SPEAKER-INDEPENDENT PHONE MODELING BASED ON SPEAKER-DEPENDENT HMMS' COMPOSITION AND CLUSTERING

Tetsuo KOSAKA Shoichi MATSUNAGA and Mikio KURAOKA †

ATR Interpreting Telecommunications Research Labs.,
2-2, Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-02 Japan
†Toyohashi Univ. of Technology, Tempaku-cho, Toyohashi-shi, 441 Japan

ABSTRACT

This paper proposes a novel method for speaker-independent phone modeling based on the Composition and CLustering method (CCL) of speaker-dependent HMMs. In general, HMM phone models are trained by the Baum-Welch (B-W) algorithm. We, however, propose a speaker-independent phone modeling in which speaker-dependent (SD) HMMs are combined to form speaker-independent (SI) HMMs without parameter re-estimation. Furthermore, by using this method, we investigate how different kinds of reference speakers influence the development of the SI models. The method is evaluated in Japanese phoneme and phrase recognition experiments. Results show that the performance of this method is similar to the conventional B-W algorithm's with great reduction of computational cost.

1. Introduction

Generally speaking, HMM phone models are trained with the Baum-Welch (B-W) algorithm [1]. In this algorithm, unfortunately, it is necessary to retrain models when new reference speakers or new data are added to the training data. This becomes a serious problem when speaker-independent phone models have to be modified because an extremely high computational cost is required when the number of training speakers becomes large.

To solve this problem, we propose a speaker-independent phone modeling in which speaker-dependent (SD) HMMs are combined to form speaker-independent (SI) HMMs without parameter re-estimation. By doing so, the computational cost to create SI HMMs becomes much less than that of the conventional approach because the computational cost to train each SD model is small.

In this method, N SD models are clustered and merged to form K-mixture models ($N \geq K$) by using a composition and clustering method (CCL). Since this method allows us to add or remove SD models easily, we can create SI models, gender-specific models (e.g. [2]) or speaker cluster models [3] quickly if the SD models are given.

This paper is organized as follows. First, an SI phone modeling procedure is described. Then, a comparison is made between the conventional B-W algorithm and the proposed method in Japanese phoneme recognition experiments. Finally, by using this method, we investigate how different kinds of reference speakers influence the development of the SI models.

2. Algorithms

2.1. Composition and clustering method (CCL)

The algorithm for the SI phone modeling proposed here is based on the composition and clustering of SD HMMs. In this paper, HMnets (Hidden Markov networks) [4] which efficiently represent phoneme context-dependent HMMs are used for the phone models instead of conventional HMMs. The details of the HMnets are given in subsection 2.2.

An outline of the modeling procedure is as follows:

STEP 1 To allow a single speaker to determine the structure of an HMnet from a large number of word utterances, use the previously proposed SSS algorithm[4]. The derived structure is assumed to be common to all speakers.

STEP 2 Create N speaker-dependent single-Gaussian HMnets using training samples uttered by N reference speakers.

STEP 3 Cluster all of the speaker-dependent HMnets by the method described in subsection 2.3; K clusters are obtained.

STEP 4 At every cluster, combine HMnets to form a single-Gaussian HMnet as follows:

$$\mu_j = \sum_i w_j^{(i)} \mu_j^{(i)} \qquad (1)$$

and

0-7803-2431-5/95 $4.00

$$\hat{S}_j = \sum_i w_j^{(i)} S_j^{(i)} + \sum_i w_j^{(i)} (\mu_j^{(i)} - \hat{\mu}_j)^2, \quad (2)$$

where $\mu_j^{(i)}$ and $S_j^{(i)}$ are the mean and variance of the output pdf., respectively, at state j in the i-th HMnet, and $\hat{\mu}_j$ and $\hat{S}_j$ are the mean and variance of the composed HMnet, respectively.

The weights $w_j^{(i)}$ are given by

$$w_j^{(i)} = n_j^{(i)} / \sum_i n_j^{(i)}, \quad (3)$$

where $n_j^{(i)}$ is the number of training samples for state j in the i-th HMnet.

STEP 5 Create a K-mixture HMnet from K HMnets (from STEP 4) by the speaker-mixture method [5] described in subsection 2.4.

