

AN EFFICIENT TECHNIQUE FOR ISOLATED WORD RECOGNITION OF MONOSYLLABIC LANGUAGES

P.C.Ching^{*}, W.M.Lai^{*}, Y.T.Chan^o

ABSTRACT

An automatic speech recognition system is discussed in which the energy-time profiles at several frequency bands are used to represent an input utterance and then compared with a reference set obtained during training with many different speakers. To reduce considerably the number of misrecognitions as well as the overall matching time, a zero-crossing count front end is used for a voice/fricative initial classification. The recognition scheme is most suitable for monosyllabic languages and has the advantages of being very simple, avoiding time-warping and permitting low-cost implementation on a microcomputer. The system was evaluated for speaker-independent isolated word recognition of the ten Cantonese digits. A mean recognition accuracy of about 90-95% was obtained.

INTRODUCTION

Although the goal of machine recognition of continuous speech remains elusive, a greater degree of success has been achieved in recognition of isolated words from a fixed vocabulary. Indeed, in the past decade, a wide variety of talker-independent isolated word recognition systems have been built and used successfully in many applications (ref 1,2). However, the temporal aligning techniques engaged in most template-based recognizers require a large amount of computations (ref 3). Speech recognizers based on hidden Markov models, on the other hand, have less computation but more complicated parameter estimation procedures for model generation (ref 4). Large storage and computing requirements, together with complex system configurations, have limited the possibility of a high speed, low-cost implementation of these algorithms even for a small vocabulary.

This paper presents a new isolated-word recognition (IWR) scheme which is essentially template-based, but avoids the necessity of utilizing dynamic time warping to align a test and reference pattern. The method uses the energy-time profiles (ETP) of a word at different frequency bands as the parameter for recognition. Each of the bandpass filtered signals is divided into a fixed number of segments regardless of the duration of the utterance, and the energy of each of these segments is recorded and normalized to form the ETP matrix. Recognition is performed by comparing the ETP of a test word to each of the reference templates which were created in a training session using an iterative clustering technique. To reduce the number of word confusions and also the matching time, each utterance is classified into two groups depending on whether it is voiced initial or fricative initial. Now a word, after classification, will only be matched against one of the two groups of references. The algorithm is designed for the recognition of Cantonese, which is a monosyllabic tonal language commonly used in Southern China as well as in Hong Kong. A description of the recognition technique together with the experiments performed for its evaluation will be given in the following sections.

^{*}Dept. of Electronics, Chinese University of Hong Kong, Shatin, Hong Kong
^oDepartment of Electrical Engineering, Royal Military College of Canada, Kingston, Ontario, Canada K5K 7L0

OUTLINE OF THE RECOGNITION SYSTEM

A functional block diagram of our IWR system is shown in Figure 1. An input utterance is first sampled and its beginning and end is found by a detection algorithm similar to that described by Rabiner et al. (ref 5). The method primarily uses the energy and zero-crossing contour of the recorded signal and an appropriate set of thresholds to estimate the endpoint locations. It is noted that the phonetic structure of monosyllabic words is in the form of "initial + single vowel + final". Hence the mid-section of a word will always have high energy which provides a good starting point to search backward or forward for the beginning or end of the utterance. The result is an effective way of locating the endpoints of an isolated word with little uncertainties due to background noise or speaker inconsistencies in any reasonable acoustic environment.

