DIGIT-SPECIFIC FEATURE EXTRACTION FOR MULTI-SPEAKER ISOLATED DIGIT RECOGNITION USING NEURAL NETWORKS

Danqing Zhang* and J.Bruce Millar
Computer Sciences Laboratory
Research School of Information Sciences and Engineering
Australian National University

ABSTRACT

The digit-specific feature extraction approach extracts distinguishing features of spoken digits in order to use a smaller amount of data to represent the digits. This reduced representation of the distinctive acoustics of the digits was evaluated in an isolated digit recognition task using a multi-layer perceptron neural network architecture. The acoustic-phonetic design of features for English digits is described as is the means to extract them from spoken utterances. The results of a recognition system based on this feaure set are presented for the conditions of multi-speaker dependent and speaker independent testing. The data set for this study is the ten isolated digits 'zero' to 'nine' spoken by three male and five female Australian speakers.

INTRODUCTION

The selection of the best parametric representation of acoustic data is an important task in the design of any speech recognition system. The standard approach is to use all analysis frames that are available, but it is feasible to reduce the complexity of the system if only those frames of the reference utterance which are distinctive within the overall utterance set are compared.

In this paper, a digit-specific feature extraction (DSFE) approach is explored for the multi-speaker Isolated Digit Recognition task. The novel DSFE approach is based on the philosophy that comparison is required only at such distinctive points of the utterance. This approach avoids the computationally expensive dynamic time-warp procedure.

The phonetic structure of a spoken English digit consists of at most two vowels with maybe initial, middle, and final consonants. All these digits, with the exception of "eight", have an initial consonant. Although there are many different characteristics existing in the consonants, they all can be classified as either "voiced" or "unvoiced" (Ladefoged, 1975). Therefore, a major phonetic distinction between the initial consonants of English digits is between voiceless and voiced onset. Similarly, there are two broad classes of vowels: monophthongs and diphthongs. A major distinction between the vowels is between monophthongal and diphthongal vowel types.

A study on digits by Rudnicky et al (1982) showed different recognition results using the first halves and the second halves of each utterance. Their result indicated that the first halves give a better recognition score than the second halves. Our preliminary studies (Zhang et al, 1990) showed that cepstral coefficients which were selected at the peak energy can provide important information for digit discrimination. In our data this peak always existed in the first half of each digit. On the basis of these studies, it was decided to select features from just initial consonants and vowels so that the least amount of speech data could be used to represent the distinguishing features for digits.

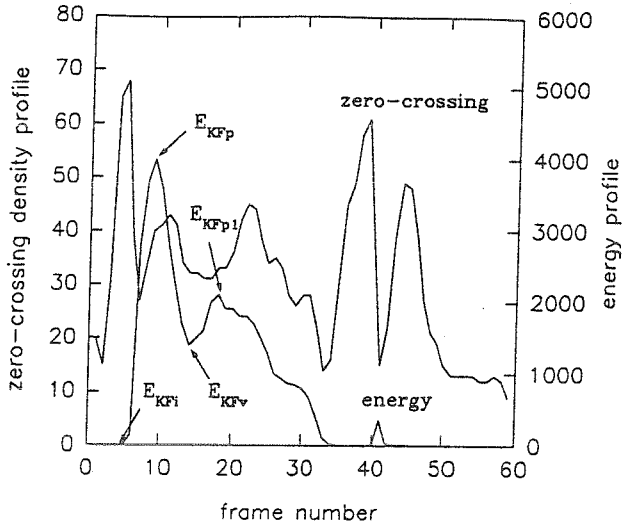

Figure 1: Energy and zero-crossing density profiles for digit "five"

SELECTION OF ACOUSTIC-PHONETIC FEATURES OF DIGITS

A temporal feature detector was used to detect several key-frames (KFs) from the speech signals of isolated digits using acoustic-phonetic rules. These rules involve the classification of spoken digits in terms of voiced or unvoiced onset and the diphthongisation of the initial vowel. The peaks and valleys of the energy-time profiles were detected in order to locate KFs which relate to the distinctive vocalic nature of each digit, and zero-crossing rate was explored to locate a KF to detect the voicing status of the onset. A small number of frames around these detected KFs were chosen in order to include enough information to fully encompass the selected features.

Selection of Key Frames to represent vowels

It is well known that the energy time-contour of a speech signal is a valuable parameter to indicate the temporal location for the extraction of input features in speech recognition (Denes, 1974; Zue and Schwartz, 1980; Rabiner et al, 1984; Lai et al, 1987; Burr, 1988). In the context of isolated digit recognition, Burr (1988) showed a good result in his system using a simple feature extraction approach in which only energy was used in the selection of input features for a neural network. The strategy adopted in his approach was to select a frame is located at the energy maximum of the spoken utterance and two additional frames located before and after at a fixed fraction of the maximum energy.

