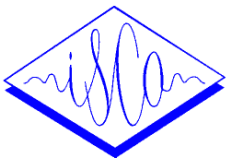

2nd International Conference on
Spoken Language Processing (ICSLP 92)
Banff, Alberta, Canada
October 12-16, 1992

SPEAKER-INDEPENDENT SPOKEN DIGIT RECOGNITION IN NOISY ENVIRONMENTS USING DYNAMIC SPECTRAL FEATURES AND NEURAL NETWORKS

Tadashi KITAMURA, Satoshi ANDO and Etsuro HAYAHARA

Electrical and Computer Engineering, Nagoya Institute of Technology,
Gokiso-cho, Showa-ku, Nagoya-shi 466 Japan

10.21437/ICSLP.1992-237

ABSTRACT

This paper describes a speaker-independent word recognition method in noisy environments using dynamic and averaged spectral features of speech and neural networks. Spectral features of speech are obtained from a two-dimensional mel-cepstrum (TDMC). TDMC is defined as the two-dimensional Fourier transform of mel-frequency scaled log spectra in the frequency and time domains. In this paper, several regions of dynamic and averaged spectral features of TDMC word are used as training data of neural networks. Neural networks are feed-forward networks with three layers and learn automatically by a back propagation training algorithm.

In order to improve the recognition performance in noisy environments, the learning order and SNR of the training data are considered in this study. Experimental results of speaker-independent word recognition for Japanese ten digits show that the proposed method gives better results especially in low SNR environments than a usual method.

1. INTRODUCTION

Instantaneous spectral features represented by LPC, LSP, LPC-cepstral coefficients etc. have been used in speech recognition. However, in noisy environments, as noise increases, spectral peaks and valleys become unclear, and they are sometimes destroyed by noise spectrum. Therefore, it becomes very difficult to recognize speech in noisy environments using only instantaneous spectral features. Some word recognition methods using instantaneous and dynamic spectral features were demonstrated to be effective for speaker-independent word recognition and speaker recognition etc. [1]-[6]. Furthermore, dynamic spectral features also were demonstrated to be less affected than instantaneous spectral features in noisy environments[4].

This paper describes a speaker-independent word recognition method in noisy environments using dynamic and averaged spectral features based on a two-dimensional mel-cepstrum(TDMC) and neural networks. TDMC is defined as the two-dimensional Fourier transform of mel-frequency scaled logarithm spectra in the frequency and time domains, and it consists of averaged spectral features and dynamic spectral features of the two-dimensional mel-log spectra. In this paper, a speaker-independent word recognition method for Japanese ten digits is discussed. The proposed method uses neural networks and averaged and dynamic spectral features of TDMC as training data for neural networks. Furthermore, noise-added reference template sets are used to improve recognition performance in high noise environments.

2. TWO-DIMENSIONAL MEL-CEPSTRUM

Let the time varying short-time spectrum in the discrete time be X(k,m) (k=0,1,..., N), where m is the frame number, k is the frequency number and N is the frame length of analysis frame. Therefore, mel-frequency scaled logarithm spectrum S(k,m) is given by

$$S(k,m) = Tw[\ \ln |X(k,m)|] \qquad (1)$$

where Tw[.] is a warping function that transforms linear frequency to mel-frequency. Two-dimensional mel-cepstrum(TDMC) C(q,p) is defined as the two-dimensional Fourier transform of the mel-log spectra S(k,m) in the frequency and time domains as follows.

$$C(q,p) = \sum_{m=0}^{M-1} \sum_{k=0}^{N-1} S(k,m)\ W_1^{kq}\ W_2^{mp}/MN$$

$$(W_1=\exp(-j2\pi/N),\ W_2=\exp(-j2\pi/M)) \qquad (2)$$

Because C(q,p) has the symmetric cahracteristics, it is enough to consider only a quarter of TDMC. It consists of some regions representing averaged and dynamic spectral features[4].

