A METHOD TO EVALUATE THE PRE-PROCESSING STAGE OF ISOLATED-WORD RECOGNITION SYSTEMS

Hiroaki Oasa and Michael Wagner
Department of Computer Science
University College
University of New South Wales

ABSTRACT - This paper describes a cluster analysis method developed to evaluate pre-processors for isolated-word recognition. The method derives various statistical measures on inter-word and intra-word distances, whose ratios are analysed to evaluate the relative effectiveness of the pre-processors.

The study uses a speech database comprising (a) the 36-word set of alphabetic letters and digits and (b) a phonetically balanced set of 36 CVC words, and evaluates three pre-processors which extract FFT log-power spectrum coefficients, linear-prediction based cepstral coefficients, and critical band energies from the speech signal.

INTRODUCTION

Evaluating an automatic speech recognition system by stating the overall recognition rate of the system for certain vocabularies, speaker modes and syntactic support is often unsatisfactory. This stems from the fact that the overall recognition rate is a highly non-linear function of the distribution of the test and reference tokens in the chosen parameter space. As such the recognition rate often provides little enlightenment as to why a system performs as well or as badly as it does.

A better approach to evaluate automatic speech recognition systems is the separate examination of the major modules of the system, for example the parameter or feature extraction module, the pattern recognition module or the syntax, semantics or pragmatics modules. For each module to be thus examined it is necessary to derive suitable measures against which the module can be evaluated in order (1) to provide developers with more useful information on the strengths and weaknesses of the system and (2) to give potential users of a speech recognition system a robust and detailed yardstick to compare between systems.

This study presents an evaluation of the pre-processing stage of an isolated word recognition system. In any pattern recognition system it is important that a suitable set of parameters or features is extracted from the signal in order to provide a good separation of the classes of tokens which are present in the signal and which are to be distinguished by the recognition system. The study therefore proposes a number of statistical measures which relate to the clustering power of the pre-processing module, that is the ability of the system to cluster tokens of the same class closely together while separating clusters of tokens from one class as widely as possible from clusters of tokens of another class.

As the automatic speech recognition system in question is a speaker-dependent isolated-word recognition system, tokens are isolated words from a given vocabulary which are spoken by a given speaker. There are as many classes of tokens, and therefore clusters of tokens, as they are words in the vocabulary and each cluster contains as many tokens as repetitions of the word were recorded by the speaker.

This study introduces a number of statistical measures which are closely related to the well-known F-ratio, that is the ratio of inter-cluster over intra-cluster variance. These statistical measures are then used to evaluate three different pre-processors, namely, the log power spectrum, the LPC cepstrum and the critical band spectrum.

EXPERIMENTAL PROCEDURE

Figure 1 illustrates the experimental procedure. The speech data used in this study comprises two vocabularies recorded by five male and five female native speakers of Australian English. The first vocabulary is the 36-word set of alphabetic letters and digits (the alphadigit or AD set), which has been subject to extensive research before and is alleged (e.g., O'Shaughnessy 1987, p.415) to contain highly confusable subsets, e.g., the E-set, the A-set and the EH-set. The other vocabulary is a set of 36 common CVC words chosen to provide for phonetic diversity and balance. The vocabulary items were carefully selected from a larger list (Clark and Fraser 1982) and represents a phonetically richer (PR) set.

Six recordings were made of each vocabulary on three separate days, using a SONY PCM-2000 digital audio tape recorder. The data were low-pass filtered at 7.6 kHz, 12-bit digitised at 16 kilo-samples per second, and segmented automatically (Rabiner and Sambur 1975) with manual correction.

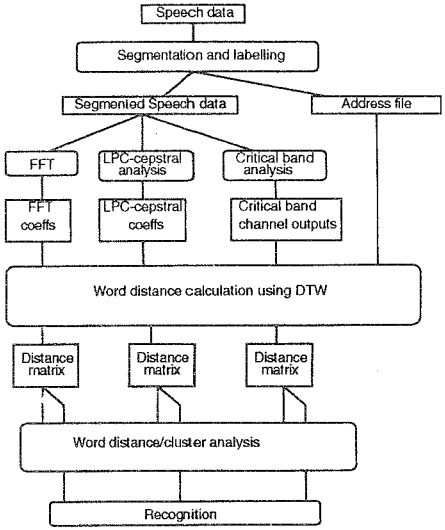

Figure 1. The experimental procedure

Three pre-processors were evaluated using the present method. They are (a) 129 log-power-spectrum coefficients from an FFT analysis, (b) 18 linear-prediction based cepstral coefficients and (c) a critical-band pre-processor which performs a non-linear frequency transformation (Zwicker 1961) and returns energies from 42 non-overlapping critical bands of width 0.5 Bark.

