coustics'08
Paris
June 29-July 4, 2008
www.acoustics08-paris.org

Dependency of recognition rate on number of words for text-independent speaker recognition using Vector Quantization

Hidenori Shimizu and Tetsuo Funada

Div. of Electronic Eng. and Computer Sci., Kanazawa Univ., Kakuma-machi, 920-1192
Kanazawa-shi, Japan
funada@t.kanazawa-u.ac.jp

In this research, we discuss the speaker recognition using the Kohonen's feature map. The model for each speaker consists of a feature map. Two kinds of feature vector MFCC and FTTSS of each speaker are used for training the map, and they are quantized into a specific vector on the feature map. The feature FTTSS is used to develop a robust speaker recognition system under noisy condition. Using the map, we conduct speaker recognition(identification and verification) based on vector quantization (VQ) distortion. In particular we examine dependency of recognition rates on number of utterance words for recognition using the administrative division name of Japan as the utterance words. According to speaker identification experiments, increasing the number of words in recognition more than three words, this system can attain a correct rate of 100% for input speech of 40 speakers. We also examine the influence of the difference of uttering period. Moreover, we show the results of speaker recognition by using Hidden Markov Model (HMM) for comparison with VQ method.

1 Introduction

Recently, the problem of the information security often becomes important in an information-oriented society. Various methods are proposed for the personal identification. There is a problem that the reliability of personal identification by speech is lower than other biometric identification such as fingerprints and blood vessels while it is easy. In this research , aiming to raise the reliability of the personal identification by speech, we examine dependency of recognition rate on number of words for speaker recognition using VQ. Two kinds of feature MFCC and FTTSS are used. The former is a conventionally used one and the latter is a noise-robust one proposed by the present authors. Kohonen's self-organizing feature map is used for the VQ, and VQ distortion is used as a measure for identification.

2 The Kohonen Feature Map

Kohonen's self-organizing feature map is a two-layered network that can organize a topological map from a random starting point. The resulting map shows the natural relationships among the patterns that are given to the network. The network combines an input layer with a competitive layer of processing units, and is trained by unsupervised learning. All interconnections go from the first layer (input layer) to the second (competitive layer); the two layers are fully interconnected, as each input unit is connected to all of the units in the competitive layer. Figure 1 shows this basic network structure.

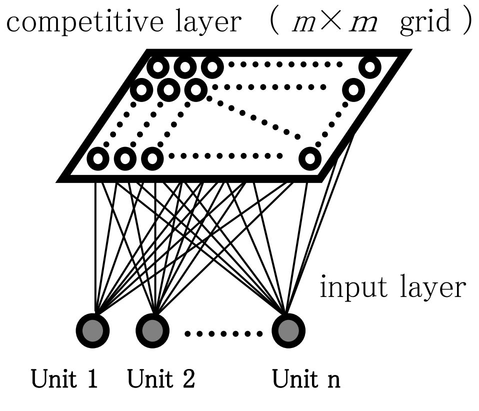

Fig. 1 The basic network structure for the Kohonen feature map[4]

When an input pettern is presented, each unit in the first layer takes on the value of the corresponding entry in the input pattern. The second layer units then sum their inputs and compete to find a single winning unit. Each interconnection in the Kohonen feature map has an associated weight value. The initial state of the network has randomized values for the weights. The weight values are updated during the training of the network. Neighborhood preservation from input layer space to output competing space is an essential property of the Kohonen feature map.

3 FTTSS

The typical feature used to recognize speech includes Mel LPC cepstrum and FFT Mel cepstrum, etc. However, the recognition rate decreases remarkably when speech is recognized by using these feature under noisy condition. In previous paper[6], we proposed a new feature based on power spectral derivatives to develop a robust speech recognition system under noisy condition. The proposed feature is calculated by the following three steps:

(1) Converting the power spectral derivative at each frequency to the ternary scale {+1, 0, -1} for 64 channel frequencies in mel-scale,

(2) Smoothing the ternary values in the time domain at each frequency,

(3) Inverse Fourier transforming the smoothed values in the frequency domain at every 10ms.

The resultant time sequences of Fourier coefficients are the proposed feature, which we call "mel-frequency Fourie Transform of Ternarized Spectral Slope (FTTSS)".

