

# RESEARCH ON SPEECH UNITS MODELING IN CONTINUOUS SPEECH RECOGNITION

HE Xiaodong, LIU Jian, ZHOU Jianlai, YU Tiecheng

Speech Processing Laboratory, Institute of Acoustics, Chinese Academy of Sciences

P.O. Box 2712, Beijing 100080, P.R.CHINA

Email:  [{hexd, lj, zjl, tcyu}@speech1.ioa.ac.cn](mailto:{hexd, lj, zjl, tcyu}@speech1.ioa.ac.cn)

<http://www.speech.ioa.ac.cn>

## ABSTRACT

It is often expedient to consider using more than one single HMM to characterize a speech unit. In this paper, we suggest a new speech units modeling method based on analysis of parameters of HMMs obtained by preliminary training. By analyzing the emission probability function of a state of a HMM obtained by segmental k-means training, we can obtain the distribution of the source data and determine the splitting of that model. The experimental results, based on totally 264,500 phoneme occurring in the 9180 sentences from 60 speakers, showed that approximate 10% improvement of the recognition rate of the basic phoneme was achieved.

## I. INTRODUCTION

Speech unit modeling is a main problem of speech recognition. Because of the variability in the speech production and/or the processing, there are some advantages to use more than one single HMM to characterize one speech unit. Normally, there are two kinds of unit modeling strategies: model clustering (a bottom-up approach) and model splitting (a top-down approach). The former is based on the concepts of starting from a large set of units and merging similar units based on some type of clustering procedure. A general procedure for model merging is based on the consideration of speech source likelihood, which has been implemented successfully by Lee [1]. Moreover, some variations on model clustering have been proposed, including knowledge-based allophonic clustering [2], and CART-based phonetic clusters, in which a decision tree is used to choose the most reasonable clustering

sequence based on phonetic considerations. The later is based on the concepts of starting from a small set of units and iteratively splitting the units. There exist two or more clusters in the speech source data, which are labeled with a same phoneme symbol. These clusters represent several units with different acoustic properties and should be identified. Once the separation is achieved, we have effectively created multiple models of that unit [3]. Several ways to create the clusters for each unit has been proposed [4].

In traditional segmental k-means training procedure, speech feature vector sequences were aligned to several states of HMMs by Viterbi decoding, and these feature vectors aligned to the same state were used to estimate the model's emission probability function. So, the clustering details of the speech source data were reflected in the propriety of the model's emission probability function. In our study, a new speech unit splitting method was suggested based on the analysis of the parameters of DHMM (Discrete Hidden Markov Model) of speech units and segmental k-means training algorithm.

In section II, we present the fundamentals of speech units selection, training, and splitting. Section III describes the implement of the unit splitting method we proposed. The experiment results will be shown in section IV, and a conclusion is finally given in section V.

## II. SPEECH UNITS MODELING

### A. *Speech unit selection and training*

In many speech recognition systems, phonelike unit was adopted as the speech unit, in which the basic phoneme set of speech is used. It has an inherent

advantage in lexicon building. However, because the phoneme unit is defined based on linguistic similarity but modeled based on acoustic similarity, there normally exists some inconsistency. Assuming that for a unit symbol, represent a phoneme, there is some inherent internal distribution of training tokens that naturally clusters into two or more groups, which represent classes of speech data that are all associated with that same phoneme symbol, but have different spectral properties. In order to reflect the real situation of speech variability, multiple models of that unit are needed.

Once we obtain a preliminary set of HMMs associated with the same phoneme, each of these HMMs represents a sub-class of that phoneme's source data, we can build a multi-lexicon Viterbi decoding and obtained the final HMM by segmental k-means training. The problem is how to get that preliminary set of HMMs.

### B. Clustering Analysis and splitting of DHMM

In case of using DHMM, a codebook is needed. First we need to classify these entire source speech feature vectors to  $M$  clusters by vector quantization [5] technology. After that, we calculate the center vector for each cluster and assign an index to each center vector. These center vectors and indexes compose an  $M$ -code codebook. These  $M$  center vectors can represent the entire speech feature vectors roughly.