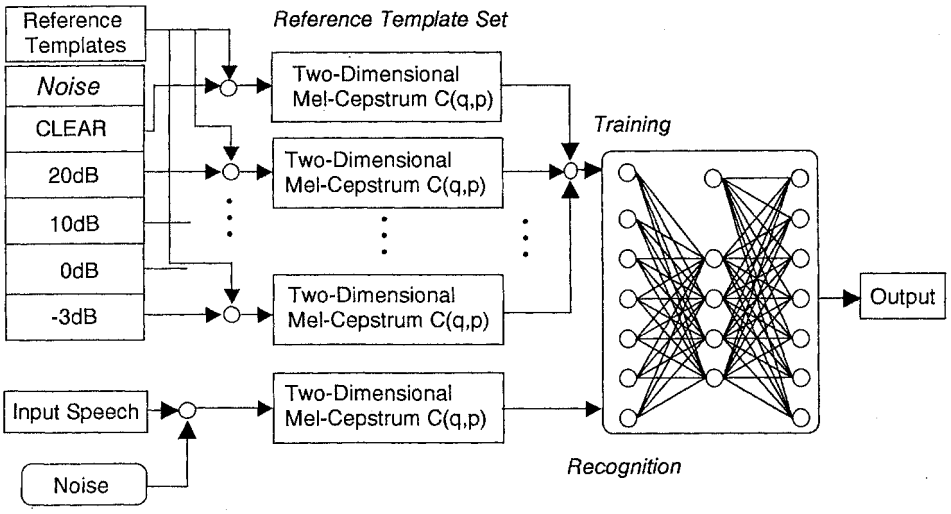

Fig.1 Basic structure of recognition System

In this study, the speech utterance is sampled at 10 kHz with 12 bits. For the digitized speech the beginning and ending points of the each utterance are semiautomatically determined by manual inspection using power of speech. In order to use a linear time scaling method for time-normalization, the reference speech and the incoming speech are divided into the same M frames. When the number of speech sample is LT, a frame period L is LT/M. A 25.6 ms Blackman window shifted every L samples is used to compute TDMC coefficients as in reference[4].

In the previous study, it has been found that smoothed dynamic spectral features of TDMC are less affected by the additive noise than the instantaneous spectral features[9]. Therefore, we utilize some regions of TDMC that represent smoothed dynamic spectral features and averaged spectral features.

3. RECOGNITION SYSTEM

Fig.1 shows a basic structure of the recognition system. In the learning step, this system analyzes TDMC coefficients of each noise-added reference template using noise-free speech and noise. Five sets of noise-added reference templates with noise-free, 20, 10, 0 or -3 dB SNR are used in this study. Then obtained dynamic and averaged spectral features of TDMC of each reference template are used for training neural networks. In the recognition step, TDMC coefficient of input speech with background noise is analyzed and neural networks classify input speech into a specified category among ten digits.

Neural networks used in this study have a three layered feedforward neural network and learn using a back-propagation training algorithm. Each layer is fully interconnected with the higher layer. According to preliminary experiments, a hundred and five TDMC coefficients of Japanese digit were used as the input elements of the network. The numbers of hidden elements and output elements were set to twenty and ten, respectively.

The momentum and the learning rate of neural networks are set to 0.9 and 0.2, respectively. A training procedure of the neural network is finished when the error decreases less than 0.01 or iteration times increases greater than 5000. As neural networks, an artificial neural network accelerator named "Neuro Turbo" is used for training and recognition. "Neuro Turbo" is implemented using four general purpose 24bits floating point Digital Signal Processors (MB86220)[9].

4. EXPERIMENTS

4.1 Data base

We used an isolated Japanese digit data base of ten male speakers. Each speaker recorded 10 Japanese digits (1:/ichi/, 2:/ni/, 3:/saN/, 4:/yoN/, 5:/go/, 6:/roku/, 7:/nana/, 8:/hachi/, 9:/kyu/, 0:/rei/), five utterances for each digit. A noise database was obtained from gaussian

white noise generated by computer and colored noise recorded in Nagoya station. Signal to noise ratio (SNR) is defined as the ratio between the powers of speech and noise in the analysis interval.

Reference templates are five utterances spoken by five speakers among ten speakers. One reference template set consists of all reference templates of ten digits with specified SNR, such as -3, 0, 10, 20 dB or noise-free. Therefore, we can get five reference template sets, four of which are noise-added reference sets and one is a noise-free reference set. The remaining data of other five speakers are used as input data for speaker-independent word recognition experiments.

4.2 Experiments

Speaker-independent word recognition experiments for Japanese ten digits in noisy environments were carried out to evaluate the effectiveness of this method.

