# GENERALIZED PHONE MODELING BASED ON PIECEWISE LINEAR SEGMENT LATTICE

*Hiroaki KOJIMA* *Kazuyo TANAKA*

Electrotechnical Laboratory, AIST, MITI
1-1-4 Umezono, Tsukuba, Ibaraki 305, Japan

## ABSTRACT

The goal of this work is to model phone-like units automatically from spoken word samples without using any transcriptions except for the lexical identification of the words. In order to implement this task, we have proposed the "*piecewise linear segment lattice (PLSL)*" model for phoneme representation. The structure of this model is a lattice of segments, each of which is represented as regression coefficients of feature vectors within the segment. In order to organize phone models, operations including division, concatenation, blocking and clustering are applied to the models. This paper mainly report on blocking and clustering. Experimental results for isolated word recognition task is that the recognition rate is significantly improved by blocking the segments and by clustering the segments within a block. We get sufficient performance for the task with the models consist of at most 128 clusters of segment patterns.

## 1. INTRODUCTION

The ultimate goal of this work is to generate robust speech recognition models. In the paradigm of traditional stochastic method of speech recognition, the process to improve robustness sometimes tend to be a wrong spiral of making precise models and increasing size of training samples. On the other hand, a human infant seems to be able to acquire various knowledge of phonological system properly. Motivated by this fact, we adopt the task to acquire phone-like unit and its structure inductively from speech samples. The task is to form phoneme models and a phoneme set from spoken word samples without using any transcriptions except for the identification of each word in a lexicon. We call this task "*phonological concept formation*"[1].

The basis of this idea is that an appropriate models should better be formed throughout interactive communication than be defined a priori. We assume that robustness is derived from a flexible task in which the least necessary knowledge is provided. In our approach, we assume that the essential factor in acquiring a phonological system is to discover the relationships between utterances and their meanings. In order to make this practical for experimental purposes, they are simplified to the relationships between isolated spoken word samples and their lexical identification.

Related studies are for example: [3][4] as spoken language acquisition, and [6][7][8][9] as automatic design of speech recognition or speech coding units.

## 2. PIECEWISE LINEAR SEGMENT LATTICE (PLSL) MODEL

In order to implement this task, we have proposed the "*piecewise linear segment lattice (PLSL)*" model as a framework for phoneme representation[2]. A spoken word sample is modeled by dividing it into several segments, each of which is represented as regression coefficients of feature vectors within the segment, that is, $\{a_k(k = 1, .., K), b_k(k = 1, .., K)\}$ in the following equation.

$$\hat{y_k}(t) = a_k(i,j)(x(t) - \overline{x(t)}) + b_k(i,j)$$

where $\hat{y_k}(t)$ is the least square estimation of $k$-th component of feature vector $\boldsymbol{y}(t)$ at the $t$-th frame, $x(t) = S \cdot t$ ($S$ is a constant), and $\overline{x(t)}$ is a mean value of $x(t)$.

An initial word model of PLSL is obtained by bundling the models of the samples which are belong to the same word (**Fig.1(a)**). The lattice of a word model is then transformed to be a more phone-like structure by matching and aligning between the sequences of the segments.

The optimum segmentation of each sample which minimize the total distortion within the sample can be efficiently calculated using a dynamic programming (DP)