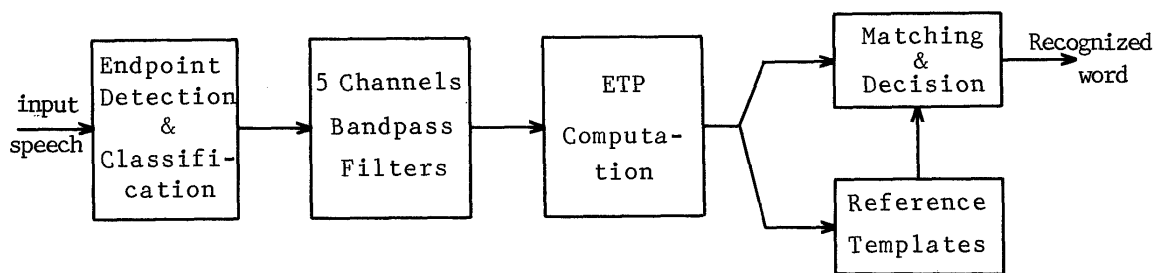


Figure 1

In order to reduce the number of comparisons during template matching, a preliminary classification is used in which the vocabulary is divided into two groups according to the phonetic labelling of their initial regions. Actually there are five different types of initials for Cantonese including nasals, glottals, plosives, fricatives and vowel-like initials and complicated analysis is needed to make an exact distinction between them. However, except for fricatives, they all have a common feature of possessing low frequency components. Thus a simple voice/fricative initial classification is employed which depends on the zero-crossing rate (ZCR) of the utterances (ref 6). Consequently, the members of the voiced-initial (VI) group possess low ZCR whilst those of the fricative-initial (FI) group have high initial ZCR. Additionally, since most of the members within the VI group are characterised by a sharp rise in energy, an utterance is also classified to the VI group if this condition is detected. An input utterance is said to be unclassified if neither of these conditions are satisfied and it will be compared with the whole vocabulary during matching. This classification scheme has been found to be very efficient and for more than two thousand test utterances, less than 10% were unclassified and no word was identified to the wrong group.

After endpoint detection and classification, the input speech samples are sent to a bank of five bandpass filters with passbands (i) 150-500Hz, (ii) 500-850Hz, (iii) 850-1.2kHz, (iv) 1.2k-1.8kHz and (v) 1.8k-3.2kHz. These filter bands are carefully chosen to provide uniform spacing and finer division in the most important 150-1.2kHz region. The outputs of these filters together with the wide-band signal form six different sequences of the utterance and their energy-time profiles are used as the parameter for recognition. Of course, more filters can be used to give a better representation of the spoken word, but this will obviously involve more computation and extra hardware. Each sequence is then evenly divided into a fixed number of segments, say N, with 50% overlapping. Hence, the segment length will vary according to the duration for different words.

Let $E(i)$ be the energy of each time segment $i=1,2,\dots,N$, of the sequence $q=1,2,\dots,6$, which is given by the sum of squares of every sample values within that frame. Assume E_{1M} is the maximum segmental energy of the wide-band sequence, then the self-normalized energy of each segment

$$\overline{E}_q(i) = \frac{E_q(i)}{E_{1M}} \quad (1)$$

is calculated which forms the energy-time profiles (ETP) of the word. Thus each utterance is now parameterized by 6 vectors of ETP, each with N elements and, therefore, dynamic time warping for pattern alignment is not necessary. We have chosen $N=16$ in our tests on an ad hoc basis.

The ETP distance D between a test utterance and a reference word is computed by

$$D = \sum_{q=1}^6 \sum_{i=1}^{16} \frac{[\overline{E}_q(i) - R_q(i)]^2}{\overline{E}_q(i) + R_q(i)} \quad (2)$$

where $R_q(i)$ is the reference ETP of block i and sequence q . The squared difference of each segment energy is divided by the sum so that the difference is normalized nonlinearly. Since the energy of the vowel in the middle of a monosyllabic word is much higher than that of the initial and final section, so even a small percentage of deviation of energy in the vowel region will dominate the overall distance measure. This undesirable masking effect is overcome by using the normalization as defined in (2). In our recognition system, we have used a multiplicity of ETP templates to characterize the variability of the features for a single word across different speakers. Therefore, a distance vector will be created for each reference word under comparison. A two-pass decision based on the K -nearest neighbour (KNN) rule is used in which the vocabulary item whose average distance of the K -nearest neighbours to the unknown utterance is minimum is chosen as the recognized word. The task of the first-pass is to determine whether there is more than one vocabulary word which is acoustically similar to the test token. In this case, we set $K=1$ and check if the differences between the minimum distance and the others exceed a pre-defined threshold. If not, we move on to the second-pass to resolve these confusions. In the second-pass decision, we perform the KNN rule with $K=2$ and find the recognized word. However, if the difference between the smallest and next smallest distances corresponding to two different words is less than a threshold, the test utterance will be rejected.