Experiment 1

Fig.2 shows the recognition results when neural networks learn only one of the reference template sets, which is obtained from one specified SNR data, i.e., noise-free or 20, 10, 0 or -3 dB. As seen in Fig.2, this method gives good results (greater than 95%) in noise-free, 20 and 10 dB SNR environments when a reference set of 20 dB SNR is used for training neural networks. Therefore, it is found that neural networks tuned by 20 dB SNR data gives the best recognition results in white noise environments. In the colored-noise environments, almost same results are obtained. However, as noise increases, recognition rate becomes worse. Therefore, it is necessary to improve recognition performance especially in low SNR environments.

Experiments 2

In order to improve the recognition performance in low SNR environments, five reference template sets were used for training neural networks. Reference template sets are obtained from noise-free or 20, 10, 0 and -3 dB training data, respectively. Furthermore, the learning order and the learning rate of the training data were considered. In this study, the data rate of training data is 5:3:1:1:1 for clear(noise-free), 20dB, 10dB, 0dB and -3dB. The learning order is clear, 20dB, clear, 10dB, clear, 20dB, clear, 20dB, clear and -3dB. Fig.2 shows the recognition results.

As seen in Fig.2, the recognition performance is much improved especially in low SNR environments. For comparison, some experiments by a previous method also were carried out. The method uses a combination distance of two distance measures defined in the regions of TDMC representing dynamic and averaged spectral features[8]. The experimental results of the distance method using all reference sets are shown in Fig.2 too. This method gives better results than a distance method in low SNR environments(0 and -3dB). The distance method gives better results than this method in high SNR environments(clear and 20dB). This method gives almost constant recognition rates(98-95%) than a distance method(99-90%).

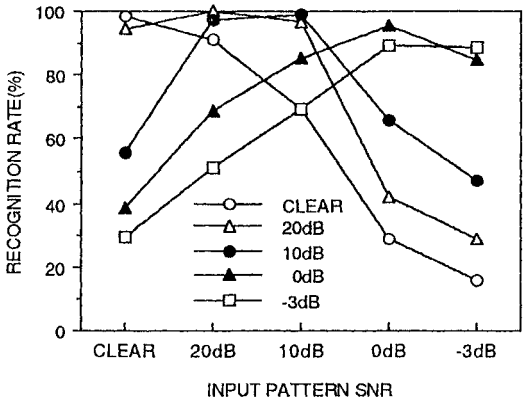

Fig.2 Recogntion rates vs. input data SNR when only one of the reference template sets is used for training neural networks.

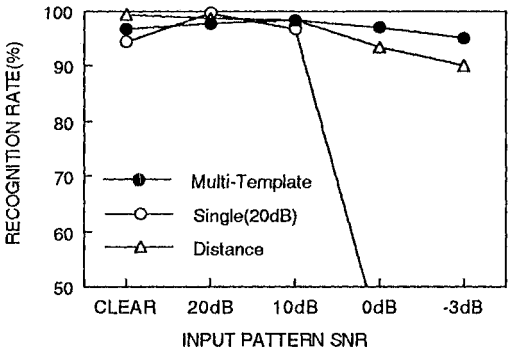

Fig.3 Recognition rates when neural networks learn a multi-template set, a single template set and use distance measure.

Experiments 3

From the experiments 1 and 2, this method is found to be robust in the noisy environments. However, noise sometimes changes in characteristics as well as in noise level. Therefore, we have to consider not only the

change of noise level but also noise characteristics. In order to study the effect of the change of noise, some recognition experiments were carried out. Fig.4 shows the recognition results when neural networks learn white-noise-added reference template sets and classify color-noise-added input data. In these experiments, data rate of training data is 3:2:1:1:1 for clear, 20dB, 10dB, 0dB and -3dB, respectively. The learning order is clear, 20dB, 10dB, clear, 0dB, 20dB, clear and -3dB, respectively. Experimental results are much affected by the change of noise characteristics. Especially, the recognition rates in low SNR are much affected. Therefore, both of color-noise-added reference template sets and white-noise-added reference template sets are used for training neural networks. The experimental results are shown in Fig.4.

