# ORGANIZING PHONE MODELS BASED ON PIECEWISE LINEAR SEGMENT LATTICES OF SPEECH SAMPLES

*H. Kojima* and *K. Tanaka*
Machine Understanding Division
Electrotechnical Laboratory, AIST, MITI
1-1-4 Umezono, Tsukuba, Ibaraki 305, Japan
Tel: +81-298-58-5937, FAX: +81-298-58-5939, E-mail: hkojima@etl.go.jp

## ABSTRACT

Aiming at robust speech recognition, we have proposed a framework for "*phonological concept formation*," which is the task of acquiring an efficient representation of phonemes from spoken word samples without using any transcriptions except for the lexical classification of the words. In order to implement this task, we propose 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. Feasibility of the method is discussed with experimental results for isolated word recognition. The recognition rate is improved by applying these operations.

## 1. INTRODUCTION

The ultimate goal of this work is to generate robust speech recognition models. In the traditional paradigm of the stochastic method for speech recognition, the process to improve robustness sometimes tends to be a wrong spiral of making precise models and increasing the size of the training samples. In order to overcome such a limitation, we have proposed a framework for "*phonological concept formation*," which is the task of acquiring an efficient representation of phonemes from spoken word samples without using any transcriptions except for the lexical classification of the words[1][2].

The basis of this idea is that a phonological system 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 element 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 classification. Related studies are [3][4] as spoken language acquisition and [5][6][7] as robust phone modeling.

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

In order to implement this task, we propose the "*piecewise linear segment lattice (PLSL)*" model as a framework for phoneme representation.

Fig.1 shows the structure of PLSL. 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 (**Fig.1(a)**). An initial word model of PLSL is obtained by bundling the models of the samples which are belong to the same word (**Fig.1(b)**). 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 (**Fig.1(c)**).

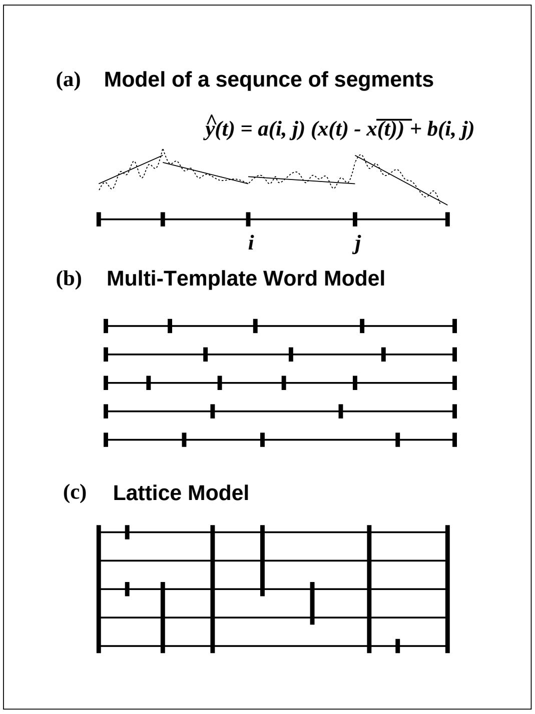
(a) Model of a sequnce of segments
$\hat{y}(t) = a(i, j)\ (x(t) - \overline{x(t)}) + b(i, j)$
i
j
(b) Multi-Template Word Model
(c) Lattice Model

**Figure 1: Structure of PLSL**

The optimum segmentation of each sample, shown in **Fig.1(a)**, can be efficiently calculated using a dynamic pro-