Distances were computed between all 216 x 216 word pairs of each speaker using dynamic time warping and a Euclidean distance measure. For each speaker we thus obtained a distance matrix of size 216 x 216.

THE EVALUATION METHOD

The method involves the derivation of the following sets of statistical measures. The first set involves the ratio of the intra- and inter-cluster medians, and the overall ratio (2 and 3 below) can be regarded as closely related to the F-ratio.

(1) Inter-cluster distance median / intra-cluster distance median (calculated for each cluster)
(2) Median of all inter-cluster distances / Median of all intra-cluster distances (calculated for the whole data)
(3) Median of all ratios in (1)

The second set calculates the ratio of the minimum inter-cluster distance to the maximum intra-cluster distance. This creates the most demanding test of cluster separation, for if a single aberrant token exists in a given class, giving rise to a large maximum intra-cluster distance, the ratio for that class or cluster becomes very small. It follows then that, while the median ratios above are more tolerant of data aberrations and can be said to be more robust, and therefore, suitable for evaluating the performance characteristics of pre-processing systems, the min/max ratio here may be regarded as better suited for trouble-shooting the systems (in the sense that it focuses on the areas where the pre-processed data do not cluster well) as well as identifying data characteristics.

(4) Minimum Inter-cluster distance / maximum intra-cluster distance maximum (calculated for each cluster)

(5) Minimum of all inter-cluster distances / maximum of all intra-cluster distances (calculated for the whole data)
(6) Median of all ratios in (4)

The function of the evaluation system as a trouble-shooter will be enhanced if the data-bound causes for poor clustering can be neutralised. The following two groups of measures are designed to achieve this separation by removing a portion of the data, i.e., above the ninety percentile and third quartile points, respectively, as outliers.

(7) 90% minimum inter-cluster distance / 90% maximum intra-cluster distance (calculated for each cluster)
(8) 90% minimum of all inter-cluster distances / 90% maximum of all intra-cluster distances (calculated for the whole data)
(9) Median of all the ratios in (7).

(10) 75% minimum inter-cluster distance / 75% maximum intra-cluster distance (calculated for each cluster)
(11) 75% minimum of all inter-cluster distances / 75% maximum of all intra-cluster distances (calculated for the whole data)
(12) Median of all the ratios in (10).

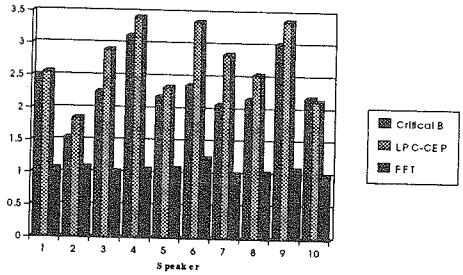

Figure 2. Median of 36 intra-/inter-cluster median ratios (AD data).

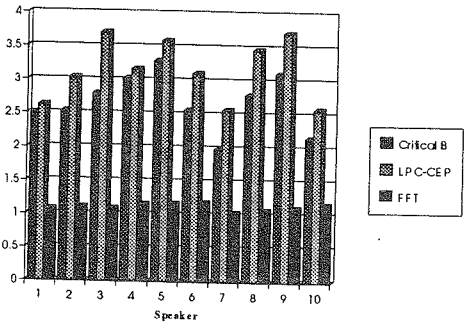

Figure 3. Median of 36 intra-/inter-cluster median ratios (PR data).

SELECTED RESULTS

The evaluation of the pre-processed data was performed with the full range of the statistical measures stipulated above. To show the general performance differences between the three pre-processors, however, we focus our attention to the median of all intra-/inter-cluster median ratios (Measure (3) above). The trends observed in these overall median ratio results were also confirmed in other measures. Recognition accuracy rates were also obtained from the distance matrices for the purpose of comparison.

The medians of inter-/intra-cluster median ratios for each vocabulary set are shown in Figures 2 and 3. It is apparent that the performance of the pre-processors as manifested in these ratios is highly speaker-dependent. Speakers 1 to 5 are males and 6 to 10 are females, and there is no clear correlation with the sex of the speaker. The FFT log power spectrum is seen as the poorest performer. Both the linear-prediction based cepstral processor (LPC-CEP) and the critical-band processor (Critical B) are seen to perform well, but the LPC-CEP seems consistently to perform marginally better than the Critical B (except for Speaker 10 in the AD data).

Generally higher ratios are obtained for the PR data than for the AD data. The same relative performances of the three pre-processors are seen in both vocabulary data.