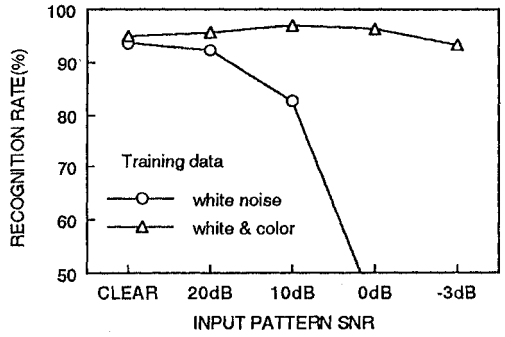

Fig.4 Recognition rates in color noise environments when white and colored noise-added reference sets or white noise-added reference sets are used for training neural networks.

As seen in Fig.4, neural networks that learn both reference template sets are less affected by the change of noise level and characteristics. The recognition rates of 97-93% are obtained in all SNR environments. Therefore, it is found that this method is effective for word recognition in noisy environments where noise level and characteristics sometimes change.

In this study, we discussed training data of neural networks and recognition rates and good results were obtained. However, there are many kinds of noise, so that it is necessary to consider noise subtraction, Lombard effect and other modification to this method.

5. CONCLUSIONS

In this paper, a speaker-independent word recognition method in noisy environments is proposed. This method uses dynamic and averaged spectral features based on TDMC and neural networks. TDMC is defined as the two-dimensional Fourier transform of mel-frequency scaled log spectra and it consists of averaged and dynamic spectral features. Some regions of TDMC representing dynamic and averaged spectral features are used as training data for neural networks. Neural networks are feed-forward networks with three layers and learn automatically by a back propagation training algorithm. In order to improve recognition performance in low SNR environments, neural networks learn noise-added reference template sets.

In order to evaluate the effectiveness of this method, speaker-independent word recognition experiments for 10 Japanese digits uttered by ten male speakers were carried out. For comparison, this method was compared with a distance method using distance measure. Experimental results have shown that this method gives recognition rates of 98-95% in noise-free to -3dB SNR white noise environments. Therefore, this method is effective for word recognition in noisy environments and robust for the change of level and characteristics of noise.

REFERENCES

[1] F.K.Soong and A.E.Rosenberg,"On the Use of Instantaneous and Transitional Spectral Information in Speaker Recognition", Proc. ICASSP, pp.877-880, 1986.

[2] S.Furui,"Speaker-independent isolated word recognition using dynamic features of speech spectrum", IEEE Trans. Acoust., Speech, Signal Processing, vol.ASSP-34, 1, pp.52-59, 1986.

[3] S.Furui,"A VQ-Based Preprocessor Using Cepstral Dynamic Features for Speaker-Independent Large Vocabulary Word Recognition", IEEE Trans. Acoust., Speech, Signal Processing, vol.ASSP-36, 7, pp.980-987, 1988.

[4] T.Kitamura and E.Hayahara,"Digit Recognition using Static and Dynamic Features of Two-Dimensional Mel-Cepstrum ", Trans. IECE, vol.J-72-A, 4, pp.640-647, 1989(in Japanese).

[5] R.Oka,"Phonetic Recognition of Each Frame with Vector Field Feature Using Continuous Dynamic Programming", Proc. ICASSP, pp.2291-2294, 1986.

[6] J.M.Baker and D.F.Pinto,"Optimal and Suboptimal Training Strategies for Automatic Speech Recognition in Noise, and the Effects of Adaptation on Performance", Proc. of ICASSP'86, pp.745-748, 1986.

[7] T.Kitamura and E.Hayahara,"Word Recognition using a Two-Dimensional Mel-Cepstrum in Noisy Environments", J. Acoust. Soc. Amer., vol.84, suppl.1, paper PPP.6, 1988.

[8] T.Kitamura, E.Hayahara and Y. Shimazaki, "Speaker-Independent Word Recognition in Noisy Environments using Dynamic and Averaged Spectral Features Based on a Two-Dimensional Mel-Cepstrum", Proc. of ICSLP90, pp.1129-1132, 1990.

[9]Y.Sato, A.Iwata, N.Suzumura, S.Matsuda and Y.Yoshida,"A Neural Network Accelerator Using General Purpose Floating Point Digital Signal Processors", Paper of Technical Group, TGMBE 88-134, IECE Japan, 1989.