In preliminary experiments (Zhang et al, 1990), an approach similar to that of Burr (1988) was used. A group of three frames were chosen with one frame at the energy maximum and two frames before and after which were closest to half this energy. These experiments showed that the maximum energy from each digit was located centrally over the vowel for a monophthongal digit, or the first vowel quality of a diphthongal digit. This measure was therefore used in the DSFE approach to locate the first key-frame (KF*p*) (Figure 1 and Figure 2) within each digit.

In addition, it has been found that for the eight speakers examined, two peaks often occur in the energy time-contour parameter where the spoken digit contains a diphthong. This feature was represented in

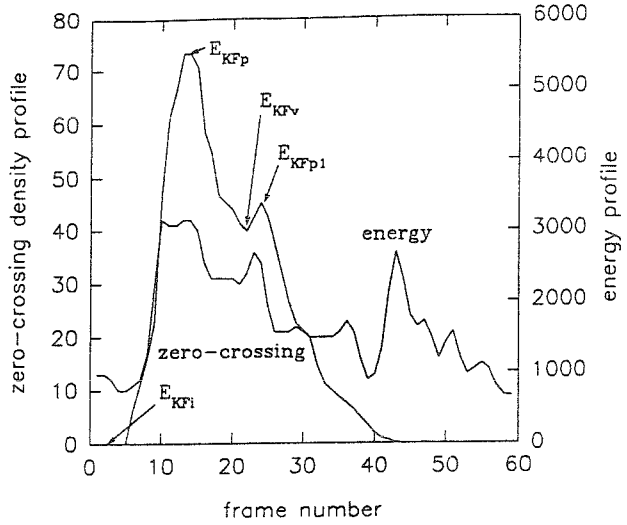

Figure 2: Energy and zero-crossing density profiles for digit "nine"

the DSFE by two key-frames, KFv and KF$p1$, which were selected respectively at the valley between the two peaks and at the later peak of the time-contour on right-hand side of KFv. The peaks and valleys were detected as simple maxima and minima of a smoothed version of the contour. These two frames, together with the key frame at the overall maximum energy, were used to locate the distinctive vocalic nature of each digit. Note that when only one energy peak exists the KFv and KF$p1$ are equated to the end of the analysis record which represents background acoustic ambience.

Selection of Key Frames to represent initial consonants

Zero-crossing density has been widely used to provide important information for distinguishing between voicelessness and voicing. A voiced consonant has a low zero-crossing density reflecting the presence of energy at a low frequency typical of voicing, while an unvoiced consonant has a much higher one as no such low frequency energy is normally present. Generally, both unvoiced and voiced consonants have a lower energy than a stressed vowel. Sometimes the energy of a weak unvoiced consonant, such as the initial consonant of "five", can be very similar to the energy characteristics of the background noise. However, as the weak fricative typically has higher frequency energy than background noise, a high zero-crossing density can often distinguish such a weak fricative from a background noise even though they cannot be resolved using energy alone. These characteristics were explored in order to derive rules to locate and broadly categorise the initial consonant for each spoken digit. Figure 1 and Figure 2 show the energy and zero-crossing density time-contours across all frames for utterances of the digits "five" and "nine", which contain the same diphthongised vowel. From these two figures, it can be seen that features appear in the initial regions of the zero-crossing profiles which distinguish between the two digits. The zero-crossing peak which is present for digit "five" is totally missing for digit "nine".

The algorithm used in the DSFE approach was to search forward towards the beginning of a utterance until both the amplitude of energy and zero-crossing density were greater than the boundary values of the energy and the zero-crossing density between background noise and the onset of speech. These boundary values were pre-determined from statistics of the initial frames of the energy and zero-crossing

density time-profiles for all utterances of the digit data which were used in our study and the specific background noise of their recording. A key-frame for initial consonants (KFi) was selected when energy and zero-crossing density exceeded these values.

Selection of Regions for Each of the Selected Key Frames

Once four KFs (KFi, KFp, KFv and KF$p1$) were initially defined, it was necessary to determine how precisely the key frames located distinctive information about the digits. This was done by selecting a region about each key frame and searching for the minimal size of that region that produced good digit discrimination. It was found that the system gave the optimal scores for both multiple speaker dependent (MSD) and multiple speaker independent (MSI) tests when a group of 12 frames were selected as inputs to a neural network which was designed to recognise all the digits. The group of 12 frames comprised a region of 3 frames which started one frame after KFi, a region of seven frames which extended three frames before and three frames after KFp, and single frames KFv and KF$p1$. These frames would appear to include spectral information about consonantal onset, extensive spectral information about the initial vowel, and brief spectral information about the trailing part of the utterance.

Data processing and the recognition engine

The digits were originally recorded at a rate of 16,000 samples per second following processing by a low-pass anti-aliasing filter with a -3dB point at 7.2 kHz. The data used were resampled at 8,000 samples per second following digital low-pass filtering to 3.6kHz.

After each utterance was segmented and stored into one file, it was analysed using a frame of 256 sampled-data points (32 ms) and a 50% (16 ms) overlap between adjacent frames. Each frame was subjected to a 10th order LPC autocorrelation analysis using a pre-emphasis factor of 0.98. The 10 low-order cepstral coefficients were then derived from the reflection coefficients.