As mentioned previously, our speaker-independent IWR scheme is based on the use of multiple templates for each word in the vocabulary. The word templates are generated from an iterative clustering analysis of a large database containing many replications of each word. The technique employed in our system is very similar to the modified K -means (MKM) clustering algorithm proposed by Wilpon and Rabiner (ref 7). But, the condition under which clusters are split will depend on the largest intracluster distance instead of the average. This has the advantage of permitting the isolation of outliers while still maintaining the property that within each cluster the word patterns are highly similar. Three methods which rely on different sets of feature parameters derived from the ETP matrix have been applied to create the reference templates. The first trial, denoted by M1, is to generate references by clustering whole ETP matrices. The second and

third trial, denoted by M2 and M3, is actually based on the same principle except that the ETP vectors of individual bands or individual time frame are now used separately as the object for clustering respectively.

RECOGNITION RESULTS AND DISCUSSIONS

The recognition system was implemented on an IBM PC/XT and two experiments were conducted to evaluate the performance of the system for speaker-independent recognition with a vocabulary consisting of the ten Cantonese digits. The voice input came from a close-talking microphone to a cassette recorder and the environment was quiet with only little ambient noise. The recorded utterances were later band-pass filtered at 100-3.3kHz, digitized with a 12-bit A/D converter at 8kHz sampling rate and stored in memory for training and testing. In the training mode, fifteen speakers, eight females and seven males, were requested to utter each of the digits twice to provide data for clustering. We have allowed eight templates for each individual word in the vocabulary.

In the first experiment, we performed the recognition test on the same group of talkers who had taken part in the training process. Every one of them was asked to utter each of the ten digits ten times again. Therefore, there were altogether 1500 input tokens for testing. The mean accuracy obtained was 95.2%, 93.6% and 95.2% for M1, M2 and M3 with a rejection rate of 2.4%, 4.3% and 2.7% respectively. The results for M1 and M3 are comparable and are very satisfactory whilst for M2, the result is slightly worse. We have also examined the effects of reducing the number of templates for each reference word. In this case, the number of templates for each clustered parameter set can be varied between 2 to 8 by assigning a threshold in the splitting procedure within the clustering algorithm. Therefore, for M2, we might have eight templates for band 1 but only six for band 2, for example. We found that with 20% reduction in the total number of templates used in the reference set, there is only a marginal decrease in the overall recognition accuracy for M2 and M3. However, for M1, we have recorded a drop of roughly 1.5% in accuracy. In the second experiment, fifteen new talkers and ten trained talkers were invited to provide data for testing. Each of them gave five replications for each digit and thus a total of 1000 tokens were used. The recognition rate obtained was 91%, 89.7% and 90.7% for M1, M2 and M3 respectively. Again, M1 and M3 gave very comparable results.

A relatively simple speaker-independent isolated word recognition system has been presented in this paper. The recognition scheme is based on the ETP of a word that avoids the time-warping process, and it is particularly suitable for monosyllabic languages. Because of its simplicity, it is implementable on a microcomputer. In fact, a real-time system may be achieved by apportioning a large part of the pre-processing tasks to external hardware. Although it has been evaluated using a limited vocabulary and very few number of speakers and tokens, the preliminary results show that the methodology is sound.

REFERENCES

1. F.Itakura, IEEE Trans. ASSP-23, p67, Feb. 1975
2. J.Wilpon, L.Rabiner, A.Bergh, J. Acoust. Soc. Amer. 72, p390, Aug. 1982
3. H.Sakoe, S.Chiba, IEEE Trans. ASSP-26, p43, Feb. 1978
4. L.Rabiner, S.Levinson, M.Sondhi, BSTJ 62, NO. 4, p1075, 1983
5. L.Rabiner, M.Sambur, BSTJ 54, No.2, p297, 1975
6. W.M.Lai, P.C.Ching, Y.T.Chan, Int. J. Electronics (to be published)
7. J.Wilpon, L.Rabiner, IEEE Trans. ASSP-33, p587, June 1985