High-Speed Speaker Adaptation Using Phoneme Dependent Tree-Structured Speaker Clustering

M. SUZUKI[1], *T. ABE*[2], *H. MORI*[2], *S. MAKINO*[1], *H. ASO*[2]

[1] Computer Center / Graduate school of Information Sciences, TOHOKU Univ.
[2] Graduate school of Engineering, TOHOKU Univ.

ABSTRACT

The tree-structured speaker clustering was proposed as a high-speed speaker adaptation method. It can select the model which is most similar to a target speaker. However, this method does not consider speaker difference dependent on phoneme class. In this paper, we propose a speaker adaptation method based on speaker clustering by taking speaker difference dependent on phoneme class into account. The experimental results showed that the new method gave a better performance than the original method. Furthermore, we propose the improved method which use a tree-structure of a similar phoneme as the substitute for the phoneme which does not appear in the adaptation data. From the experimental results, the improved method gave a better performance than the method previously proposed.

1. INTRODUCTION

One of the most efficient speaker adaptation method is the tree-structured speaker clustering algorithm proposed by Kosaka et. al[1]. In this method, Hidden Markov Network (HMnet) is constructed for a typical speaker. Then, the HMnet is adapted to each training speaker using a small size of training data. A tree-structure consisting of HMnets corresponding with each training speaker is constructed using a clustering method based on similarity between two HMnets. The root node represents a speaker-independent model constructed with HMnets of all speakers, and each leaf node represents a speaker-dependent HMnet for each speaker. When speech data for adaptation is given, the node with the maximum likelihood for the data is picked up. The recognition is carried out using the HMnet of the node.

This method has the following advantages:

1. Various models from a speaker-independent model to a speaker-dependent one are available. If an input speaker is similar to one of training speakers, the model close to a leaf node is chosen, otherwise, the model close to the root node is chosen.
2. Adaptation speed is very high because this method is based on speaker selection.

This method assumes that the same amount of speaker difference is appeared in all phonemes. However, amount of speaker difference is different dependent on kind of phoneme. The method proposed by Kosaka et. al. does not consider the above-mentioned facts.

To solve this problem, we propose a new high-speed speaker adaptation method using phoneme-dependent tree-structured speaker clustering.

2. PHONEME DEPENDENT TREE-STRUCTURED SPEAKER CLUSTERING

We propose a high-speed speaker adaptation method using phoneme-dependent tree-structured speaker clustering. This algorithm has two steps: construction step and adaptation step. Details of these two steps are described in the following subsections.

2.1. Construction algorithm for tree-structured speaker clustering

The algorithm is as follows:

1. Train a speaker-dependent HMnet using SSS-free algorithm[2]. SSS-free is one of the construction algorithm of HMnet, and it needs a large amount of training data.
2. Build another speaker-dependent HMnets from the speaker-dependent HMnet using Vector Field Smoothing (VFS) algorithm[3]. VFS is one of the speaker adaptation algorithm, and it can adapt the HMnet to a new speaker with a small size of adaptation data.
3. Split every speaker-dependent HMnets to sub-HMnets corresponding to each phoneme.
4. For all phonemes, construct tree-structure from all sub-HMnet using tree-structured speaker clustering algorithm[1].
   a) Assign all speaker-dependent sub-HMnets to one cluster.
   b) Choose a sub-HMnet pair with the maximum distance each other from all sub-HMnets assigned to a cluster having more than one sub-HMnets. Distance between sub-HMnets is defined as a sum of Bhattacharyya distance between corresponding states.

c) Split the cluster into two new clusters. Cluster center is set to the sub-HMnet chosen at the step b), and other sub-HMnets are assigned to the cluster with the nearest distance.

d) Go to the step b) until the number of sub-HMnet assigned to each cluster becomes only one.

After the construction of tree-structure, each representative sub-HMnet is computed from all sub-HMnets assigned to each cluster. Output probability distribution ($b^{(i)}(x)$) of a state of the representative sub-HMnet is set to the weighted sum of states in each speaker-dependent sub-HMnets as follows:

$$b^{(i)}(x) = \sum_{s} \frac{n_s^{p(i)}}{\sum_{s'} n_{s'}^{p(i)}} b_s^{(i)}(x)$$

where, i indicates a state, s indicates a speaker, $n_s^{p(i)}$ indicates number of training samples at the state i of a phoneme p.

