AN EFFICIENT CLUSTERING ALGORITHM AND ITS USE IN PHONEME SYNTHESIS

Jin Zijian*, Ke Youan*

ABSTRACT

An efficient clustering algorithm called ECA based on long sequences of speech data is proposed for designing vector quantizer. The ECA algorithm, which uses a new splitting scheme and applies the K-means method only in a subset of the training data, greatly reduces the computation requirement. Its special use in phoneme speech synthesis is investigated. Comparisons between the new algorithm ECA and some algorithms presented earlier like LBG, MKM are made based on 12,000 frames of speech data. Simulation results show the advantages of ECA when both the performance and the computation time are considered.

INTRODUCTION

It is well known that language is made up of finite phonemes. Based on phonemes, all the words in language can be formed. However, the speech quality of a synthesis system is very bad if we simply combine these phonemes together. This is because that in human pronunciation, the physical organism naturally smoothes the change from one phoneme to another, and forms many transitional frames between them. In order to improve the speech quality, we must find additional "phonemes" to represent the transitional frames.

Suppose that there is a map from the pronunciation region to the hearing region. Here the "hearing region" is defined as the region in which the distance is proportional to the difference discerned by human ear. In the same way, we can define the term "pronunciation region". Experiments have shown that this map satisfies the mel-log characteristics, i.e., it is logarithm in magnitude, and in frequency, it can be expressed as:

\[ f' = 1000 \times \log(1 + f/1000) \]  \hspace{1cm} (1)

In this formula, \( f \) is the actual physical frequency or frequency in pronunciation region, and \( f' \) is the frequency discerned by ear or frequency in hearing region. If we divide the hearing region into \( M \) clusters, within each cluster take the center to approximate all the points in it, and if the clusters and \( M \) are properly chosen, the speech can be restored well based only on these centers.

Now the problems can be restated as following:

(a) Give \( N \) speech samples \( X = \{x_1, x_2, \ldots, x_N\} \) in hearing region, how to select the codebook \( C = \{c_1, c_2, \ldots, c_M\} \) so that the average intracluster distortion

\[ E = \frac{1}{N} \sum_{i=1}^{N} ||X_i - Q(X_i)|| \]  \hspace{1cm} (2)

reaches a minimum. Here \( Q \) is a map from \( X \) to \( C \).

(b) How to use the codebook in speech synthesis and what is the resulting performance.

* Prof. Ke and Mr. Jin are with the Dept. of E.E., Beijing Inst. of Tech., P.O. Box 327, Beijing, P.R. China
OUTLINE OF THE PROCEDURE

Fig.1 gives the procedures for phoneme speech synthesis. (a) Generation of the codebook, the main part of which—the codebook generation algorithm, will be discussed in the next section. (b) Generation of the word library which includes the code form of pitch, energy and linear prediction parameters of each training frame in each word. (c) The speech synthesizer. It is the final existing form.

The procedure presented here has two distinguished aspects. First, the mel-log parameter has been adopted. This is because of its close correspondence to human ear characteristic and because it is much easy for computation. We only need to compute the Euclidean distances, i.e., if \( x_1 = (x_{11}, x_{12}, \ldots, x_{1L}) \) and \( x_2 = (x_{21}, x_{22}, \ldots, x_{2L}) \) are the mel-log parameters of two speech frames, the distance between them is

\[
\| X_1 - X_2 \| = \frac{1}{L} \left[ \sum_{l=1}^{L} (X_{1l} - X_{2l})^2 \right]^{1/2}
\]

(3)

The simplicity in computation is also of great benefit to the word library generation. Second, since linear prediction parameters have been successfully used in speech synthesis, the codebook2 (in 10-order reflection coefficients) is taken other than codebook1 (in mel-log parameters).

CODEBOOK GENERATION ALGORITHM

How to generate the codebook is of great importance to phoneme speech synthesis. It has a direct influence on the quality of the synthesized speech. The criterion of the codebook quality can be defined as the average intracluster distortion \( E \) in formula (2). A clustering algorithm called K-means method has been widely used in this process because of its many attractive properties. Like the computer minimization of a two-variable function, when an initial value is given, the K-means method iteratively refines clusters and cluster centers to achieve a minimum or a local minimum. Gray et al. (ref.1) have proposed a splitting scheme to obtain the initial guess by means of a perturbation vector in the cluster center. A very simple algorithm can be formed by splitting \( X \) \( M \) times and then proceeding the K-means method, here we call it the simplest method (SM) algorithm for convenience.
The K-means cluster method outputs only a local optimal value. The final result depends on how well the initial guess is made. This dependence is even apparent on speech data because of their discrete distribution. In some algorithms presented earlier like LBG(ref.1) and MKM(ref.2), the K-means method has been inserted into the splitting process in order to improve the initial guess. For example, in MKM algorithm, the K-means method is used after every splitting. It is obvious that the MKM algorithm gives the best result, but the computation load is also prohibitive. Here we propose a new algorithm in which the K-means method is still used. It has a higher efficiency and is based on two assumptions: (a) The radius nonunity of each cluster. Thus, splitting using a perturbation vector always along the longest radius seems leading to the best result. This assumption can be approved by observing the advantage of SM2 over SM in Fig.4. SM2 has the same structure as SM except for adopting the new splitting scheme. (b) The attenuating tendency of the splitting perturbation of K-means method. In the algorithm of MKM, when a cluster is splitted and a new K-means process begins, the movement of elements from one cluster to another is called the splitting perturbation. It is easily imagined that the perturbation attenuates as the distance to the splitted cluster increases, i.e., perturbation is utmost in clusters which near the cluster just been splitted, while there is no perturbation at all where distances are remote. Therefore the K-means method should be proceeded only in the nearest K clusters other than all of the current M clusters. In this way the computation requirement is reduced about the square of M/K times while the performance remains almost unchanged.

The new algorithm called ECA is illustrated in Fig.2, in which the parameter D, is the product of intracluster distortion and number of elements in ith cluster, and MARK is the number of clusters which are not change in the current K-means process. The parameter K is estimated according to MARK, if MARK=0, then K=K-1, otherwise K remains unchanged. This technique fits the actual case well, as we can see it from Fig.3.
SIMULATION RESULTS

Simulations have been proceeded to compare the relative performance of ECA and testify the feasibility of the phoneme synthesis procedures. The ECA algorithm is compared with LBG, MKM with respect to the decreasing rate of average distortion and the computation complexity. The results are shown in Fig.4 and Fig.5. Simulations are done at IBM-3031 with about 8 minutes of speech data of a male speaker. The figures show that the new algorithm—ECA outputs almost the best results, and on the other hand, computation requirement is greatly reduced, especially when the final number of clusters $M_{max}$ is big. In fact, in the case of $M_{max} = 512$, the computation load of ECA is nearly the same as that of LBG. The improvement of SM2 over SM shows the effectiveness of the new splitting method. As the next step, synthesis experiments have been conducted based on the resulting codebook($M_{max}=512$). The phoneme synthesis system cuts down the bitrate of 1500 bits per second of original scalar coded system to about 950 bits per second, while the speech quality remains as good as the original one, judged independently by a group of individuals.