In DHMM, the symbol density  $b_{\lambda_j}(k)$  represents the probability of that feature vector, which is represented by center vector with code number  $k$ , appears at the state  $j$  of model  $\lambda$ . Then the discrete symbol density matrix  $B_{\lambda_j}$  represents the distribution of coded feature vectors in state  $j$  of model  $\lambda$ . According to the clustering analysis of  $B_{\lambda_j}$ , we can determinate if one unit should be split and how to split it.

For example, as to phoneme  $p$ , we only build one preliminary model  $\lambda$  for it at first. However, assuming that there exists an inherent distribution of source speech data which all labeled with  $p$ , it could be divided into two or more clusters. The clustering distribution of speech data can be reflected by the discrete symbol density matrix  $B_{\lambda_j}$ . If we examine the histogram of probability of each code for state  $j$  of model  $\lambda$ , we get curves similar to those show in Figure 1.(a). The two peaks appearing in that histogram indicate that the

distribution of source speech data labeled with phoneme  $p$  cluster to two clusters. Accordingly, two models are needed to build for phoneme  $p$ . As shown in figure 1, the unit splitting is to create two new B distributions such as figure 1 (b) & (c) from the preliminary B distribution shown as figure 1 (a).

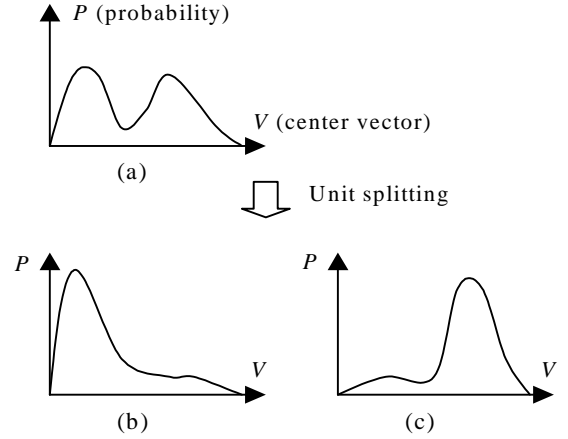


Figure 1: Illustration of unit splitting

This is for state  $j$  of model  $\lambda$ .  $V$  means the center vector of codebook, which represent the real speech feature vector.  $P$  is the probability of that the feature vector which has code number  $k$  appears at the state  $j$  of model  $\lambda$ .  $P(V_k) = b_{\lambda_j}(k)$ .

Focus our consideration on the discrete symbol density matrix  $B_{\lambda_j}$  of state  $j$  of model  $\lambda$ . These  $M$  center vectors compose a data set, each of them is associated with weight  $b_{\lambda_j}(k)$ . Many succeed clustering algorithms [6] can be used to divide that set to several clusters. Assuming that these  $M$  center vectors cluster to  $N$  clusters, then  $N$  new matrix  $B$  can be created as following:

As to the  $n^{\text{th}}$  matrix  $B_n$ , use  $b_k$  represent  $b_{\lambda_j}(k)$ , let  $P_1 = \sum b_{k'}$ ,  $P_2 = 1 - \sum b_{k'}$ ,  $k'$  belongs to cluster  $n$ . Then

$$\hat{b}_k = \begin{cases} b_k + \frac{b_k}{p_1} \cdot p_2 \cdot r & k \in \text{cluster } n \\ b_k \cdot (1 - r) & k \notin \text{cluster } n \end{cases} \quad 1 \leq k \leq M$$

$$\sum_{1 \leq k \leq M} \hat{b}_k = 1$$

$r$  represents a scale coefficient.  $r$  is between 0.5 to 0.9.

With these new matrix  $B$ ,  $N$  DHMMs are built for that single phoneme. They have the same parameter as the original DHMM, except that each of them has a new matrix  $B$  at state  $j$ , then these speech data clusters labeled

with the same phoneme can be distinguished. According to this, the original mono-model lexicon was extended to multi-model lexicon as figure 2.