2.2. Hidden Markov networks (HMnets)

The proposed SI phone modeling method has been successfully applied to produce an HMnet considered to be a highly generalized form of the HMM. This HMnet incorporates context-dependent variations of phones and state sharing among different allophones. It contains a finite number of states, each containing Gaussian distributions, that are connected to each other to form paths representing context-dependent phone models. This network is automatically derived by using the Successive State Splitting (SSS) algorithm, which simultaneously solves three problems: network topology, allophone clusters, and the acoustic distribution for each state.

2.3. Clustering algorithm for HMnet

In the clustering process of STEP 3, all of the distances between every two models are calculated in advance and a distance table is created. Bhattacharyya distance measures are used in order to deal with the stochastic models. The cluster with the maximum sum of distances is divided step-by-step using the distance table, similar to the clustering method proposed in SPLIT word recognition [6], which is one of the modification methods of the LBG algorithm [7]. In this method, only the number of clusters is required.

The distance measures between HMnets are described below. We start by introducing the following set of notations:

$b_{jk}^{(i)}$: the k-th state observation probability at state j in the i-th HMnet.
N: the total number of states.
L: the total number of output probabilities.

The distance between two HMnets M_1 and M_2 is defined by

$$D(M_1, M_2) \triangleq \frac{1}{LN} \sum_{j=1}^{N} \sum_{k=1}^{L} d(b_{jk}^{(1)}, b_{g(j)k}^{(2)}), \quad (4)$$

where $g(j)$ is the state permutation that minimizes the value of this equation. In this case, the output probability distribution is used instead of the output string probability. Provided the two HMnets have the same structure, we can assume that

$$g(j) = j. \quad (5)$$

Generally, the values of $d(b^{(1)}, b^{(2)})$ are given by using stochastic measures such as Kullback information measures [8], Chernoff distance measures or Bhattacharyya distance measures. In this work, Bhattacharyya distance measures are used. When the output probability $b^{(i)}$ is assumed to be given by Gaussian pdf. $N(\mu_l, S_l)$, the Bhattacharyya distance between two Gaussian pdfs. $b^{(1)}$ and $b^{(2)}$ is

$$d(b^{(1)}, b^{(2)}) = \frac{1}{8}(\mu_1 - \mu_2)^t \left(\frac{S_1 + S_2}{2}\right)^{-1} (\mu_1 - \mu_2) + \frac{1}{2} \ln \frac{|(S_1 + S_2)/2|}{|S_1|^{1/2} |S_2|^{1/2}}. \quad (6)$$

2.4. Speaker-mixture method

We have already proposed a speaker-mixture method [5] that can yield highly accurate speaker-independent phone models. This method is used in STEP 5 of the SI phone modeling. Suppose that a set of speaker-dependent continuous-mixture HMnets is given for each of several reference speakers. Then, a speaker-mixture phone model is constructed by merging all of the corresponding states of each speaker with equal speaker weights. This model is a kind of hierarchical mixture model that contains speaker mixture weights and intra-speaker mixture weights for mixture components.

3. Recognition Experiments

3.1. Experimental conditions

The SI modeling method was tested on Japanese phoneme and phrase recognition using phrase by phrase utterances. The experimental conditions are summarized in Table 1.

A single Gaussian 200-state HMnet was trained with 50 Japanese sentences uttered by each of 285 speakers. Not only the Baum-Welch algorithm but also Vector Field Smoothing (VFS) [9] was used for the training. Since the VFS algorithm has a smoothing procedure, we could decrease the amount of training data.

Phrase recognition experiments were carried out by using a generalized LR parser, which could cope with

the context-free grammar. The task included 1,035 words and its phoneme perplexity was 5.9.

Table 1: Experimental Conditions

Analysis conditions	
Sampling rate	12 kHz
Window	Hamming window (20 ms)
Frame period	5 ms
Analysis	log power + 16-order LPC-Cep + Δlog power + 16-order ΔLPC-Cep
Training data	
Speakers	139 female + 146 male speakers
Samples	50 Japanese sentences
Recognition data	
Speakers	5 males + 5 females
Samples	279 Japanese phrases

3.2. Investigation on speaker variety

By using the proposed method, we investigated the influence of the number and kinds of reference speakers on the development of the SI models. In conventional SI modeling, while the number of reference speakers has been investigated, the issue of speaker variety has not been considered. We tried a method to create SI models from several reference speakers who were selected from among many reference speakers by a speaker cluster method. This allowed us to decrease the number of reference speakers as well as the computational cost for training.

