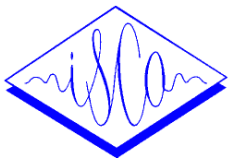
ISCA Archive
http://www.isca-speech.org/archive



First International Conference on
Spoken Language Processing (ICSLP 90)
Kobe, Japan
November 18-22, 1990

Phoneme Recognition Using a Hierarchical Time Spectrum Pattern

Kei Miki

Human Interface Laboratory, OKI Electric Ind. Co., Ltd.
550-5 Higashi-asakawa-cho, Hachioji-shi, Tokyo, 193 Japan

Abstract

This paper describes a new vector-quantization-based phoneme recognition method which uses the hierarchical time spectrum pattern (TSP). The phonetic feature is discussed by a mutual information and a posteriori probability between vector quantization codes and phoneme label codes. The TSP of Mel-scaled LPC cepstra and the power-change pattern (PCP) are used as acoustic parameters. Input speech is firstly vector-quantized by the PCP codebook. Secondly it is vector-quantized by the TSP of which the codebooks are classified by the PCP-VQ code. Hierarchical TSP-VQ improves performance of phoneme classification compared with only the TSP-VQ. A frame-label matching experiment on a speaker-independent condition with the JEIDA Japanese speech database of connected 4-digit uttered by 16 males and 16 females, shows 79.4% of recognition accuracy using the method. The experimental result indicates that the proposed method is highly effective.

1. Introduction

To achieve the goal of a large vocabulary, speaker-independent and continuous speech recognition, high-performance phoneme recognition method has been required. For a system to yield this high performance, dynamic features of input speech must be extracted, in addition to statical features. A method to extract effective dynamic features, for example the power-change pattern (PCP), and to create its codebook on the criterion of mutual information, is proposed by K. Shirai et al. and high performance was attained on the speaker-independent phoneme recognition experiments[1]. On the other hand, several effective dynamic features, the time spectrum pattern (TSP) and regression coefficients of spectral parameters are proposed and their effectiveness for improving both word and phoneme recognition accuracy, has been verified[2-9]. This paper proposes a phoneme recognition method based on vector quantization (VQ) using the PCP codebook and the hierarchical TSP codebooks. The TSP codebooks are multiply created from learning data which are previously classified speaker characteristics, mainly concerned with age and sex[10]. Furthermore, one of the TSP codebooks is made for each group classified by the PCP-VQ code, or hierarchically structured by PCP-VQ code. This paper is divided as follows. In section 2, we overview the phoneme recognition system. Section 3 describes acoustic features and their codebooks. In section 4, we explain the phoneme score calculation referred to by the table of probabilities and conditional entropies between phonemes and the VQ codes. In section 5, the evaluation results for the phoneme recognition experiments are described. Section 6 is the conclusion.

2. System Overview

Fig. 1 describes the block diagram of the proposed phoneme recognition method. The recognition method consists of five parts. They are acoustic feature extraction, power-change pattern VQ , hierarchical time spectrum pattern (TSP) VQ by plural TSP codebooks, optimal TSP VQ code selection and scoring phoneme.