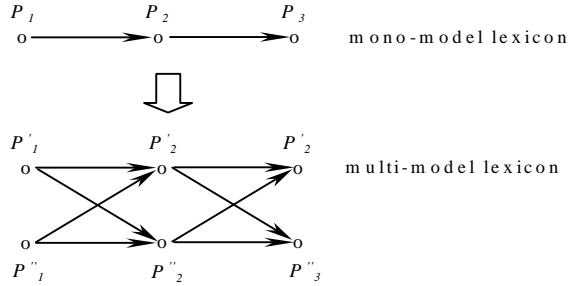


Figure 2: Illustration of lexicon extending

Eventually, a segmental k-means training procedure with multi-model lexicon Viterbi decoding is employed to obtain the final modeling and training result.

### III. IMPLEMENT OF UNIT SPLITTING

First, we need to select a set of basic phoneme. In Chinese, a syllable is normally made up of two parts, an initial and a final. The entire consonants, vowels, and silence, totally 60 phonemes, compose the basic phoneme set. Consider that some phonemes have several distinct context-dependent pronunciations that can be labeled with different phoneme symbol directly, the basic phoneme set is composed by 60~70 phonemes. In our study, 65 phonemes were selected to compose the basic phoneme set. It includes 38 vowels, 26 consonants, and 1 silence.

In order to split DHMM, we implement clustering analysis of the discrete symbol density matrix  $B$  of DHMM as following.

For convenience, we use  $b_k$  to represent  $b_{i,j}(k)$ . As to the data set composed by these  $M$  center vectors  $V_k$ , assuming that they cluster to  $N$  clusters.

*Step 1: Centroid calculation:*

Define  $V_{cn}$  as the centroid of cluster  $n$ ,

$$V_{cn} = \frac{\sum b_{k'} V_{k'}}{\sum b_{k'}}; \quad k' \in \text{cluster } n$$

*Step 2: Classification:*

$$V_k \in \text{cluster } n^*; \quad n^* = \arg \min_{1 \leq n \leq N} (|V_k, V_{cn}|).$$

*Step 3: Judging of convergence:*

Define  $V_c$  as the centroid of the whole data set,

$$V_c = \frac{\sum b_k V_k}{\sum b_k} = \sum b_k V_k; \quad 1 \leq k \leq M$$

Define  $SNR$  as a splitting criterion.

$$SNR = \frac{\sum_{1 \leq n \leq N} \sum_{k' \in \text{cluster } n} b_{k'} |V_{k'}, V_{cn}|^2}{\sum_{1 \leq k \leq M} b_k |V_k, V_c|^2}$$

Iterate step1-2 until the convergence of  $SNR$ .

*Step 4: Judging of model splitting:*

If  $SNR$  is larger than a threshold  $SNR_{th}$ , it means that the source speech data has an inherent distribution clusters to  $N$  cluster, then we execute the model splitting method stated above. Otherwise, it means that the model should not be split.

Now we conclude the complete model splitting procedure here:

- 1) Select a basic phoneme set by linguistic rule, build a mono-model lexicon.
- 2) Obtain the preliminary HMMs by segmental k-means training with mono-model lexicon.
- 3) Cluster and split these preliminary HMMs and create a new set of initialized HMMs. Then build a multi-model lexicon.
- 4) Obtain the final HMMs by segmental k-means training with multi-model lexicon.

### IV. EXPERIMENTS

Several experiments were conducted to prove the validity of this method. These test results are based on totally 264,500 phonemes occurring in the 9180 sentences from 60 speakers. 6120 sentences by 40 speakers were used for training, the remain for testing. All these speech data are selected from the "863 National Project Chinese Mandarin Speech Corpora". We studied the relation among split phoneme number, clusters number  $N$  for each phoneme, and the splitting threshold  $SNR_{th}$ . The result is given at figure 3. We also gave a result of the comparison of accuracy and average likelihood in case of before model splitting and after model splitting. The result is shown at table 1.

