SPOKEN JAPANESE DIGITS RECOGNITION SYSTEM USING LVQ

Kiyoshi Kondou, Hiroyuki Kamata and Yoshihisa Ishida

Department of Electronics and Communication, Meiji University
1-1-1, Higashi-mita, Tama-ku, Kawasaki, 214 Japan

ABSTRACT - In this paper we present a new spoken Japanese digits recognition system using LVQ (Learning Vector Quantization). LVQ has very simple algorithm and generates some good reference vectors for DP (Dynamic Programming) matching. Our overall system has two characteristics. One is that the learning and recognition systems have two modes which are for vowels and consonants. The other is that the vowel recognition system predicts consonant label. In our two recognition systems the vowel recognition system segments vowel part and recognizes vowel label. In the consonant recognition system, actually recognized label is compared with predictive label. Each system (vowel and consonant recognition system) generates series of label Japanese digit. According to the experiment, our recogniyion system has good performance.

INTRODUCTION

Conventional recognition system with DP matching has to have many reference vectors. We attempte to generate a smaller number of reference vectors with LVQ. LVQ is often used for phenomena recognition and it has high performance. But it often recognizes wrong phenomenon near correct phenomenon point when it is used continuously. So it is necessary to ascertain whether it is wrong phenomenon. Our system recognizes vowel and consonant separately. The vowel with a few category is first recognized. To improve the recognition rate for consonants, our recognition system next segments vowel part from other parts. In the consonant recognition system, the consonant to be recognized is predicted by the phoneme sequence of each digit and vowel which has already been recognized. The consonant phenomenon is recognized based on the predicted and actually recognized phenomena.

LVQ (Learning Vector Quantization)

LVQ can generate some good reference vectors from many vectors and its algorithm is very simple. Then we describe the learning rule of LVQ as following:

First we calculate Euclidean distance for all reference vectors $m_i(t)$ and then the closest reference vector $m_c(t)$ to the input vector $x(t)$ by Eq.(1).

$$\|x(t)-m_c(t)\|=\min\{\|x(t)-m_i(t)\|\} \tag{1}$$

Let the input vector x(t) belong to category S_r and the closest reference vector $m_c(t)$ to x(t) belong to S_s. If S_r is the same category as S_s , $m_c(t)$ is moved closer to x(t) by Eq.(2)

$$m_c(t+1)=m_c(t)+a(t)\{x(t)-m_c(t)\} \tag{2}$$

where the function a(t) ought to be a monotonically decreasing function of time $(0<\alpha(t)<1)$. If S_r is a category other than S_s, $m_i(t)$ is moved away by Eq.(3).

$$m_c(t+1)=m_c(t)-a(t)\{x(t)-m_c(t)\} \tag{3}$$

All reference vectors excepted $m_i(t)$ stay there.

$$m_i(t+1)=m_i(t) \tag{4}$$

Thus LVQ moves the vectors that belong to same category closer and the vectors that belong to other category away. Then it makes clearly bound surface among classes.

SPEECH ANALYSIS

Speech signals are sampled at 10 kHz and pre-emphasized by $(1-Z^{-1})$. Each spoken digit is hamming-windowed and 256-points FFT is computed every 8msec. Mel-cepstrum coefficients (16 sections) are normalized by each frame.

RECOGNITION SYSTEM

Segmentation(Vowel Recognition)

We search voiced part using speech energy E(I) as shown in Eq.(5) which is greater than N (Fig.1).

$$E(I)=\frac{\sqrt{\sum_{i=0}^{N} x(i)^2}}{N} \tag{5}$$

I:Frame,x(t):speech

In the voiced part, the input vector, 3×16-dementional vector, for LVQ is generated in each vowel part (stable part). Then it is learned by LVQ and generates some reference vectors. Their reference vectors are applied to vowel label, /a/, /i/, /u/, /e/ and /o/, with DTW (Dynamic Time Warping) in whole of vowel parts. Each label d_i is compared with d_{i+j} $(j=1,2,\cdots)$ separately. If the total number of different labels

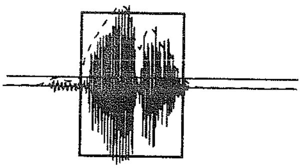
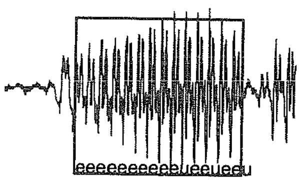


Figure 1. Voiced part (doted line: energy)



Label /e/

Figure 2. Segmentation (vowel recognition)

is greater than N and (j-i) is greater than M, d_i is recognized (Fig.2).