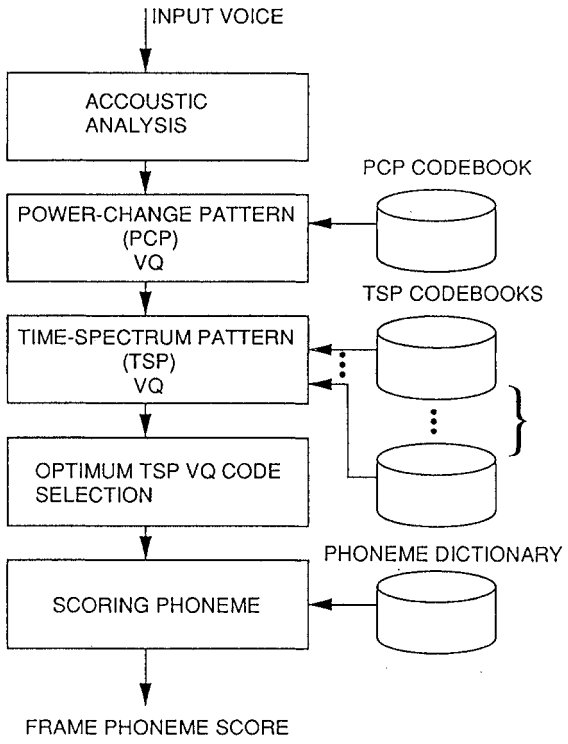


Figure. 1 Block Diagram of the system

In the first part of the method, the power-change pattern (PCP) and the time spectrum pattern (TSP) are extracted as acoustic features of the input speech. In the second part, a PCP-VQ code is derived from the PCP codebook. The PCP-VQ code classifies the input into several groups. In the third part, the several candidates of TSP-VQ codes and their quantization errors are derived from the plural TSP codebooks attached to the PCP-VQ code. In the fourth part, the optimal TSP-VQ code train, discriminated by the accumulated VQ distortion, is acquired after the end point of the input speech. Finally, after the VQ processes, the optimal TSP-VQ code train is converted into the phoneme probability vector train by looking up the phoneme dictionary that consists of two items : the posterior probabilities and the conditional entropies of phoneme labels.

3. Acoustic Features

3.1 Power-Change Pattern (PCP)

A power-change pattern shows the feature of time dynamics in input speech logarithmic power. Because the speech power is a robust feature, input speech is effectively classified into several groups by the PCP-VQ method. Moreover speech power is less dependent on a speaker's characteristics than speech spectrum, a small PCP codebook is able to accurately classify input speech.

Considering the i-th frame of an input speech, if the power of the i-th frame is denoted by B_i, the power-change pattern vector $B_i(k)$ is defined by,

$$B_i(0) = B_i \quad , \quad B_i(\pm k) = B_i - B_{(i\pm k)} \quad ; \quad k = 1,2,...,N_p \qquad (1)$$

Creating a PCP codebook as follows :

A PCP codebook is produced by the *Linde, et al.* VQ algorithm. The number of N_0 is the codebook size. After this, the PCP clusters are integrated into several clusters via the merging algorithm to minimize the loss of the mutual information between the VQ codes and the phoneme labels. N_1 is the code size after it has merged.

We created the codebook, in the condition $N_p = 4, N_0 = 128, N_1 = 16$, according to the report[1].

3.2 Time Spectrum Pattern (TSP)

As mentioned above, the TSP feature is the time variance of a spectral parameter. The TSP expresses exactly allophonic variations between spectral parameters and phonemes. However, the TSP codebook size will be very large and also the TSP-VQ will require much computation, because the codebook must cover all the allophonic variations. Therefore, to reduce the TSP-VQ computation as much as possible is very important without less deterioration of recognition accuracy. We discuss three reduction methods. Firstly, we select spectral parameters which express phoneme features compactly. Because the parameters compactly express the features, the TSP of them also expresses allophonic variations compactly. We evaluate several spectral parameters and TSP features by several phoneme recognition experiments in the condition of the same codebook size. The second reduction method is a pre-classification. The effective pre-classification lead to the inhibition of vector quantizing with the mismatched TSP code vectors which are attached to different classes. Therefore, it results in the reduction of the TSP-VQ computation. Our pre-classification method is the vector quantization by the PCP codebook, mentioned in section 3.1. The effectiveness of the method is discussed in section 5.2. The last reduction method is a division of the TSP codebook according to a speaker's characteristics. The division separates the spectral variations, which result from speaker's characteristics as the vocal tract length, from the variation caused from other factors. Using the multiple TSP codebooks, higher phoneme recognition accuracy can be hoped for. In this paper, as a primitive approach we create two TSP codebooks, a male TSP codebook and a female TSP codebook.

The TSP codebook is produced as follows :

All the same speaker group data are vector-quantized by the PCP codebook, and are divided to N_1 subsets. Next, each subset is divided by phoneme label. Therefore, we have smaller $N_{ph} \times N_1$ subsets. N_{ph} is the number of phoneme categories, 18. Finally, TSP centoroids are extracted from each of the smaller subsets by the *Linde, et al.* algorithm.

