

GENERATING PHONEME MODELS FOR FORMING PHONOLOGICAL CONCEPTS

Hiroaki Kojima *Kazuyo Tanaka* *Satoru Hayamizu*

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

ABSTRACT

The aim of this work is to improve speech recognition performance by forming "phonological concepts". In order to improve speech recognition performance, the phoneme models or phone models of the system need to satisfy following two properties: 1) preciseness and 2) robustness. These two properties usually trade off against each other in traditional stochastic models. In order to satisfy the both simultaneously, we propose to simulate situations of a human infant learning a language. We call this "phonological concept formation", which is a task of acquiring knowledge of phonological system from spoken word samples without using any transcriptions except for the identification of each word in a lexicon. This knowledge includes what is the set of the whole phonemes, how the acoustic phonetic features of each phoneme are described, and how they are appropriately discriminated. The basis of this idea is specifying essential situations of speech communication instead of providing all encompassing universal samples of a spoken language. Based on this framework, this paper discusses about the method for generating both precise and robust models.

I. INTRODUCTION

The aim of this work is to improve speech recognition performance through the formation of "phonological concepts".

In order to improve speech recognition performance, the phoneme models or phone models of the system must satisfy the following two properties:

- preciseness
- robustness

These two usually trade off against each other in traditional stochastic models. Though an increase in the preciseness of models contributes to a better recognition rate when is estimated using the same closed set of data as used in a learning process, it sometimes causes a decrease in the recognition rate for a test data set which

is different from the training set.

In this case several kinds of criteria (ex. AIC, MDL) are sometimes used. They are effective to avoid over-learning by estimating the optimum precision. They are rather compromise between preciseness and robustness than taking into account substantial robustness.

In order to satisfy both the need for preciseness and robustness simultaneously, models need to represent the essential and common characteristics of the object modeled. In the case of learning a spoken language, phonological models of a learning system need to represent the structures of the phonological system of the language as well as its phonetic features. Models which are generated through a stochastic method inevitably depend on the training set used. One solution is to provide ideal data which reflect the essential and common characteristics of speech communication. However, there is no method for finding this kind of universal data. Another solution is to use existing acoustic, phonetic and phonological knowledge to design a system. However, it is not realistic to expect that this knowledge can be represented appropriately.

We think the key to this task lies in the fact that a human infant seems to be able to acquire these knowledge and concepts efficiently. Motivated by this, we created a framework which we call *phonological concept formation*[5]. This is the task which simulates the situations in which a human infant forms phonological concepts. The basis of this idea is that it is easier to specify essential situations of speech communication than to provide all encompassing universal samples of a spoken language. Specification of the situation is described in the next section.

II. PHONOLOGICAL CONCEPT FORMATION

What is the essence of the necessary situation of speech communication, especially for a human infant to acquire phonological concepts? We summarized this situation as one in which it is possible to discover correspondences between utterances and their meanings