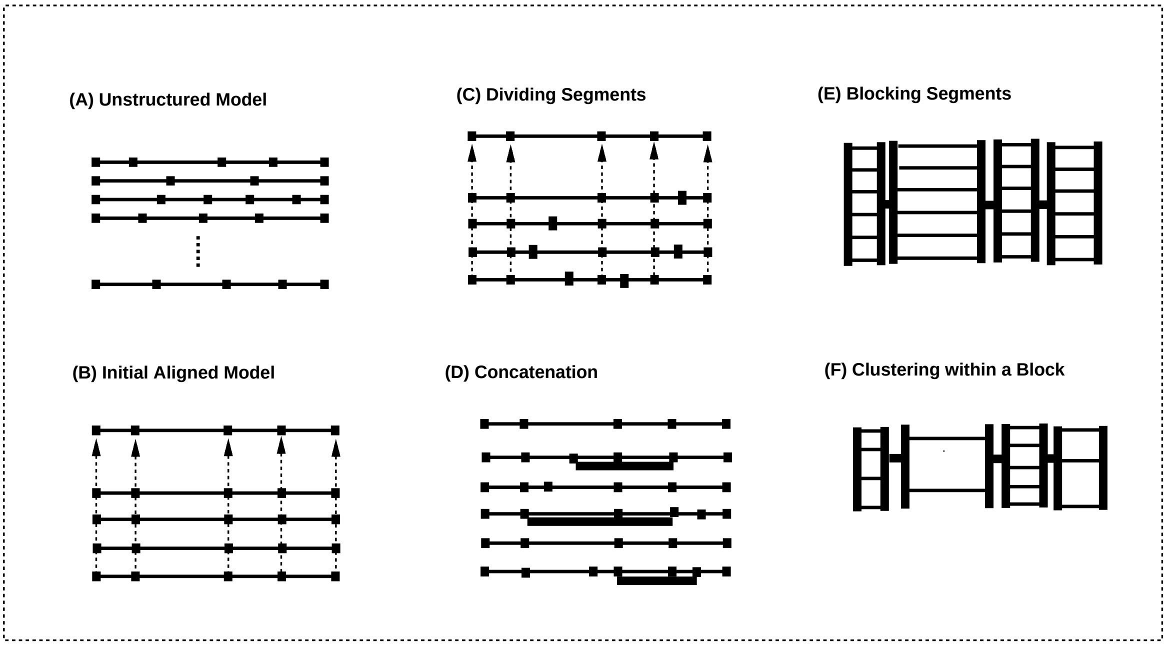

**Figure 2: Structure of Models**

gramming (DP) procedure, if the number of division is fixed. The optimum $N$ division 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 described as follows:

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

where $y_k(t)$ is the $k$-th component of feature vector at the $t$-th frame.

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 regular distribution of variance Σ, 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 $\alpha$.

$$L_{AIC} = g(N, T) + \alpha \cdot N \qquad (2)$$

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. All these computational characteristics are crucial points to derive phonetic structures.

## 3. ORGANIZING PHONE MODELS

The procedure of organizing phone models is described as follows:

### (A) Unstructured Multi-Template Model

First of all, each sample in the training set is segmented into a sequence of piecewise linear segments. Then the unstructured multi-template models are constructed. Each sample in the testing set is recognized by matching it with these models using the DP beam search. (**Fig.2(A)**)

### (B) Initial Aligned Model

The above model is unstructured and not suitable for organizing phonetic 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.2(B)**)

### (C) Dividing Segments

The next step introduces division to the initial aligned model. This division is done when the matching scores between segments in the alignment process exceeds a threshold. (**Fig.2(C)**)

### (D) Concatenation

This step introduces concatenation of successive segments to the previous model. For all the pairs of successive two segments, new segments are generated by concatenating the pairs of segments. The new segments are added to the original lattice as new matching paths. (**Fig.2(D)**)

### (E) Blocking Aligned Segments

In this step, all the segments which are aligned at the same position are bundled into a single sub-lattice, which we call "*blocking*" in this paper. When matching a sample to the lattice, the matching path can cross over the segments in the different sequences of the original samples. (**Fig.2(E)**)

### (F) Clustering within a Block

Continued from **(E)**, similar segments within a block are clustered based on the LBG algorithm. (**Fig.2(F)**)

### (G) Grouping Similar Blocks

In this step, phonetic units of models are extracted by grouping similar blocks over all word models by matching sub-lattices in the models.

## 4. EXPERIMENTAL RESULTS

We have tested 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.

### 4.1. Beam Width

Results of experiments in changing the beam width are as shown in **Fig.3**. The recognition rate is almost saturated at a beam width of 128, and its recognition rate is 84.2%. Accordingly, we use this beam width in the following experiments.

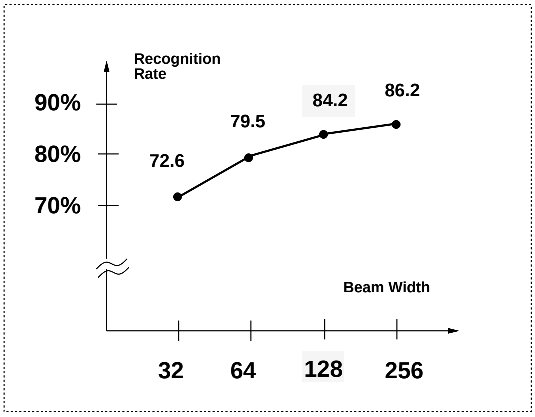

**Figure 3: Beam Width**

### 4.2. Structure of Models

The recognition rate for each step described in the previous section is as follows:

**(A)** 84.2%, **(B)** 80.7%, **(C)** 82.9%, **(D)** 84.4% and **(E)** 91.5%.

These results show that the simple alignment model does not achieve a sufficient recognition rate, but that the rate can be improved by modifying the structures of the lattices. Moreover, blocking of aligned segments significantly improves the recognition rate.