4. Phoneme Scoring

Phoneme scoring means estimating phoneme probabilities frame by frame referring to the TSP-VQ code train. The process is divided into two steps. The first step is selecting the optimal TSP codebook and the second step is the estimation of phoneme probabilities referred to by the optimal TSP-VQ code train.

4.1 Selecting the Optimal TSP codebook

In the case of using multiple TSP codebooks, to select the optimal TSP codebook is a very significant problem. The Optimal TSP codebook selection requires much greater accuracy than phoneme detection. Our criterion for TSP codebook selection is the distance between the TSP codebook and input speech. The distance is, we defined, the average value of each frame VQ-distortion distance, the VQ error.

The selecting algorithm is as follows :

TSP-VQ error calculated each frame and each codebook. After the end point of input speech, the average value of the VQ error is calculated for each codebook. And the codebook given the minimum value, which is named the optimal codebook, is selected.

Only the optimal TSP-VQ code given by the optimal TSP codebook, is used following phoneme estimation. The optimal TSP-VQ code is named simply the TSP-VQ code except where specially noted.

4.2 Phoneme Scoring Algorithm

A phoneme score is calculated by using the posterior probabilities and the conditional entropy between the TSP-VQ code and the phoneme labels, both of which are set to training data. To estimate phoneme probabilities frame by frame, it is necessary to use the information from the adjacent frame. To make the optimal phoneme decision depending on the neighboring frame, the probability by conditional entropy of each code should be weighted. Then the probability based on the effective feature is emphasized by the lower entropy. The weighted score $\Phi(x_i)$ for the phoneme x_i

$$\Phi(x_i) = \prod_{j=i-N's}^{i+N's} P(x_i \mid Y_j) \,/\, H(x_i \mid Y_j) \qquad (2)$$

Y_j is the TSP-VQ code of the j-th frame, P means probability and H is entropy. Ns (= $2N'$s+1) means considering frame width.

5. Experimental Evaluation

5.1 Speech Data and Analysis Condition

The speech database consists of connected 4-digits uttered in Japanese by 75 male and 75 female speakers. For the training utterances, 16 male and 16 female speakers are used. The reference data are the other 16 male and 16 female speakers. Input speech are passed through a telephone line, and are digitized at a rate of 8-kHz. The speech data is pre-emphasized by difference filter and Hamming window is applied. And then, linear predictive cording (LPC) analysis is performed on all frames. The LPC analysis is a 30-order analysis. The length of the Hamming window is 24 msec., and the speech data is analyzed by 8 msec. interval. The LPC Mel-cepstral coefficients (15-order), which are final spectral parameters, are derived from LPC cepstral coefficients (30-order) by using the *Oppenheim's* procedure. The sound logarithmic power also is calculated for each frame. Phoneme labels and speech segmentations are given by inspection in advance. In all the experiments, the number of discriminated phoneme categories is 18, as shown in Table 1., which appears in Japanese 10 digits.

Table 1. The discriminated Phoneme Categories.

a , i , u , e , o	Vowel
j	Semivowel
I , U	Devocalized vowel
n , N	Nasal
r	Liquid
g , k	Plosive
z , h , s , c	Fricative
.	Pause

5.2 Experimental Evaluation of PCP-VQ

We evaluate the effectiveness of the PCP-VQ and the hierarchical TSP-VQ from the view point of recognition accuracy and computations. The reference condition is the method only used for simple TSP-VQ. The experimental result is shown in Table 2. "TcP" means that non-hierarchical simple TSP codebook consists of LPC Mel-cepstral coefficients (c), the reference. "PCP+TcP" means both the PCP and the hierarchical TSP codebook are used. The result shows effectiveness of the method, the PCP-VQ and the hierarchical TSP-VQ.