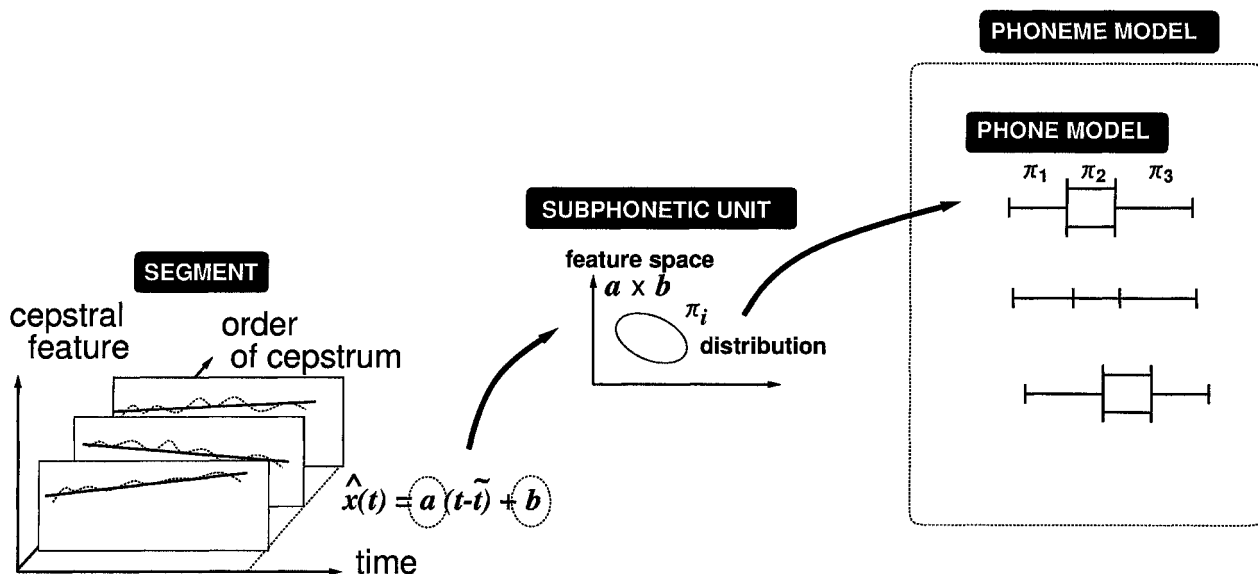


Figure 1. Structure of models

throughout an interaction with other people.

In order to make this situation practical for experimental purposes, we simplified it as follows. We dealt with isolated spoken word samples instead of complete utterances, and we distinguished between lexical differences between words instead of their meanings. This framework is limited to perception of speech even though production is also important in such interactive situations[1].

The task of phonological concept formation can be described as follows. Let $W : \{w_1, w_2, \dots\}$ be the spoken word samples used in a learning process, and let N_{all} be the number of elements in W . The lexicon $C : \{c_1, c_2, \dots\}$ which covers all the elements of W , and the projection $M : \{(w_1, c_i), (w_2, c_j), \dots\}$, which shows, for example, w_1 belongs to c_i , are both given.

In this case, the aim of the learning process is to find efficient phoneme models (i.e., phonological concepts), ρ , such that the recognition rate is maximized. The recognition rate can be defined as $R = N_{right}/N_{all}$, where N_{right} is the number of the sample elements w_i such that $(w_i, \rho(w_i)) \in M$ for each element of W where ρ is a certain function denoting that w_i belongs to the word $\rho(w_i)$.

III. GENERATING PRECISE MODELS

3.1. Representation of Phoneme Models

With our framework, models must include the following factors in order to be precise:

- (a) Sufficient degree of freedom for representing variations

- (b) Flexibility for generating arbitrary precise models with small computational costs

From the above, we defined a phoneme model as shown in Figure 1. A phoneme model consists of a set of phone models. Phone models are represented as networks of subphonetic units. Each phone model has its own individual rules which include information about context dependency. According to the rules, one phone model is chosen from the elements in a certain phoneme set.

This model includes a sufficient degree of freedom to take into account the following: sets of phonemes, sets of phones, context condition of each phone, network structure of each phone model, and representation of each subphonetic units. In addition, this model ensures flexibility by adopting network structures.

3.2. Representation of Segments

Each subphonetic unit in this model is represented by a reference pattern which consists of distribution parameters (i.e. mean values and variances) of feature vectors within a speech segment. The features we used are linear regression coefficients which represent the dynamic pattern of each component of the cepstrum vectors within a segment.

When a sequence of feature vectors in a segment is $s : \{\mathbf{x}(t) \mid t = 1, 2, \dots, T_s\}$, its feature is obtained as λ_s which is the concatenation of \mathbf{a}_s and \mathbf{b}_s such that $\mathbf{x}(t)$ can be approximated by the equation of $\hat{\mathbf{x}}(t) = \mathbf{a}_s(t - \bar{t}) + \mathbf{b}_s$ (where \bar{t} is the average of t). The distortion in s ,

d_s , is the following equation:

$$d_s = \frac{1}{nT_s} \sum_{t=1}^{T_s} (\mathbf{x}(t) - \hat{\mathbf{x}}(t))' \cdot (\mathbf{x}(t) - \hat{\mathbf{x}}(t))$$

(where n is the order of the vector \mathbf{x} , and $'$ indicates the transpose operator).

