering method CLS (clustering with shared group) is based on the conception of the SNN method and the *K*-means iteration procedure [Rabiner *et al.*, IEEE Trans. Acoust. Speech Signal Process. **ASSP-27**, 336–349 (1979)]. The experimental results showed more than 95% recognition accuracy not only for ten Japanese city names but ten Chinese city names. CLS can be easily implemented on the 16-bit microcomputer system and it takes less than 20 min to make seven templates with the microcomputer system.