5.3 Relationship between TSP Dimension and Phoneme Recognition Accuracy

The change of phoneme recognition accuracy was evaluated when the Nt, the time length of TSP, was varied from 1 frame to 13 frames. The longer time TSP strictly expresses to the dynamic features of spectrum. However, it is more disadvantaged against the variations of spoken speed. Furthermore, it is remarked that it also has the disadvantage of an increase in computational complexity because the number of dimensions is increased. Conversely, the shorter time TSP has less computation, but it cannot precisely follow dynamic spectral features. On the other hand, phoneme scoring is considered in the preceding and succeeding Ns frame TSP-VQ code, so it seems that Nt has some connection with Ns. Therefore, the recognition accuracy is evaluated changing both Nt and Ns in the experiment. The result of the experiment is given in Table 3. "TcP_k" means the time spectrum pattern, which consists of Mel-cepstra, over k(=Nt) frame. In the training data, the recognition accuracy is not improved as Nt increases in the case of fully Ns. However, an increase in Nt yields some improvement in accuracy in the open (speaker independent) data. The bigger Ns yields an improvement in accuracy in both training and open data. The tendency of improvement is seen as long as the width of TSP is short, and is also seen in the training data. The recognition accuracy improved until Ns reached 27 frames (216 ms).

Table 2. Phoneme Recognition Result with PCP-VQ. "TcP_k" means TSP of cepstra with k-th frame width. A frame phoneme score considered with neighbor 27 frame TSP codes. Recognition accuracy is calculated from the frame match, and average of all phoneme categories.

Feature	Computation	Recognition Accuracy	
		Training	Open
TcP_{13}	1.0	90.5%	66.2%
PCP+TcP_{13}	0.1	89.5%	74.2%

Table 3. Phoneme Recognition Result with Various Width TSP. "PCP + TcP_k" means PCP-VQ and TSP-VQ of cepstra with k-th frame width.

A frame phoneme score considered with neighbor Ns frame TSP codes.

(a) Recognition in Training Data

Ns (frame)	1	3	11	19	27
PCP+TcP_1	57.1%	68.2%	80.7%	85.3%	87.8%
PCP+TcP_5	64.5%	72.8%	82.7%	86.7%	88.9%
PCP+TcP_9	68.7%	76.4%	84.0%	87.3%	89.4%
PCP+TcP_{13}	71.2%	78.2%	85.1%	87.7%	89.6%

(b) Recognition in Open Data

Ns (frame)	1	3	11	19	27
PCP+TcP_1	45.8%	53.8%	62.6%	65.4%	65.3%
PCP+TcP_5	53.0%	58.7%	66.1%	68.8%	69.3%
PCP+TcP_9	57.5%	63.2%	69.7%	72.2%	73.1%
PCP+TcP_{13}	61.2%	66.3%	72.5%	74.1%	74.8%

5.4 Experimental Evaluation about Several Dynamic Features

The delta cepstral coefficient (Δc), which is calculated the regression coefficient of a cepstrum as time axis, is

proposed as one of the dynamic spectral features and the effectiveness of the parameter has been reported in many experiments. So we compared several features with various combinations cepstrum (c) and Δc. A weight ratio between c and Δc, is optimized for some preliminary experiments. The result of the experiment is given in Table 4. The result shows combined features c and Δc, especially in the case of using different time span features simultaneously, attained a higher score than with TcP, time cepstral pattern. Then we evaluated the effectiveness of the matrix quantization method with the time pattern codebook, which consists of c and Δc. This result is also shown at Table 4. The result means that time pattern quantization is also effective, in the case of using "(c,Δc)" features.

5.5 Multiple TSP Codebooks

Using the best performance feature, we created two codebooks, a male TSP codebook and a female TSP codebook, and evaluate the same condition. Each of the divided codebooks is half the size of the codebooks used in previous experiments. The evaluation result is also shown in Table 4. The multiple TSP codebook method shows a little improvement of phoneme recognition accuracy.