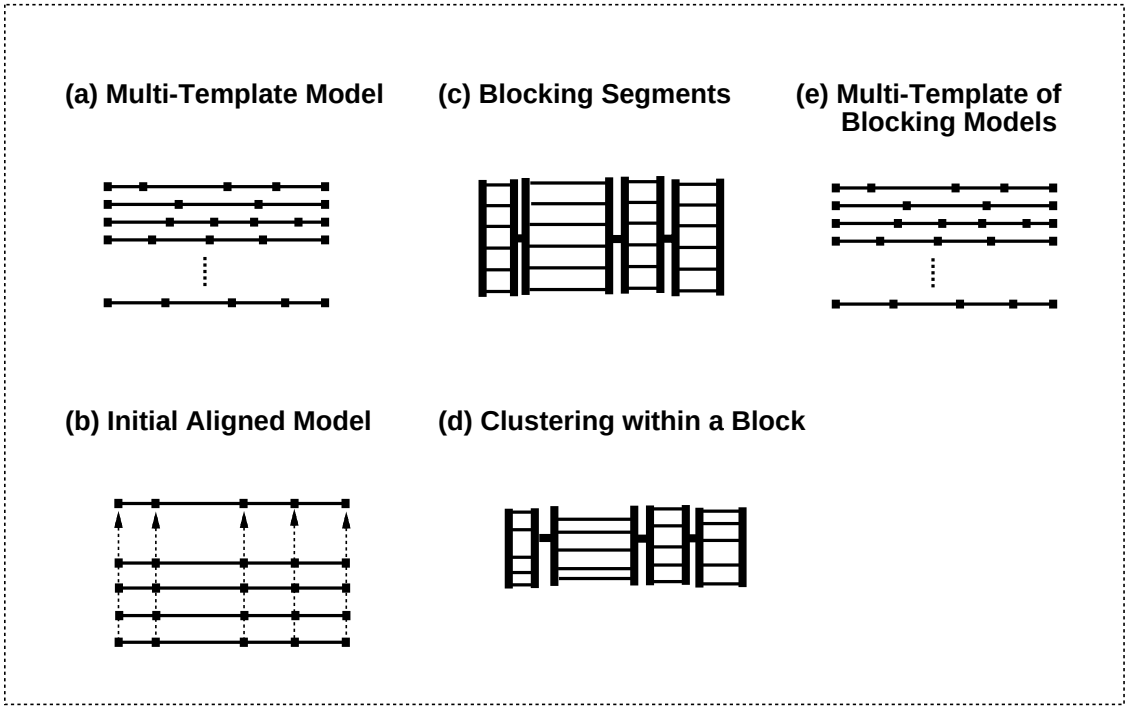


**Figure 1:** Structure of Models

procedure, if the number of division is fixed. The optimum division into $N$ segments is calculated with the following recurrent formulas as $g(N, J)$ where $J$ is the total frame size of a sample.

$$g(1,j) = d(1,j)$$
$$g(n,j) = \min_i [g(n-1,i) + d(i,j)] \quad (\text{if } n > 1)$$

The distortion within a segment from the $i$-th frame to the $j$-th frame in a sample is defined as follows:

$$d(i,j) = \frac{1}{K} \sum_{t=i}^{j} \sum_{k=1}^{K} (y_k(t) - \hat{y}_k(t))^2$$

The proper number of divisions is determined as the number $N$ which minimize the following AIC criterion. Assuming distributions of residual vectors $\boldsymbol{y}(t) - \hat{\boldsymbol{y}}(t)$ as a uniform normal distribution of variance $\Sigma$, the AIC criterion is described as follows:

$$l_{AIC} = \frac{1}{2|\Sigma|} g(N,T) + K \cdot N$$

(Items independent of division are omitted.)

We modify this criterion as follows in order to control the number of division with the parameter $\alpha$.

$$L_{AIC} = g(N,T) + \alpha \cdot N \qquad (1)$$

Matching distance is defined as the total distortion of a sample with a PLSL, which can also be efficiently calculated using DP.

The PLSL model has an ability to represent objects in arbitrary precision. And compared with typical stochastic models, PLSL has the following advantages:

**1)** model parameters can be stably estimated with less samples,
**2)** its structure can be dynamically changed with less calculation.
**3)** hierarchical structure of different precision can be consistently integrated within a lattice.

All these computational characteristics are crucial points to derive phone-like structures.

## 3. EXPERIMENTAL RESULTS

### 3.1. Experimental Conditions

We have examined this model by speaker-independent isolated word recognition. Word samples consist of 492 Japanese words uttered once by 10 male speakers. The model is trained with the samples from 9 speakers, and tested with the samples from the other one speaker. The feature vector consist of 12 cepstral coefficients and a log-

power with 5ms interval. As a reference, an experiment is conducted on the same sample set using HMM which consist of feed forward type word models including almost as many states as the segments in PLSL. The recognition rate is 78.9%.

### 3.2. Unstructured Multi-Template Model

First of all, each sample in the training set is segmented into a sequence of piecewise linear segments as is described in the previous section. In this sample set, the average number of segments per sample is 12.1, and the average length of a segment is $72.4ms$.

Then the unstructured multi-template models are constructed as is shown in **Fig.1(a)**. Each sample in the testing set is recognized by matching it with these models using the DP beam search. The recognition rate is saturated at a beam width of 128, and its recognition rate is 84.2%. Accordingly, we use this beam width in the following experiments.

### 3.3. Initial Aligned Model

The above model is unstructured and not suitable for organizing phone-like structures. By aligning the segments in each word model, this model has rudimentary structures. One sequence selected from the training set is used for a reference pattern of each word model. All the other samples of that word in the training set are segmented by aligning them to the reference. The number of segments, thus become the same for each word (**Fig.1(b)**). The recognition rate with this model is 80.7%

### 3.4. Division and Concatenation

We have conducted experiments applying the processes of division and concatenation to the model **(b)**. Division is done when the matching scores between segments in the alignment process exceeds a threshold. Concatination is the process to add to the lattice new segments for all the pairs of successive two segments. The details of these processes have been reported in in our previous paper[2]. It results 82.9% with division and 84.4% with both division and concatenation.