Comparison with Recognition Rates

The recognition rates shown in Figures 4 and 5 follow the same trends as in the ratio data above for both data sets. The PR data show generally higher recognition rates than the AD data. As for the pre-processors, LPC-CEP edges Critical B out in all cases but for Speakers 2, 4, and 9 in the PR data. In general, the differences between the two closely performing pre-processors -- Critical B and LPC-CEP -- are brought out more clearly in the ratio data.

The markedly poor recognition rate for Speaker 7 in the PR data is an enigma, and an explanation is sought in a later section.

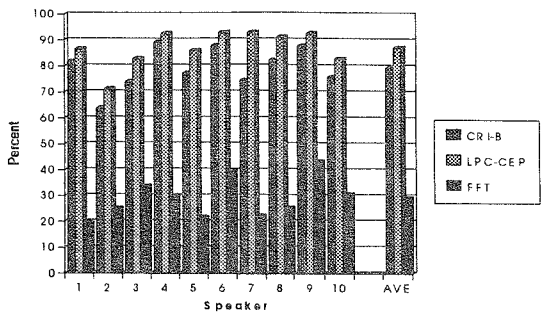

Figure 4. Recognition rates for AD data.

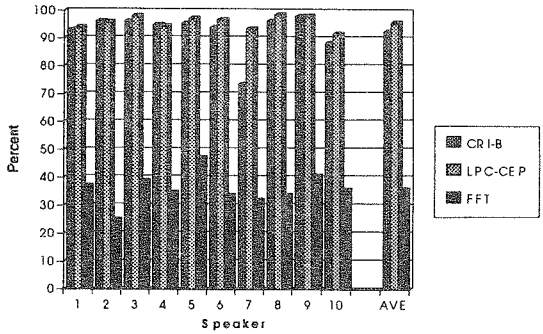

Figure 5. Recognition rates for PR data.

Word-by-word Analysis

To find out which words or groups of words are good or poor performers, and thereby seek any phonological, phonetic, or acoustic phonetic reasons for the observed trends, a word-by-word analysis is performed. Shown in Figures 6 and 7 are the inter-/intra-cluster median ratios for each word in the AD and PR critical-band pre-processed data. The ten speakers' ratios for each word category are represented by the three values indicating their minimum, maximum and average in ratio values, and the word categories have been sorted by the speaker average values.

Immediately obvious is the wide variation of the ratios for the words "R" and "Z" in the AD set. Some extremely high ratio values for "R" may be explained by the very distinctive acoustic properties of the retroflex /r/ which were produced consistently by some speakers in the citation style recording situation.

From the AD data, it can be observed that initial consonants, especially voiceless fricatives contribute to high cluster definition. Generally, highly consonantal words perform better than highly vocalic words. Words which are purely vocalic perform most poorly (e.g., "A" and "O").

The word "four" has an initial voiceless fricative, but its high ratio values may be also correlated with the fact that the vowel in this word is the only back monophthong in the entire 36-word alphadigit vocabulary. In fact, as the diphthong /oU/ is not quite so back, and /u/ is high-central and often diphthongised in the speakers' dialect, the vowel in "four" can be said solely to represent the back vowel region.

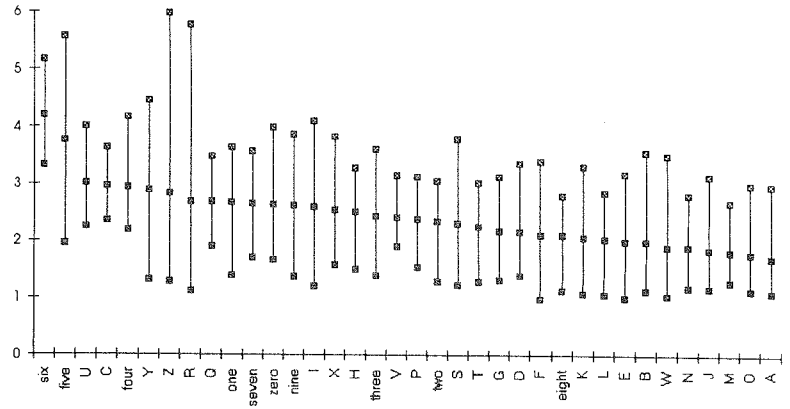

Figure 6. Sorted inter-/intra-cluster median ratios for AD data (Critical B).