Table 4. Phoneme Recognition Results with Several Dynamic Features. "(c,Δc_5)" means single codebook with multi features, cepstrum and Δcepstrum over 5 frame. "(c,Δc_9),(c,Δc_{13})" means double codebooks with (c,Δc_9) and (c,Δc_{13}), respectively. A frame phoneme score considered with neighbor 27 frame TSP codes.

Feature	Recognition Accuracy	
	Training	Open
PCP+(c,Δc_5)	87.2%	70.3%
PCP+(c,Δc_9)	89.7%	75.4%
PCP+(c,Δc_9),(c,Δc_{13})	90.6%	76.9%
PCP+TSP_9	89.4%	73.1%
PCP+TSP_{13}	89.6%	74.7%
PCP+T(c,Δc_9)P_9	90.0%	78.2%
PCP+T(c,Δc_9)P_{13}	90.2%	79.0%
PCP + multi-T(c,Δc_9)P_{13}	90.7%	79.4%

6. Conclusion

We proposed the phoneme recognition method based on vector quantization (VQ) using the power-change pattern (PCP) codebook and the plural hierarchical time spectrum pattern (TSP) codebooks. The recognition method consists of five parts. They are acoustic feature extraction, power-change pattern VQ, hierarchical time spectrum pattern (TSP) VQ by plural TSP codebooks, optimal TSP-VQ code selection and scoring phoneme using the posterior probabilities and the conditional entropies between the optimal TSP-VQ code and phoneme labels. The current method and current features have given a performance of 79.4% phoneme recognition accuracy frame by frame on speaker-independent continuous 4-digit speech in Japanese. In the next step, a performance of full phoneme recognition will be evaluated and other spectral parameters and dynamics will be investigated.

Acknowledgment

The author wish to thank Mr. Chihara, Dr. Eki, Dr. Noto, Mr. Tabei and Mr. Yazu for their help and advice. He also would like to thank Mr. Saito (director of the laboratory), Mr. Nose (general manager of the division) and Mr. Morito (manager of the section) for their advice and encouragement during the work.

References

[1] Katsuhiko Shirai, Noriyuki Aoki and Naoki Hosaka "Multi-Level Clustering of Acoustic Features for Phoneme Recognition Based on Mutual Information." Proc.ICASSP-89, pp604-607 (May 1989)

[2] Sadaoki Furui "On the use of Hierarchical Spectral Dynamics in Speech Recognition." Proc.ICASSP-90, pp789-802 (April 1990)

[3] Sadaoki Furui "Isolated Word Recognition Based on Emphasized Spectral Dynamics" IEICE Technical Report, SP85-77, pp597-604 (January 1986) (in Japanese)

[4] Masafumi Nishimura "HMM-Based Speech Recognition Using Dynamic Spectral Feature" Proc.ICASSP-89, pp298-301 (May 1989)

[5] S.Roucos and M.O.Duhum "A Stochastic Segment Model for Phoneme-Based Continuous Speech Recognition" Prco.ICASSP-87, pp73-76 (1987)

[6] S.Morial, S.Makino and K.Kido " Phoneme Recognition in Continuous Speech Using Phoneme Discriminant Filters" Proc.ICASSP-86, pp2251-2254 (1986)

[7] A.Waibel, T.Hanazawa, G.hinton, K.shikano and K.Lang " Phoneme Recognition Using Time-Delay Neural Networks" Technical Report TR-1-0006, ATR Interpreting Telephony Research Laboratories (Oct. 1987)

[8] Erik McDermott and Shigeru Katagiri "Shift-Invariant, Multi-Category Phoneme Recognition using Kohonen's LVQ2" Proc.ICASSP-89, pp81-84 (May 1989)

[9] M. Endo, S.Makino and K.Kido "Phoneme Recognition using Modified LVQ2" IEICE Technical Report, SP89-50, pp33-40 (Sep. 1989) (in Japanese)

[10] S.Nakagawa, H.Shirakata, M.Yamao and T.Sakai "Considering on Speaker Grouping By Sex and Age for Automatic Speech Recognition" Jour.IECE Jpn, J63-D, 12, pp1002-1009 (1980) (in Japanese)