The test result indicated that approximate 10% improvement of the recognition rate of basic phonemes was achieved.

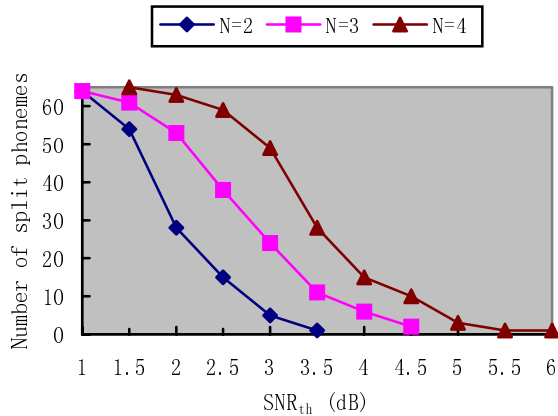


Figure 3: Curves of  $N$ ,  $SNR_{th}$  and split phoneme number  $N$  represents the cluster number of one phoneme.  $SNR_{th}$  is a threshold. When  $N$  is fixed, for each phoneme  $p$ , if  $SNR(p) > SNR_{th}$ , then phoneme  $p$  should be split. The "Number of split phonemes" in this figure represents the total of all phonemes that should be split.

	Number	Before splitting		After splitting	
		Accur-acy %	Like-hood	Accur-acy %	Like-hood
Split phonemes	28	55.4	-3.85	78.6	-3.49
Un-split phonemes	37	62.3	-3.88	60.7	-3.82
Total phonemes	65	59.1	-3.87	68.2	-3.68

Table 1: The effect of model splitting

In case of  $N=2$ ,  $SNR_{th}=2.0$ , the likelihood is the frame-average log likelihood of these phonemes.

## V. CONCLUSION

According to the experiment results, we can conclude that the speech unit splitting method presented in this paper is an efficient way of unit modeling. As shown in table 1, after model splitting, the average likelihood of these split phonemes increase greatly, this suggests that the source speech data were classified more accurate to their inherent distribution. This lead to that the values of the discrete symbol density matrix  $B$  of DHMM is concentrated. As a result, the recognition rate of these basic phonemes is increased.

On the other hand, although the recognition accuracy of these split phonemes is increased remarkably, the

accuracy of these un-split phonemes declines slightly. This is because that the segmental k-means training is not a discriminative training, for example, if A belong to model  $\lambda$  but B don't. The update of model  $\lambda$  for increasing  $P(A|\lambda)$  maybe also increase  $P(B|\lambda)$ . This will make B easy to be mis-recognized.

In the next step of our research, a more comprehensive investigation about the selection of basic phoneme set, the clustering and splitting method, and the multi-model lexicon is still needed.

## REFERENCES

- [1] K.F. Lee, *Automatic Speech Recognition--The Development of the SPHINX System*, Kluwer Academic Publishers, Boston, 1989.
- [2] L. Deng et al., "Acoustic Recognition Component of an 86,000 word Speech Recognizer," *Proc. ICASSP 90*, Albuquerque, NM, pp. 741-744, April 1990.
- [3] C.H. Lee, L.R. Rabiner, R. Pieraccini, and J.G. Wilpon, "Acoustic Modeling for Large Vocabulary Speech Recognition," *Computer Speech and Language*, 4: 1237-165, January 1990.
- [4] L.R. Rabiner and B.H. Juang, *Fundamentals of Speech Recognition*, PTR Prentice Hall, Englewood Cliffs, NJ, 1993.
- [5] Y. Linde, A. Buzo, and R.M. Gray, "An algorithm for vector quantizer design," *IEEE Trans. Commun.*, COM-28, pp.84-95, January 1980.
- [6] J. Mike, *Classification Algorithms*, Collins Professional and Technical Books, London, UK, 1985