3.3. Segmentation Algorithm

In general, the shorter the segments used for the unit of modeling, the more precise the model becomes. In order to obtain sufficient precision for the models, the system generates a number of hypotheses of segmentations. A linear regression model provides an efficient segmentation procedure using dynamic programming as is described below.

Let $d(i, j)$ be the total distortion from the i th frame to the j th frame in a sample, which is normalized by the frame size ($j - i + 1$). Segmentation is performed with the following recurrent formulas where $g(n, j)$ is the optimum n division within from the first to the j th frame.

$$g(1, j) = d(1, j)$$

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

If the optimum N division is calculated as $g(N, J)$ where J is the total frame size, the optimum $N+1$ division can be derived with as much as $O(J)$ calculations by storing the value of $g(N, i)$ and $d(i, J)$ for ($1 \leq i \leq J-1$) in advance. This is because the linear regression models are suited for generating precise and flexible models.

IV. GENERATING ROBUST MODELS

4.1. Procedure

Figure 2 shows a flow diagram of the procedure for searching for the appropriate hypotheses for robust phoneme models.

In the first step of this procedure, sequences of cepstrum feature vectors from within the spoken word samples are segmented and transformed to linear regression coefficients. Then, they are projected to a proper metric space, and vector quantized to subphonetic symbols. The sequences of subphonetic symbols generated from the samples are categorized according to the class of their words within the lexicon. By comparing the sequences with each other, network structures of phone models are constructed. Phone models are mapped to phoneme models according to context dependency rules. Finally, a phonemic transcription for each sample is obtained.

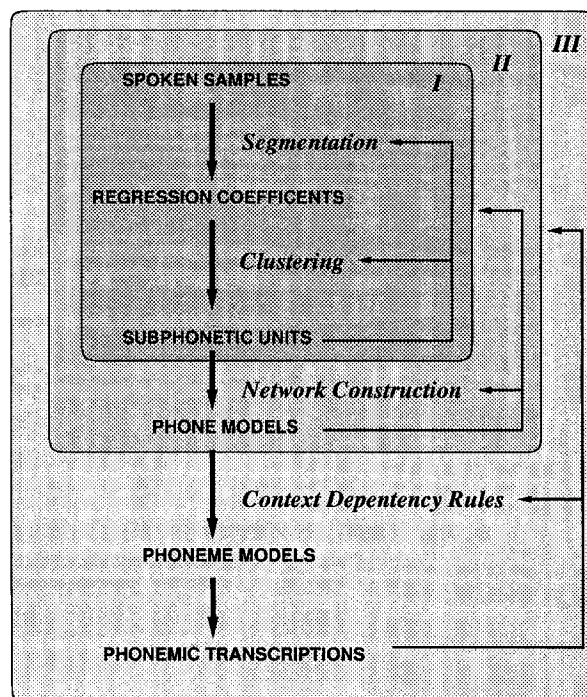


Figure 2. Flow diagram for generating models

In order to perform these processes effectively, a number of possibilities have to be considered to adopt the appropriate hypothesis in every step. These hypotheses consist of segmentation, clustering method, network structures, and context dependency rules. These hypotheses represent the structures of the phoneme models, namely what we call "phonological concept".

The main goal of this procedure is to acquire phoneme models for which all samples in the lexicon can be appropriately transcribed. This means that samples which belong to the same word have the same transcription, and that different words have different transcriptions. However, the learning procedure supervised only by this goal is not practical in terms of computational cost, because it causes an explosion of hypotheses space. Therefore, we divide this procedure into three hierarchical components. By Localizing feedback loops within each component, the total hypotheses space is reduced.