3.1 PSD Filter

Let $S_n(f_c)$ be short time power spectrum of signal x_n, where n is temporal parameter and f_c is frequency.

$$S_n(f_c) = \left| \sum_{k=-\infty}^{n} x_k w_{n-k} \exp(-j2\pi f_c kT) \right|^2 \quad (1)$$

$$w_n = \exp(-2\pi f_b nT)$$

Where w_n is a time window function, f_b determines the length of the window (or the band width of corresponding spectral window) and T is the sampling interval of input speech. Power spectral derivative $S_n'(f_c)$ can be calculated with using PSD filter proposed by the author[1] such as

$$S_n'(f_c) = 2\,\mathrm{Im}\{y_n(0)y_n^*(1)\} \qquad (2)$$

The block diagram of PSD filter is shown in Figure 2. The transfer function of the components are as follows:

$$H_A(z) = \frac{1}{1 - az^{-1}} \qquad (3)$$

$$H_B(z) = \frac{az^{-1}}{1 - az^{-1}} \qquad (4)$$

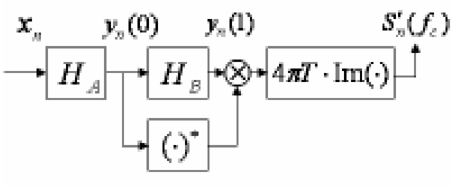

Fig. 2 Block diagram of PSD filter

3.2 Procedure for extracting FTTSS

A bank of N-channel PSD filters is constructed. Parameters f_c and f_b of PSD filters are set to equal resolution on mel scale and 100Hz, respectively. The output of each PSD filter is ternarized to -1, 0 or +1 according to the difference between the value of power spectral derivative and the threshhold. The ternarised value of channel c is smoothed by a second order low pass filter, and is denoted by symbol A_c. Accidental variations by noise or glottal volume flow is expected to be decreased by smoothing operation. The discrete Fourier transform(DFT) of the N –channel smoothed values is calculated every 10ms in order to decrease information quantity such as in the case for MFCC.

$$F_k = \frac{1}{N}\sum_{c=1}^{N} A_c \exp(-j\frac{2\pi ck}{N}) \quad (k = 0,1,\cdots,K\,) \quad (5)$$

The parameter F_k is referred to FTTSS. The procedure for extracting FTTSS is illustrated in Figure 3. Channel size N is set to 64 in the present experiment. Total 30-dimensional parameters including the FTTSS F_k up to 14-order (K = 14), logarithmic power P and each delta value are used for recognition. These parameters are passed through a band pass filter for extracting effective modulation frequency components.

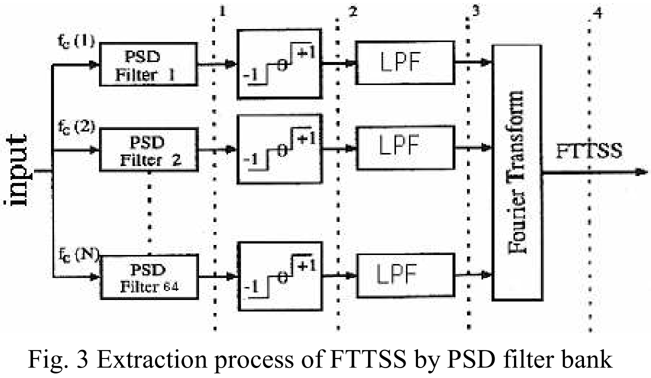

Fig. 3 Extraction process of FTTSS by PSD filter bank

4 Speech data and analysis condition

The speech data uttered by 40 Japanese male speakers is used. Each speaker utters 50 words (the administrative divisions name etc. of Japan) five times for each. These speech data were sampled at 16 kHz and 16bit. Moreover, speech data uttered at two different time (after one month and one year) are used for five speakers among these. We carry out experiment by extracting Mel-frequency spaced filter-bank cepstral coefficients (MFCC) and FTTSS from these speech data, and using it. The feature extraction condition is shown in Table 1.