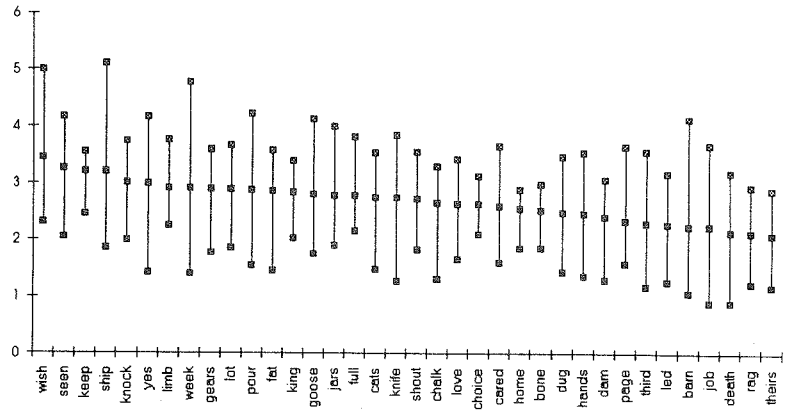

Figure 7. Sorted inter-/intra-cluster median ratios for PR data (Critical B).

The E-set is *not* seen as particularly confusable ("B" is the worst, others are not remarkably poor, but "C" is very good, which may be attributable to the effect of the initial voiceless fricative). The EH-set does show poorer cluster definition, especially when V / #_. The A-set words also seem to form poorly-defined clusters.

In the PR data, an additional observation is that voiced consonants (especially initially but also elsewhere) appear to perform poorly. The main difference between the two vocabularies lies in the difference between the best and the worst average ratio values. The range of average ratio values in the PR data is from 3.448 to 2.105, which is much smaller than that in the AD data: 4.202 to 1.697.

An attempt was made to find an underlying reason for the low recognition rate of Speaker 7 in the PR data. Figure 8 shows the speaker's ratio values plotted against the speaker means. Generally poor performance is seen throughout the phonetic categories and no plausible phonetic generalisation can

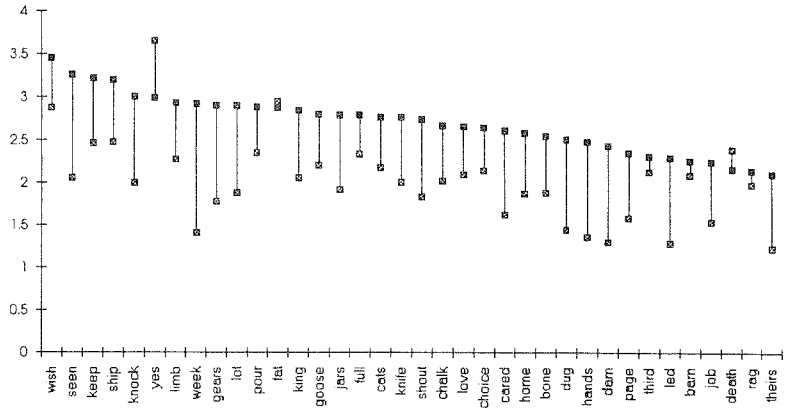

Figure 8. Sorted inter-/intra-cluster median ratios for PR data (Critical B): Speaker 7 (light box) vs ten speaker average (filled box).

be made. High vowels seem to correspond to somewhat low cluster definition but this entertains several counterexamples.

Amongst alternative explanations, the following seems the most plausible. With the PR data, the recognition rates appear to quantise in the region of 90% or above when the ratio value is above 2.0. When it is below 2.0 (which is the case with Speaker 7), the recognition rate falls into a more volatile region and can return a value substantially lower than what may be expected by linear extrapolation from the cases with the ratio value of above 2.0. This non-linear mapping between the ratio values and the recognition rates is applicable to other vocabulary data but the critical value will have to be estimated independently since factors such as phonetic balance, the number of words and the number of repetitions will affect the critical value.

Conclusions

The chief aim of this paper was to demonstrate the novel cluster analysis method designed to evaluate the performance differences of various pre-processors for speech recognition. The preceding sections have shown that the clustering information obtained by the method, while closely related to recognition rates, highlights the relative performance differences of different pre-processors more clearly.

The method also enables a closer scrutiny of the data by providing quantitative information about the degrees of cluster separation. The paper has shown that the clustering information is useful for analysing the performance characteristics of the vocabularies used for recognition.

REFERENCES

Clark, J.E. and Fraser, H. (eds.) (1982) *Australian Speech Archive*, Occasional Papers, Speech and Language Research Centre, Macquarie University, N.S.W.

Rabiner, L.. and Sambur, M. (1975) *An algorithm for determining the endpoints of isolated utterances*, Bell Sys. Tech. J., 54, 297-315.

O'Shaughnessy, D. (1987) *Speech Communication*, Addison-Wesley Publishing Company.

Zwicker, E. (1961) *Subdivision of the audible frequency range into critical bands (Frequenzgruppen)*, J. Acoust. Soc. Am., 33, 248-249.