### 4.3. Clustering Segments

Before clustering within a block, we made a preliminary experiment of clustering all of the segments in the model **(A)**. The recognition rates by changing the number of clusters are shown in **Fig.4**. The recognition rate is almost saturated at a cluster size of 128.

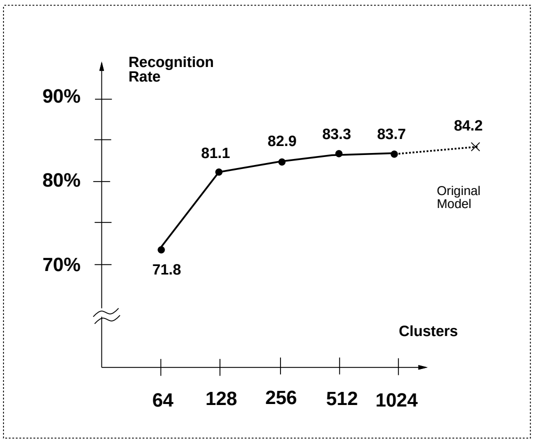

**Figure 4: Clustering Segments in (A)**

We also made a similar experiment on the model **(B)**. The results are shown in **Fig.5**. At the cluster number of 512 and 1024, the recognition rates are improved from the original model.

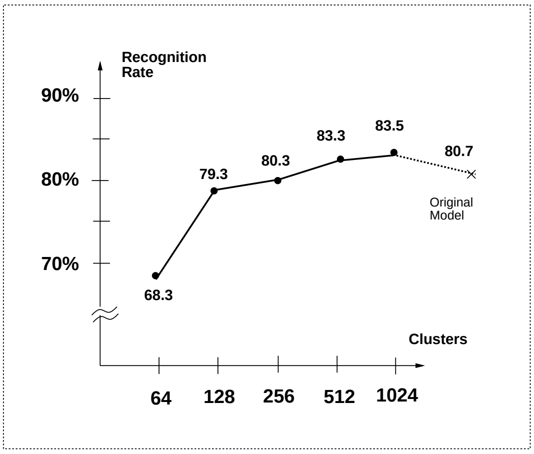

**Figure 5: Clustering Segments in (B)**

A similar experiment on the model **(E)** at the cluster number of 128 results in 90.2% recognition.

### 4.4. Clustering within a Block

An experiment of clustering within a block generates the models as shown in **Fig.2(F)**. The recognition rates by changing the number of clusters in each block are shown in **Fig.6**.

The results show that the recognition performance is improved by clustering and that the fewer the cluster the better the recognition rate. These results suggest that these models have room for improvement by integrating division and concatenation and making more precise models.

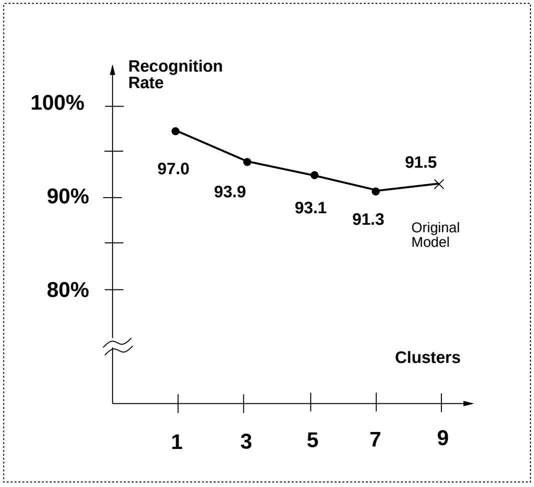

**Figure 6: Clustering within a Block**

## 5. CONCLUDING REMARKS

In order to organize robust phone models, we proposed PLSL and reported on the experiments of modifying the models.

The experimental results are summarized as follows: 1) the recognition rate is improved from the initial simple alignment model by applying division and concatenation to segments in the models. 2) The rate is significantly improved by blocking aligned segments. 3) Clustering within a block also improves the performance.

Investigating the optimum integration of these operations, and extending the matching procedure over different word models are left for future work. A human infant seems to be able to acquire robust speech recognition model. Our task is motivated by this fact. The future goal is to form a phonological system automatically from speech samples by simulating a human infant learning a spoken language.

## 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 and S. Hayamizu, *"Generating Phoneme Models for Froming Phonological Concepts"*, ICSLP 94, pp.1699-1702 (1994).

[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] M. Y. Hwang and X. Huang, *"Subphonetic Modeling with Markov States - Senone"*, Proc. ICASSP-92, Vol.I, pp.33-36 (1992).

[6] J. Takami and S. Sagayama, *"A Successive State Splitting Algorithm for Efficient Allophone Modeling"*, Proc. ICASSP-92, Vol.I, pp.573-576 (1992).

[7] 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).