Feature	MFCC	FTTSS
Number of dimension	30 (MFCC1-14, log Pow, ΔMFCC1-14, Δlog Pow)	30 (FTTSS1-14, log Pow, ΔFTTSS1-14, Δlog Pow)
Sampling frequency	16kHz	16kHz
Size of frame/ equivalent frequency resolution	25ms	100Hz
Frame shift / extraction rate	10ms	10ms

Table 1 Condition for feature extraction

5 Experiment

The feature extracted as shown in Table 1 is used as an input vector to the Kohonen feature map. The Kohonen feature map is trained by 10000 utterances (50 speakers and 200 utterances (50 words × 4 utterance) of each speaker). The network has 30 input units and a 20×20 grid of output units. The distance(VQ distortion) of each speaker's feature map is calculated by using one utterance of the remainder, and the speaker is recognized.

5.1 Speaker identification experiment

The distance(VQ distortion) between input speech and each speaker's feature map is calculated, and the speaker number in the feature map which shows the minimum distance is assigned as identified person. The case where the person in question is correctly identified is classified to be a correct answer. We examine the dependency of recognition rate on repetitions of training the map and number of words used for recognition.

5.2 Speaker verification

After ID that specifies a speaker and input speech are given to map, the distance between the feature map of the specified speaker and the input speech is calculated. The speaker is dismissed as others if the distance is above a specified threshold, or accept as the person in question if it is below the threshold. The system is evaluated by the variation of the threshold and the error rate especially by crossover error rate (Error rate that person in question dismissal rate P(N|s) equals to pretender acceptance rate P(S|n)).

5.3 Influence of utterance in different period

In general, a high recognition rate is obtained when speech data for training and for recognition is uttered in simultaneous period, and the performance of the recognition system falls when speech are uttered in different period. Therefore, we carry out speaker recognition experiment for 5 speakers using the speech data of one month and one year after the training speech has been collected, and examine the influence by the time difference on recognition rate.

5.4 Comparison of VQ with using HMM

To compare speaker recognition experiment using VQ with other method of recognition, we carry out speaker recognition using HMM. The structure of HMM is left to right model (L-R HMM) and ergodic model (EHMM) of Gaussian single mixture output probability.

5.5 Comparison of FTTSS with MFCC under noisy condition

In this experiment, speech data uttered by 10 speakers are used. Noise added speech data for evaluation was generated by adding clean speech data with car noise, crowd noise, workstation noise, and white noise at Signal-to-Noise ratio (SNR) of 10dB. These noises except white noise are obtained from JEIDA. Using the noisy speech data, we carry out the speaker identification experiment to confirm noise robustness of FTTSS.

6 Results and discussion

6.1 Speaker identification

Figure 4 shows recognition rates of the speaker identification using one word (one utterance) when repetitions of training are increased from 10 to 500 times (40 speakers , using speech uttered in simultaneous period of training). Figure 5 shows the recognition rate when the speaker identification is carried out using the distance sum of two words and three words. According to figure 4, recognition rates become high with an increase in repetitions of training. Moreover, according to figure 5 the recognition rate at which the mistake is very few in using two words was obtained. Using three words, this system can correctly recognize all speakers of the input speech. When two or more words were used, the average recognition rate of the combination of all words was used as a recognition rate.

6.2 Speaker verification experiment

Speaker verification experiment using the feature map of 500 repetitions of training and the combination distance of two or more words (each one time utterance) was carried out. Figure 6 shows the change of P(N|s) and P(S|n) when the threshold is changed (using the total distance of three words). Moreover, to evaluate the speaker verification performance with increasing number of words, numerical values of P(N|s) and P(S|n) are shown in Table2. The improvement of the speaker verification performance with an increase in the number of words was confirmed though the threshold that both P(N|s) and P(S|n) become 0 was not obtained.

6.3 Influence of utterance in different period

Figure 7 shows the result of speaker identification experiment with speech uttered in different period. In the speaker identification, the recognition rate of 100% was obtained even by one word. On the other in case of using more than five words, about 7% recognition rate decreased because of the difference of period (one year). The result of the speaker verification experiment is shown in Table 3. In speaker verification, the crossover error rate has increased 5.5% and system deterioration by the difference of period appears greatly even when ten words are used.