Consonant recognition

Our system can recognize the consonants used in Japanese spoken digits (Table.1).,/z/,/r/,/chi/,/n/,/s/,/y/,/g/,/k/,/h/ and /kyu/. The input vectors are generated by previous 7 frames to vowel part. When the

same label d_i is generated N times in M frames, the label D_i is nominated as consonant label. In many cases, the generated consonant labels are plural and different. Their nominated consonants are compared with predicted consonant as following: Our system predicts a consonant label previous to the vowel one, when our system segments a speech signal. For example, when /a/ is recognized, predictive consonant labels are /s/ and /n/ (see Table.1). If the label D_i generated by DTW belongs to the predictive consonant labels, the label D_i is recognized as correct label. If it is not so, the vowel label (segmentation) has been recognized uncorrectly.

0	ZERO	1	ICHI	2	NI	3	SAN	4	YON
5	GO	6	ROKU	7	NANA	8	HACHI	9	KYUU

Table 1. Japanese digits

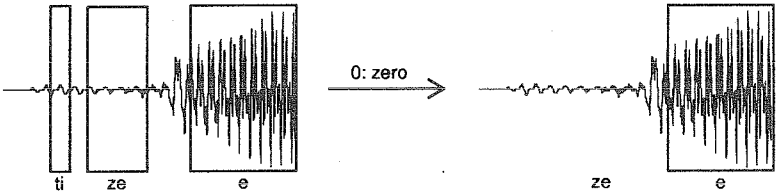


Figure 3. Consanant recognition example (zero)

RESULT

We experimented for the speech except the unvoiced part and the vowel part to be too short for our system.

Vowel	a	i	u	e	o
Rate [%]	97.1	100	95.0	100	100

Table 2. Vowel recognition result

The result of vowel recognition (Table 2) shows that our system segments correctly and assigns correct labels. The label /u/ is often inserted on the semivowel part. In consonant recognition (Table 3) label /y/ and /r/ reduce recognition rate because of /u/ inserted. The second /n/ of `NANA` is stable phenomenon and /n/ used to learning system is a transition part from the consonant to the vowel. For this reason /n/ reduce recognition rate, too. The digit recognition result (Table 4) has high performance because our system estimates a correct digit from the sequence of labels generated in which phoneme change is considered.

	z	r	chi	n	s	y	g	k	h	kyu
Rate [%]	100	83.3	92.9	85.7	90.0	90.9	100	90.0	100	81.8

Table 3. Consonant recognition result

	0	1	2	3	4	5	6	7	8	9
Rate [%]	83.3	100	100	90.0	83.3	100	83.3	90.0	100	81.8

Table 4. Japanese digit recognition result

CONCLUSIONS

In the paper we have presented a new method for the spoken Japanese digits recognition using LVQ. The proposed method reduce the number of reference vectors for DTW, and sequence label (like digit letter) and vowel recognition predict the consonant. The experimental results have shown that this method is effective for word recognition.

REFERENCES

E.McDermott, S.Katagiri, *Shift-Invariant, Multi-Category Phoneme Recognition Using Kohonen's LVQ2*, IEEE ICASSP88,9.S3, pp.81-84, (1986.5)

Sakoe.H, Chiba.S, *Dynamic programming algorithm optimization for spoken word recognition*, IEEE Trans.Acoust.,Speech,Signal Processing. ASSP-23, 1, pp.67-72, (1978)

T.Kohonen, *The neural phonetic typewriter*, IEEE Computer, 21, 3, pp.11-22, (1988)

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.

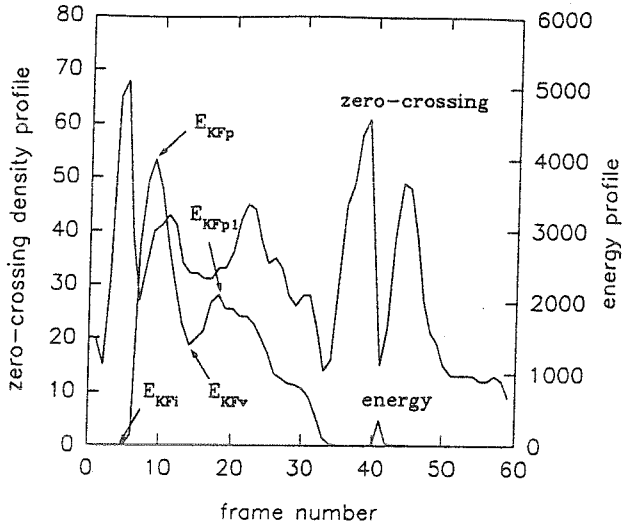


Figure 1: Energy and zero-crossing density profiles for digit "five"

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