Each representative sub-HMnet is assigned to each node in a tree-structure. The sub-HMnet assigned to the root node is corresponding to a speaker-independent phoneme HMnet, and the sub-HMnet assigned to a leaf node is corresponding to a speaker-dependent phoneme HMnet.

2.2. Adaptation algorithm

When speech data are given for adaptation, we choose an optimum sub-HMnet for each phoneme independently using the following algorithm.

1. Calculate a likelihood for adaptation data using a sub-HMnet assigned to the root node. Mark on the root node.
2. For all child sub-nodes under the marked node, calculate a likelihood using a sub-HMnet assigned to the sub-node, and choose the sub-node with the maximum likelihood.
3. Mark on the chosen sub-node. Go to the step 2 until there is no sub-node at the marked node.

The sub-HMnet with the maximum likelihood of the all marked node is chosen.

When we cannot choose a sub-HMnet of a phoneme because the phoneme data are not existed in adaptation data, we use substitute for the sub-HMnet of the phoneme. Selection algorithm of the substitute is given as follows:

1. Construct a tree structure from all speaker-dependent HMnets corresponding with all phonemes. This tree-structure is the same as that obtained from the original algorithm[1], and it is called "all-phoneme tree" in this paper.
2. Choose an HMnet with the maximum likelihood using the same algorithm for picking up a sub-HMnet.
3. Split a chosen HMnet corresponding with all phonemes to a sub-HMnet of the phoneme which is not included in adaptation data..

In this paper, this method is called "method 1".

2.3. Phoneme recognition experiment

To confirm effectiveness of our algorithm, we carried out a phoneme recognition experiment. We constructed a speaker-dependent HMnet using 400 sentences uttered by a male speaker. Eight speaker-dependent HMnets were built using 50 sentences uttered by each speakers (four male, four female). We carried out adaptation experiments for four speakers (two male, two female), and one sentence per speaker is used as adaptation data.

Table 1: Phoneme recognition accuracy

	original	method 1
vowel	75.1%	76.2%
consonants	61.5%	61.3%
total	68.7%	69.5%

Table 1 shows phoneme recognition accuracy. Vowel recognition accuracy was improved from that of the original[1], on the other hand, consonants recognition accuracy was similar to that of the original. Tree-structure is much different dependent on kind of vowel, but it is not different dependent on kind of consonants. Amount of speaker difference is different dependent on kind of vowels, however it is not different dependent on kind of consonants.

Totally, phoneme recognition accuracy was increased by 0.8% in comparison with the original tree-structured speaker clustering algorithm.

We investigate the obtained phoneme tree-structures and "all-phoneme tree". Dendrogram of typical phonemes are shown in figure 1 to 4. A diverging point of a branch indicates a distance between speaker clusters. For example, a distance between speaker MTK and MMY is about 180 in figure 1. In these figures, speaker M** and F** indicate a male and a female, respectively, and the cluster in the shadow box is a selected

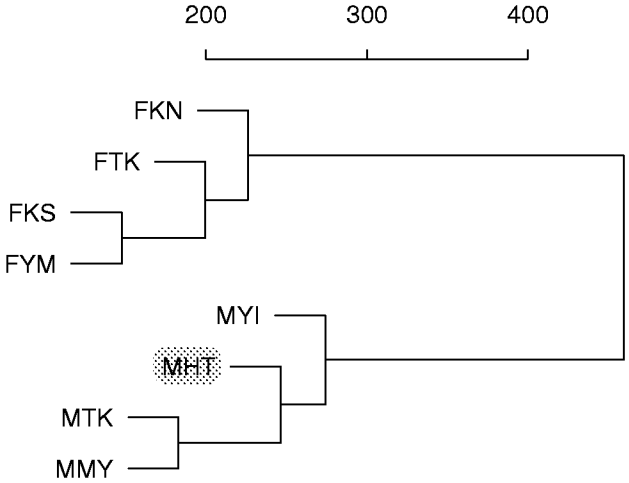

Figure 1: Tree structure corresponding with an HMnet for all phonemes

cluster at the adaptation step. In about half of number of tree-structures for each phonemes (figure 2 and 3) and in the "all-phoneme tree" (figure 1), the first level node (root node) is split into male cluster and female cluster, and lower level node is split to various clusters. It shows that amount of speaker difference is different dependent on kind of phonemes. On the other hand, in tree-structures of some phonemes, for example /s/, /b/, /p/(figure 4) etc., the first level node is not split into male and female clusters. It shows that these phonemes have no speaker difference.