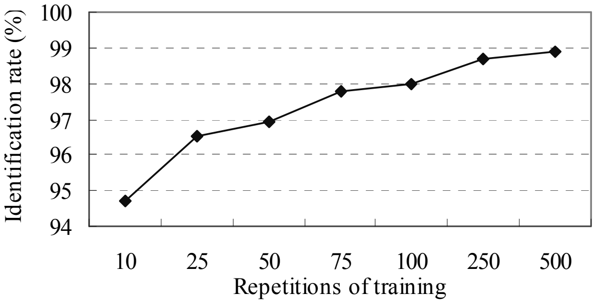

Fig. 4 Dependency of recognition rate on repetitions of training

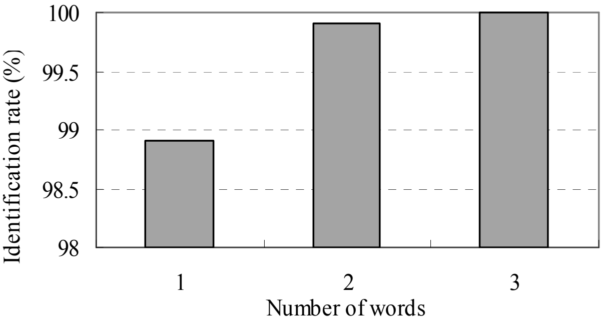

Fig. 5 Dependency of recognition rate on number of words

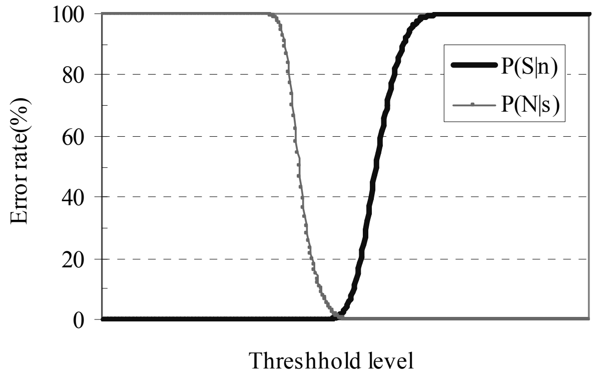

Fig.6 Threshold versus error rate (using 3 words)

	P(N|s) (%)		
Number of words	3	5	10
P(S|n):0%	58.3	28.2	10.9
P(S|n):1%	2.8	0.8	0.2
P(S|n) = P(N|s)	1.7	0.9	0.5

Table 2 Error rate of speaker verification

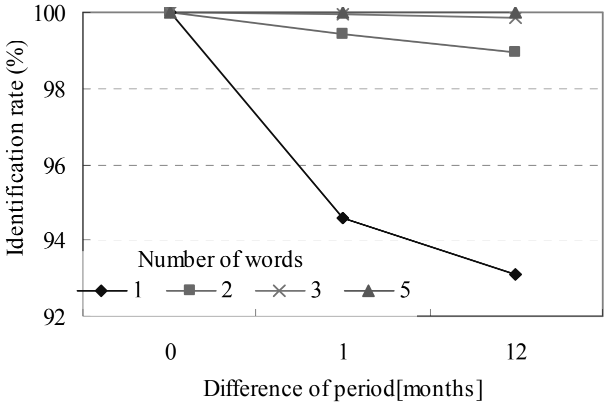

Fig. 7 Speaker recognition with speech uttered in different period

	Difference of period (months)		
	0	1	12
Number of words	Crossover error rate (%) (P(N|s) = P(S|n))		
3	0.07	3.5	9.5
5	0.005	1.8	7.5
10	0.0002	0.8	5.5

Table 3 Speaker verification with speech uttered in different period

6.4 Comparison of speaker recognition with using HMM

Dependency on state number and model topology are calculated on the same condition as VQ (training : 200 utterances, recognition : 1 utterance). Because considerably low recognition rates were obtained in speaker identification using L-R HMM, only recognition rates using EHMM are shown in Figure 8. The improvement of the recognition rate can be confirmed by increasing the number of states of HMM. Dependency on state number and model topology are calculated on the same condition as VQ (training : 200 utterances, recognition : 1 utterance). Improvement of the recognition rate can be confirmed by increasing the number of states of HMM.

