ORGANIZATION OF PHONEMIC SPACE REPRESENTED BY THE UNITS OF SPECTRA AND SPECTRAL CHANGES

Katsuhiko Shirai and Kazunori Mano

Department of Electrical Engineering,
Waseda University
3-4-1, Ohkubo, Shinjuku-ku, Tokyo 160, JAPAN

ABSTRACT

This paper describes a method of organization of phonemic space for phoneme recognition. Phonemic space is obtained by clustering speech spectra and spectral changes. Power change, LPC cepstral coefficients and the differences of LPC cepstral coefficients are used to represent the characteristics of the spectral contour and spectral change. The efficiency is shown by an experiment of phoneme recognition.

INTRODUCTION

There are many factors which make it difficult to extract phonemic features precisely. Some of the factors are as follows.
(1) In continuous speech, boundaries between adjacent phonemes are uncertain and it is difficult to segment correctly.
(2) There are many variations in phoneme patterns.
(3) As the characteristics of phonemes exist not only in spectral contours but also in spectral changes, both static and dynamic properties in speech signals must be considered as acoustic features.

Vector quantization (VQ) method is an efficient method to encode speech signals[1]. We have used the VQ technique as a clustering method to extract phonemic features frame by frame[2][3]. In this paper, an organization of phonemic spaces with a VQ technique is discussed and we consider the relation between acoustic features represented by VQ codes and their phonemic features which belong to the clusters of the VQ codes.

REPRESENTATION OF ACOUSTIC FEATURES AND PHONEMIC FEATURES FOR CLUSTERING

Acoustic Features

Acoustic features defined in each frame are LPC cepstral coefficients called Level 1 feature, changes of LPC cepstral coefficients called Level 2 feature and power change. The Level 1 feature is calculated in a frame and denoted by the following.

Level 1 feature : $(C1(1),\ldots,C1(n))$,

where n is the order of LPC analysis. The Level 2 feature and the power change are defined as the differences between the parameters in the first half and the second half of the frame. If the LPC cepstral coefficients in the first half and the second half are denoted by $(C21(1),\ldots,C22(n))$ and $(C22(1),\ldots,C22(n))$ and the powers P1 and P2, the Level 2 feature and the power change in the frame are defined as follows.

Level 2 feature : $(\triangle C2(1),\ldots,\triangle C2(n))$, where

$$\triangle C2(i) = C21(i)-C22(i). \quad (i=1,\ldots,n)$$

Power change :

$$\triangle P = (P2 - P1) / P1$$

The Level 1 feature shows a spectral contour which represent a static property in a frame. The Level 2 feature corresponds to the change of the spectrum. This feature is efficient to describe the precise movements of spectrum in a frame, especially in transient parts of speech such as consonant-to-vowel(CV) sounds. The power change shows a global changes such as the change from silence or unvoiced sound to voiced one.

Phonemic Features

A label called a frame label which is composed of three phonemic symbols is assigned to each frame by visual inspection before clustering. For example, if a frame belongs to a transient part of speech /.pa/, where /./ means silence, the frame labels such as /..p/, /.pp/, /ppa/, /paa/ or /aaa/ are sequentially yielded according to the position of the frame. The frame label of /..p/, means that the frame contains silence /./, in more than half part of the frame and a sound of /p/ is following the silence in the frame. The /aaa/ means the frame exists only in vowel part, that is, the frame is almost stationary.

CLUSTERING METHOD BASED ON VQ ALGORITHM

Phonemic features are related to acoustic features by clustering. The main reason of using clustering method is that it makes the speech frames into groups which have both acoustically and phonemically similar properties. Each frame is characterized by code numbers of the produced cluster and the frame labels in the cluster.

As for the clustering, vector quantizer design method which is a slightly modified one proposed by Linde, Buzo and Gray[1] is adopted. The modified points are that the centroids to be split are determined by considering kinds of the frame labels for effective distributions of centroids. That is, more centroids are assigned to the clusters which have a lot of kinds of frame labels and less centroids to the clusters which have only one or two frame labels. By this modification, the quasi-optimality of the VQ method is not kept any more, but it is more useful to extract phonemic features.

For example, if a cluster has the frames which have the same frame labels, the centroid of the cluster is not split in the preceding procedure because the phonemic features of the cluster is sufficiently represented by the frame label. Such clusters appear in stationary parts. On the other hand, if a cluster has various kinds of frame labels, the phonemic features in the domain of the cluster are not described by the centroid and it means that more centroids are necessary to obtain phonemically unified clusters. Such clusters mainly exists in transient parts.

ORGANIZATION OF PHONEMIC SPACE

The above clustering method is applied to each set of frames to organize phonemic space.

Before clustering, all the speech frames are classified into three parts called the ascending, flat and descending parts by the degree of power change ΔP in each frame. The ascending part contains such sounds like consonant-to-vowel, silent-to-consonant or vowel-to-stronger vowel. The flat part contains almost stationary parts of vowels, nasals and fricative consonants. In the descending part, the sounds such as vowel-to-consonant, vowel-to-silence or vowel-to-weaker vowel are contained. By this pre-classification, it is possible to avoid grouping of the frames which have entirely different frame labels, even in the case that the acoustic distortion between the frames is small.

Clustering is performed in each part of the three parts with Level 1 features and Level 2 features, respectively and six codebooks composed of centroid vectors and sets of frame labels are produced. The phonemic space is organized by the distributions of the centroids and the frame labels which belong to the corresponding cluster in each part and each level.