### 3.5. Blocking the Segments

In this experiment, all the segments which are aligned at the same position in the model **(b)** are bundled into a single sub-lattice as is shown in **Fig.1(c)**, which we call "*blocking*" in this paper In this case, the matching path for the input sample is allowed to cross over the segments in the different sequences of the original samples. The recognition rate is with this model is 91.5%

### 3.6. Clustering the Segments within a Block

Similar segments within a block in the model **Fig.1(c)** are reduced by clustering based on the LBG algorithm, as is shown in **Fig.1(d)**. The recognition rates according to the number of clusters are shown in **Table1**.

| Number of Clusters | 1 | 3 | 5 | 7 |
|---|---|---|---|---|
| Recognition Rates(%) | 97.0 | 93.9 | 93.1 | 91.3 |

**Table 1:**

The recognition rates are usually improved by the increase of the number of clusters. Therefore, relationships between these two shown in the table seems to be irregular in the ordinary viewpoints. In order to investigate the cause of this phenomenon, we added the following experiments.

The first experiments is to change the speaker of the reference pattern in the alignment process. With the models using a single cluster within each segment, the recognition rates for the 9 speakers are between 94.7% to 97.0%, and the average is 95.6%.

The second experiment is to double the beam width (i.e. 256), but there is no significant differences.

These experiments show that this phenomenon is not caused by the dependency on the initial pattern, nor beam width. Another possibility is that the segmentation criterion we adopt here is optimized for the models for few clusters. Although we have not tested yet, we suppose that we need a more temporally precise model in order to utilize the multi-cluster blocking model.

### 3.7. Multi-Template of Blocking Models

By bundling the 9 blocking models described in the previous paragraph, a multi-template model is generated as is shown in **Fig.1(e)**. The recognition rate is 94.1%.

### 3.8. Clustering All of the Segments

We try to reduce the segment patterns by clustering all segments in all of the word models, then a model is generated based on this VQ codebook. This codebook is used as the basis of the phone-like units in the process of phone modeling.

The clustering is applied not only to the model **(e)** but also to **(a)** and **(b)** for comparison. **Table2** shows the

recognition rates according to the number of clusters.

| Number of Clusters | | 64 | 128 | 256 | 512 |
|---|---|---|---|---|---|
| Recognition | **(a)** | 71.8 | 81.1 | 82.9 | 83.3 |
| Rates(%) | **(b)** | 68.3 | 79.3 | 80.3 | 83.3 |
| | **(e)** | 89.6 | 94.3 | 93.7 | 96.1 |

**Table 2:**

This results shows that the multi-template model of blocking models **(e)** is more robust to reduction of codebook size, comparing with the rudimentary models like **(a)** and **(b)**, and that these models are sufficient in the performance by using as many number of units as ordinary phonological units.

## 4. CONCLUDING REMARKS

we have reported on the experiments based on the PLSL model for establishing generalized phone modeling.

The experimental results are summarized as follows:

**1)** the recognition rate is significantly improved by blocking the segments which are aligned at the same position.
**2)** Clustering within a block also improves the performance.
**3)** Multi-template of blocking models does not make significant improvement, but it is robust to reduction of the segment patterns.

We are planning to investigate the optimul integration of temporal division and clustering, and also extend the matching procedure over different word models.

## 5. REFERENCES


1. H. Kojima, K. Tanaka and S. Hayamizu, "Formation of phonological concept structures from spoken word samples," *Proc. ICSLP 92*, pp.269-272(1992).
2. H. Kojima, K. Tanaka: "Organizing phone models based on piecewise linear segment lattices of speech samples," *Proc. EuroSpeech'97*, pp.1219-1222 (1997).
3. F. Fallside: "On the acquisition of speech by machine ASM," *Eurospeech 91*, Keynote 2, (24 Sep 1991).
4. A. L. Gorin, "An experiment in spoken language acquision," *IEEE Trans. SAP*, Vol.2 No.1 (1994).
5. K. K. Paliwal, "Lexicon building methods for an acoustic sub-word based speech recognizer," *Proc. ICASSP-90*, pp.729-732 (1992).
6. M. Y. Hwang and X. Huang, "Subphonetic modeling with markov states - Senone," *Proc. ICASSP-92*, Vol.I, pp.33-36 (1992).
7. J. Takami and S. Sagayama, "A successive state splitting algorithm for efficient allophone Modeling," *Proc. ICASSP-92*, Vol.I, pp.573-576 (1992).
8. M. Bacchiani, M. Ostendorf et al., "Design of a speech recognition system based on acoustically derived segmental units," *Proc. ICASSP-96*, pp.443-446 (1996).
9. M. Saito. M. Masukata and S. Nakagawa, "Automatic aquisiton of speech units for recognition and very low bit coding," *Proc. ASA and ASJ Third Joint Meeting*, pp.1061-1066 (Dec 1996).