6.5 Comparison of speaker identification using FTTSS and MFCC under noisy condition

Figure 9 shows the result of a comparison of speaker identification using FTTSS and MFCC under noisy condition. At clean speech, the recognition rate of MFCC is higher than FTTSS. However, under every kind of noisy condition, the recognition rate by FTTSS is higher than MFCC. The difference appeared to be remarkable especially in case of the white noise.

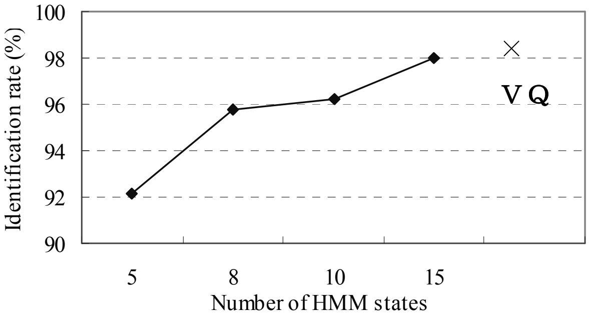

Fig. 8 Comparative result of VQ and HMM

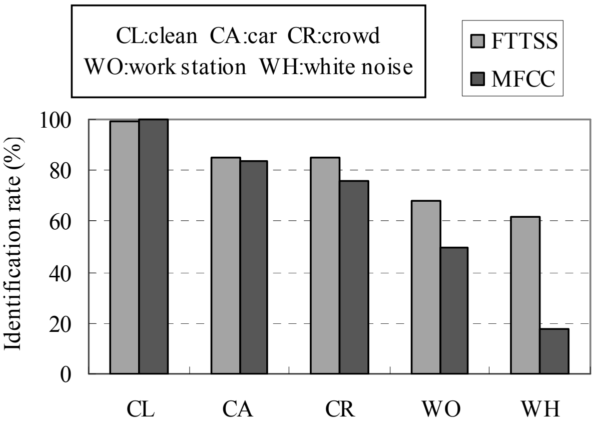

Fig. 9 Recognition rates for noisy speech data(SNR 10[dB])

7 Conclutions

The purpose of present research was to confirm the effectiveness of using VQ by Kohonen feature map to speaker recognition. Features MFCC and FTTSS are used for recognition. The feature FTTSS is used to develop a robust speaker recognition system under noisy condition.

From the experiment we have confirmed that the recognition performance using VQ by Kohonen feature map is better than using HMM and the recognition rates increase by increasing repetitions of training and increasing the number of words for recognition. Moreover, in speaker verification experiment using FTTSS, we have confirmed noise robustness of FTTSS compared with MFCC.

Though 200 utterances (50 words ×4 utterances) were used to learn in this research, it is necessary to examine the use of continuous of speech sentences than the word utterance in future research. Examining dependency of the threshold on difference of period between training and recognition in the speaker verification is also necessary.

References

[1] S.B.Davis, P. Mermelstein, “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences”,ICCC Trans . ASSP-28, pp.357-366(1980)

[2] Sadaoki Furui, “Speech information processing”, MORIKITA PUBLISHING Co.Ltd. in Japanese (1998)

[3] Sadaoki Furui, “Advances in Speech Signal Processing”, Marcel Dekker,Inc.(1992)

[4] J. Dayhoff, “Neural Network Architectures”, Van Nostrand Reinhold (1990)

[5] T.Kohonen, “Self-Organizing Map”, Springer Verlag,Tokyo. (2005)

[6] Richard O. Duda, “Pattern Classification, 2/Edition ”, New Technology Communications Co. Ltd.(2001)

[7] Tetsuo Funada,Tomoaki Ookubo, Hideyuki Nomura, “An Application of a Feature from Ternarized Spectral Derivatives to Word Speech Recognition” , ICA, Tu.P2.13 (2004)

[8] Megumi Umeno, Tetsuo Funada, Hideyuki Nomura, “Noisy Word Recognition Using a Feature Based on Ternarized Spectral Slope”,IEEE . ISCIT ,pp.1575-1579 (2007)