EXPERIMENT OF PHONEME RECOGNITION

For an evaluation of the phonemic space which is represented by codebooks and frame label sets, an experiment of phoneme recognition is carried out. Figure 1 shows the diagram of extracting phonemic features. When a frame is analyzed, the power change is calculated and one of the part number of the power change is assigned to the frame and according to the LPC cepstrum and the cepstral differences, codes of Level 1 and 2 are given to the frame. Output of the frame labels is obtained from the intersection of the sets of frame labels in Level 1 and 2. By symbolic processing the sequences of the frame labels, phoneme sequences are produced.

The result of the cumulative recognition rates of phonemes for one male speaker is shown in Figure 2. In the experiment, the codebooks are generated from 800 syllables and 100 city names are used for the recognition. The sampling frequency is 12.5[kHz]. The frame length in Level 1 is 32[ms] and the interval of analysis is 16[ms]. The number of VQ codes in each part is about 256. In Fig.2 the phoneme recognition rates are about 91% in vowel sounds and 73% in consonants in the first candidate. Within 3 candidates, the rates increase to 99% in vowels and 89% in consonants.

CONCLUSION

A method to make phonemic space base on the spectrum and the spectral difference was proposed. The efficiency of this method was evaluated.

REFERENCES

[1] Linde, Y., Buzo, A. and Gray, R. M., "An algorithm for Vector Quantizer Design," IEEE Trans. Commn., Vol.COM-28, No.1, pp.84-95, January 1980.

[2] Shirai, K. and Mano, K., "A Clustering Experiment of the Spectra and the Spectral Changes of Speech to Extract Phonemic Features", Signal Processing, Vol.10, No.3, April 1986.

[3] Shirai, K. and Mano, K., "Feature Extraction of Phonemes by clustering the spectra and the spectral changes in continuous speech. Proc. of IASTED Inter. Symp. of Applied Signal Processing and Digital Filtering, pp.201-204, June 1985.

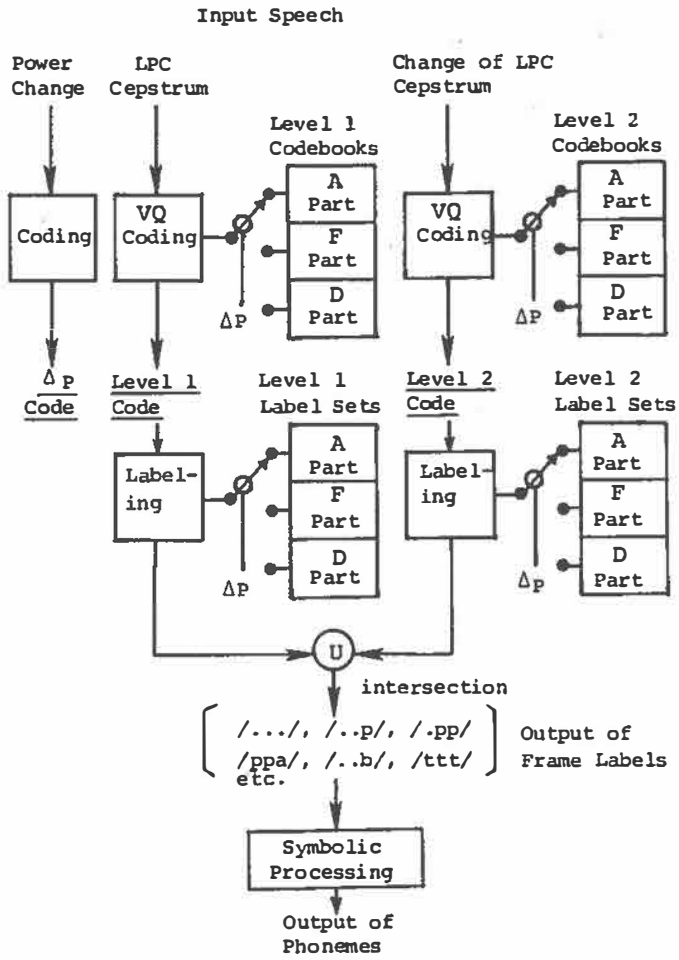

Fig.1 Diagram of Extracting Phonemic Features

A Part : Ascending Part
F Part : Flat Part
D Part : Descending Part

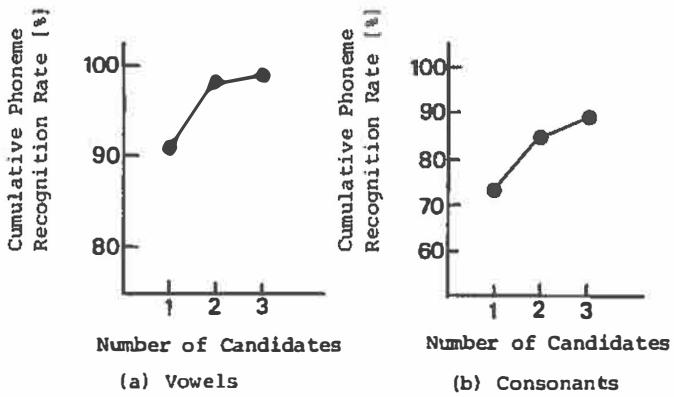

Fig.2 Cumulative Recognition Rate of Phonemes