In the experiment, the following two methods were compared: (1) By using the speaker clustering method described in subsection 2.3, 285 models for each reference speaker were clustered to make N classes. The centroid speaker's models for each class were selected. In other words, N speaker models were selected from among 285 reference speakers. (2) N speakers were selected from among 285 reference speakers randomly. The selected speakers created SI HMnets by using CCL. In method (2), five trials were made and the final recognition rate was calculated by averaging the five results.

The results of phoneme recognition experiments are shown in Figure 1. The results show that the HMnet created with selected speakers based on the clustering method outperforms the HMnet created with randomly selected speakers. This suggests that the number of reference speakers for parameter estimation can be reduced when those speakers are chosen by the clustering method.

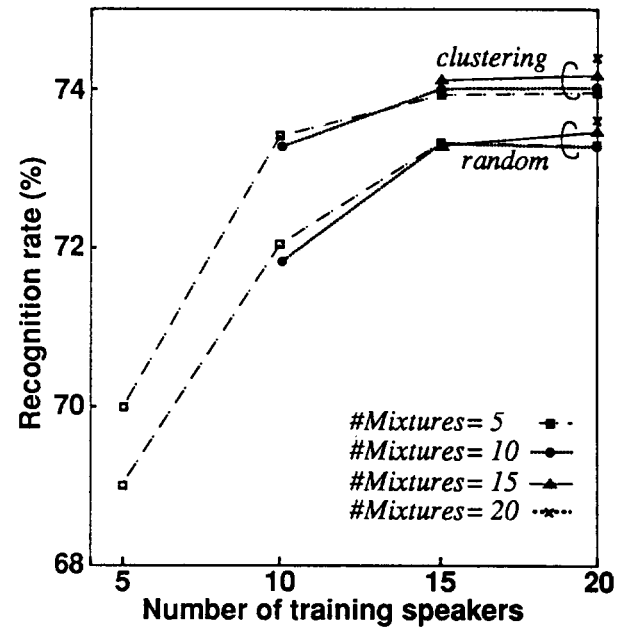

Figure 1: Recognition rate versus the number of training speakers for a comparison between randomly selected speakers and speakers selected by a clustering method in phoneme recognition experiments.

3.3. Recognition results by using composed models

The SI models formed by the SD models were tested on Japanese phoneme recognition.

Every SD HMnet was trained with 50 Japanese sentences uttered by each of 15 speakers. These speakers were selected by the method described in subsection 3.2. Fifteen SD HMnets formed 5, 10 and 15 mixture SI HMnets by the proposed method (Experiment 2). By using these formed HMnets as initial models, additional parameter training by the conventional B-W algorithm (Experiment 3) or VFS (Experiment 4) was carried out for comparison purposes. In addition, parameter re-estimation by the conventional B-W algorithm only was tested (Experiment 1). In this case, the initial values of output pdf. were given by a VQ method. The flowchart for these experiments is shown in Figure 2.

As shown in Table 2, the results indicate that the proposed method achieves better performance than that of the conventional B-W algorithm only. Note that the performance is actually better when the B-W algorithm or VFS is added to the composed model, than when the composed model itself is evaluated. But the difference in performance among B-W retraining of the composed model, VFS retraining of the composed model and the composed model itself is small.

In Table 2, parentheses represents the computation time in hours. Since the computation time for creating the structure of an HMnet is common to all experiments, it was eliminated. The calculation cost of the composed model is much smaller than that of the B-W algorithm. In fact, compared with the B-W algorithm, the CCL reduced the calculation cost to between approximately 1/20 and 1/50.

Table 3 shows the results of phrase recognition by using the composed HMnet. In this experiment, 5 and 10 mixture HMnets were created with 285 reference speakers by the proposed method, and they achieved good performance.