4.2. Generating Subphonetic Units

The subgoal of Level I (in Figure 2) is generating a plausible set of subphonetic units. A spoken word sample is segmented to the sequence of the vector of regression coefficients. Subphonetic units are generated by clustering these vectors. The variable parameters to be determined in this level are as follows:

(I-a) threshold of total distortion within a sample

- (I-b) number of segments in a sample
- (I-c) threshold of distortion within a segment
- (I-d) metric space for clustering
- (I-e) threshold of distortion within a cluster
- (I-f) number of clusters (i.e. subphonetic units)
- (I-g) estimation of similarity between the sequences of subphonetic units which are belong to the same word.

(I-a) and (I-b) define preciseness of segmentation. (I-c) effects to eliminating non stable segments like glides. (I-d) also eliminates peculiar segments from the class of subphonetic units. (I-f) defines preciseness of subphonetic units. The learning process in this level is supervised by (I-g). Within this level, all the subphonetic units are not necessary to be discriminated each other. Feedback information from other levels indicates which pair of units must be discriminated, and modify (I-d).

4.3. Generating Phone Models

The subgoal of Level II is generating adequate phone models. The sequences of subphonetic units are divided to phone level segments. The segments decided to be in the same phone are merged. Then, network structures of phone models are constructed. The variable parameters in this level are as follows:

- (II-a) phone level segmentation of subphonetic sequences
- (II-b) criterion to merge the sequences of subphonetic units into a network structure.
- (II-c) number of phones
- (II-d) estimation of similarity between the sequences of phones which are belong to the same word.
- (II-e) discrimination between the sequences of phones which are belong to different words.

(II-a), (II-b) and (II-c) define preciseness of phone models. In this level, the learning process is supervised (II-d) and (II-e), which require all the individual words to be differently transcribed with phonetic units.

4.4. Generating Phoneme Models

The goal of Level III, which is the main goal, is generating appropriate phoneme models. The phone models are classified to phoneme models according to the context dependency rules formed at the same time. The variable parameters in this level are as follows:

- (III-a) criterion to merge phones into a phoneme
- (III-b) context dependency rules

- (III-c) number of phonemes
- (III-d) discrimination between the sequences of phonemes which are belong to different words.

In this level, the learning process is supervised (III-d), which requires all the individual words to have different phonemic transcription. In the result of this process, the parameters in this level represent phonological knowledge.

V. CONCLUDING REMARKS

We proposed a method for generating both precise and robust models based on the framework of phonological concept formation. The hypotheses space in which optimum parameters are searched for is reduced by adopting the hierarchical strategy. More efficient heuristics and adequate criteria for determining parameters are left for future work

REFERENCES

- [1] F. Fallside: "On the acquisition of speech by machine ASM", Eurospeech 91, Keynote 2, (24 Sep, 1991).
- [2] O. Fujumura: "Phonology and phonetics - A syllable-based model of articulatory organization", JASJ (E) Vol.13 pp.39-48 (1992)
- [3] M. Y. Hwang and X. Huang: "Subphonetic Modeling with Markov States - Senone", Proc. ICASSP-92, Vol.I, pp.33-36 (1992)
- [4] J. Takami and S. Sagayama: "A Successive State Splitting Algorithm for Efficient Allophone Modeling", Proc. ICASSP-92, Vol.I, pp.573-576 (1992).
- [5] H. Kojima, K. Tanaka and S. Hayamizu: "Formation of Phonological Concept Structures from Spoken Word Samples", ICSLP 92, (1992)
- [6] S. Bird: "Finite-state phonology in HPSG", 15th Int. Conf. on Computational Linguistics (1992)
- [7] A. Kannan and M. Ostendorf: "A Comparison of Trajectory and Mixture Modeling in Segment-based Word Recognition", Proc. ICASSP-93, Vol.II, pp.327-330 (1993).
- [8] L. Deng and D. Sun: "Phonetic Recognition using HMM Representation of Overlapping Articulatory Features for All Classes of English Sounds", Proc. ICASSP-94, Vol.I, pp.45-48 (1994).
- [9] S. Krishnan and P.V.S.Rao: "Segmental Phoneme Recognition Using Piecewise Linear Regression", Proc. ICASSP-94, Vol.I, pp.49-52 (1994).