Once all frames were processed, a total number of 120 LPC cepstral coefficients for the 12 representative frames described above were selected for subsequent use as input data for the recognition engine.

The architecture of the multi-layer perceptron used as the recognition engine for the experiments to demonstrate the DSFE process in action has three-layers with two hidden layers and one output layer as described in Zhang and Millar (1992). It was designed and trained using the fast training algorithm of Brent (1991).

EXPERIMENTAL STUDIES

Ten classes of isolated digits from "zero" to "nine" spoken by three male and five female Australian speakers were selected for this study. Every class of digits was repeated ten times by each speaker yielding a total of 800 utterances.

Two methods were adopted to choose training and testing data: (1) 50% of data (400 utterances) comprising five utterances for each class of digit from all the eight speakers were selected for training and the other 50% were used for testing: this was referred to as the multi-speaker-dependent (MSD) test; (2) 75% of the speakers (600 utterances) whose data contain ten utterances from each of six different speakers (four females and two males) for each class of digit were selected for training and other 25% of speakers whose data contain ten utterances from each of the other two different speakers (one female and one male) were used for testing: this was referred to as multi-speaker-independent (MSI) test. No rotation of the utterances was performed in MSD testing but all the possible 15 rotations of the speakers in MSI testing were performed.

The experimental results for both MSD and MSI tests are given in Table 1. Each experiment was run

Experiments	*No.Hidden Units in 1st Layer*	*No.Hidden Units in 2nd Layer*	*Training Time (sec)*	*Recognition Rate*	*Variation of Results*
MSD	11	12	1095	93.0%	1.0%
MSI	16	17	1772	92.5%	1.3%

Table 1: Results for multi-speaker dependent and independent tests

several times using different random starting points in the search space of Fast Training Algorithm (Brent, 1991) based MLP training. The results shown in the tables give the mean recognition rate and its range of variation about the mean over five such runs.

No empirical results were obtained with full-time sampling of the analysis frames. As the average number of frames was 60, a factor of 5 in the size of the input layer of the MLP would have caused a major increase in the complexity of the network.

CONCLUSION

This paper has outlined a novel specification of the input to an isolated digit recognition engine implemented as a neural network where the adoption of a digit-specific feature extraction approach has significantly reduced the complexity of the neural network architecture required.

The benefit of the DSFE approach to input signal pre-processing can be evaluated by the theory of neural networks. It has been shown (Brent, 1991) that the number of inputs is the major determinant of the overall number of connections in a network which can successfully classify its input patterns. The overall number of connections determine the total number of weights, which in turn determine the computational complexity of the network. The DSFE approach extracts the distinguishing features of spoken digits in order to use a smaller amount of data to represent the digits. The value of DSFE in minimising the number of inputs to the NNs and hence reducing their internal complexity is therefore self-evident.

ACKNOWLEDGEMENT

The authors would like to thank P. Duke and Telecom Australia for providing the spoken digit data for this research.

REFERENCES

R.P. Brent (1991), "Fast training algorithm for multi-layer neural nets", *IEEE Trans. Neural Networks*, Vol. 2, No. 3, pp. 346-354.

D.J. Burr (1988), "Experiments on neural net recognition of spoken and written text", *IEEE Trans. ASSP*, Vol. 36, No. 7, pp. 1162-1168.

P.Denes (1974), "Speech recognition: old and new ideas", in *Speech Recognition*, ed. by D.R. Reddy (Academic Press), pp. 73-95.

P. Ladefoged (1975), *A Course in Phonetics* (New York, Harcourt Brace Jovanovich).

W.M. Lai, P.C. Ching and Y.T. Chan (1987), "Discrete word recognition using energy-time profile", *International Journal of Electronics*, Vol. 63, No. 6, pp. 857-865.

L.R. Rabiner, K.C. Pan and E.K. Soong (1984), "On the performance of isolated word speech recognizers using vector quantization and temporal energy contours", *AT&T Bell Lab. Tech. J.*, Vol. 63, pp. 1245-1260.

A.I. Rudnicky, A.H. Waibel and N. Krishnan (1982), "Adding a zero-crossing count to speech information in template-based speech recognition", *Technical Report*, CMU-CS-82-140, Carnegie-Mellon University.

D. Zhang, J.B. Millar and I. Macleod (1990), "Multi-speaker digit recognition using neural networks" *Proc. 3rd Australian International Conference on Speech Science and Technology*, Melbourne, pp. 28-33.

D. Zhang and J.B. Millar (1992), "A hybrid approach to isolated-digit recognition", *Proc. 3rd Australian Conference on Neural Networks*, Canberra, pp. 278-281.

V.W. Zue R.M. Schwartz (1980), "Acoustic processing and phonetic analysis", in *Trends in Speech Recognition*, ed. by W.A. Lea (Prentice-hall, Inc., Englewood Cliffs, New Jersey), pp. 101-124.

END NOTE

The first author is now with the Computer Centre, University College, University of New South Wales, Australian Defence Force Academy, Canberra, ACT 2600, Australia.