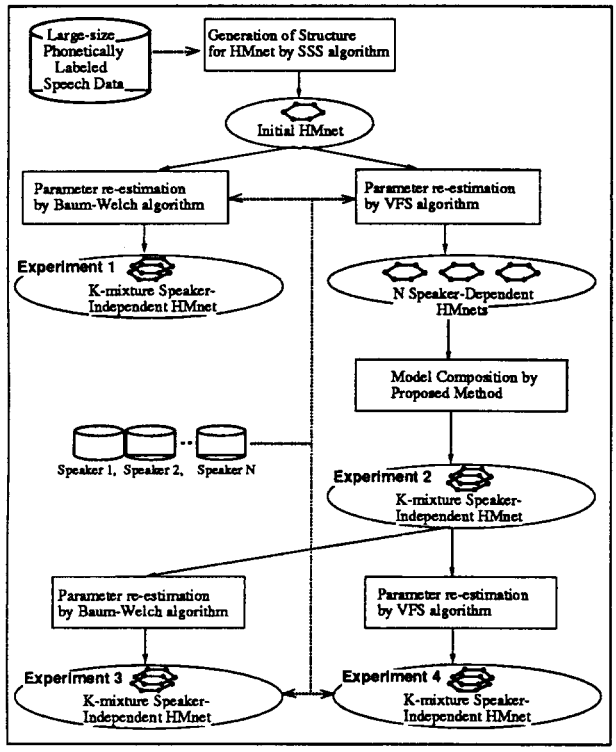

Figure 2: Flowchart of recognition experiments

4. Conclusion

We have proposed a novel method for speaker-independent phone modeling based on the composition and clustering of speaker-dependent HMMs. Since this method allows us to add or remove SD models easily, we can create SI models, gender-specific models or speaker cluster models very quickly. It has been evaluated in Japanese phoneme and phrase recognition experiments. The results show that the performance of this method is equivalent to the conventional B-W algorithm's in spite of a drastically reduced computational cost. In fact, compared with the B-W algorithm, this method reduced the calculation cost to between approximately 1/20 and 1/50.

We plan to expand the number of speakers to several hundred speakers for the SI modeling, and to apply this modeling approach to conventional continuous and semi-continuous HMMs.

Table 2: Phoneme recognition rate by using CCL itself, with Baum-Welch or VFS (%). ( ) represents computation time (hours).

method / #mixtures	5	10	15
1. Baum-Welch	64.6 (102.0)	67.4 (188.5)	70.2 (276.1)
2. CCL	73.6 (4.4)	74.0 (4.4)	74.1 (4.4)
3. CCL + Baum-Welch	77.2 (100.7)	77.8 (179.1)	78.1 (259.6)
4. CCL + VFS	74.5 (34.3)	76.0 (62.5)	76.1 (88.9)

Table 3: Phrase recognition rate by using CCL (%)

candidates/#mixtures	5	10	15
top 1	76.6	77.4	77.6
top 5	94.9	95.2	95.1

5. References

[1] L. E. Baum and J. A. Eagon: "An Inequality with Applications to Statistical Prediction for Functions of Markov Processes and to a Model fr Ecology," Bull. Am. Math. Soc., 73, (1967).

[2] L. F. Lamel and J. L. Gauvain: "High Performance Speaker-Independent Phone Recognition Using CDHMM," Proc. of Eurospeech93, pp. 121-124, (1993).

[3] T. Kosaka and S. Sagayama: "Tree-Structured Speaker Clustering for Fast Speaker Adaptation," Proc. of ICASSP'94, pp. 245-248 (1994).

[4] J. Takami and S. Sagayama: "A Successive State Splitting Algorithm for Efficient Allophone Modeling," Proc. of ICASSP92, (1992).

[5] T. Kosaka, J. Takami and S. Sagayama: "Rapid Speaker Adaptation Using Speaker-Mixture Allophone Models Applied to Speaker-Independent Speech Recognition," Proc. of ICASSP93, pp. 570-573, (1993).

[6] N. Sugamura, K. Shikano and S. Furui: "Isolated Word Recognition Using Phoneme-Like Templates," Proc. of ICASSP83, (1983).

[7] Y. Linde, A. Buzo and R.M. Gray: "An Algorithm for Vector Quantizer Design," IEEE Trans. Commun., COM-28, 1, pp. 84-95 (1980.01).

[8] B.-H. Juang and L. R. Rabinaer: "A Probabilistic Distance Measure for Hidden Markov Models," AT&T Technical Journal Vol. 64, No. 2, (1985.02).

[9] K. Ohkura, M. Sugiyama and S. Sagayama: "Speaker Adaptation Based on Transfer Vector Field Smoothing with Continuous Mixture Density HMMs," Proc. of ICSLP92, We.fPM.1.1, pp. 369-372 (1992).