Moreover, in tree-structures of some phonemes which is few appeared in Japanese, the first level node is not split into male and female cluster. One of the reason is why there is few training data at building a speaker-dependent HMnet using VFS algorithm. If there is few training data for VFS, we cannot obtain reliable parameter for HMnet. Tree-structure of the phoneme does not describe speaker difference.

We also investigated a speakers assigned to a selected node of tree-structures at the adaptation step. In tree-structures of many consonants, speakers assigned to a selected node were the same as those assigned to a selected node of the "all-phoneme tree". This is one of the reason why consonants recognition accuracy with our algorithm is similar to that with the original. On the other hand, in tree-structures of vowels, speakers assigned to a selected node were different dependent on kind of vowels. This shows that construction of tree-structure for each phoneme is effective to improve phoneme recognition performance.

3. A NEW SELECTION ALGORITHM FOR SIMILAR SUB-HMNET

In the proposed method previously mentioned, a sub-HMnet assigned to the node of "all-phoneme tree" is used as a substitute for a sub-HMnet of the phoneme when training data for the phoneme is not existed. However, "all-phoneme tree" does not consider differences between speakers. The utilization of a similar phoneme tree-structure is desirable instead of utilization of "all-phoneme tree". Now, we assume that a distance between phonemes with similar tree-structure is close. Then, we propose a construction method of substitute for a sub-HMnet of a phoneme based on a distance between phonemes.

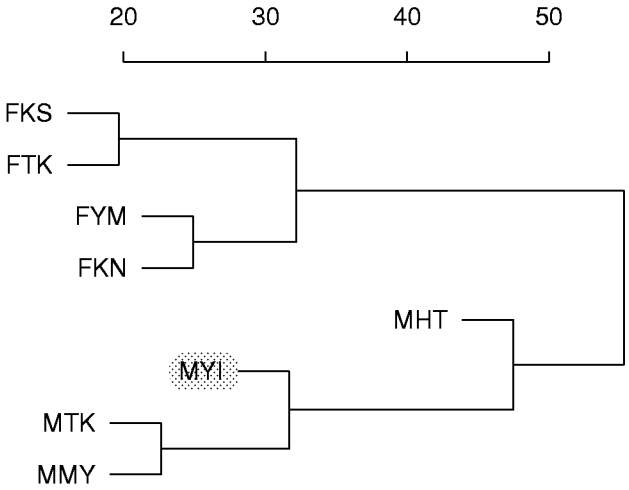

Figure 2: Tree structure corresponding with an HMnet for phoneme /a/

3.1. New construction method of substitution sub-HMnet

The distance between two phonemes is at first calculated at the construction step, and then a substitute for a sub-HMnet of a phoneme is newly constructed at the adaptation step. The distance between two phonemes is calculated from parameters of sub-HMnet as follows:

1. We calculate a distance between two paths in a phoneme sub-HMnet using the dynamic time-warping method, and the path(M) with the minimum distortion($d(M)$) is

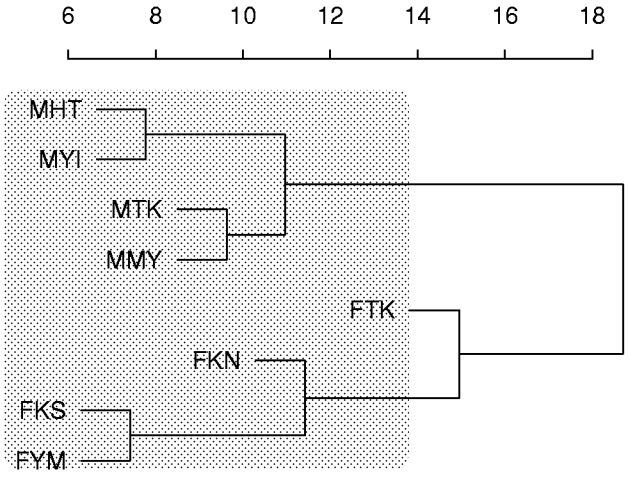

Figure 3: Tree structure corresponding with an HMnet for phoneme /g/

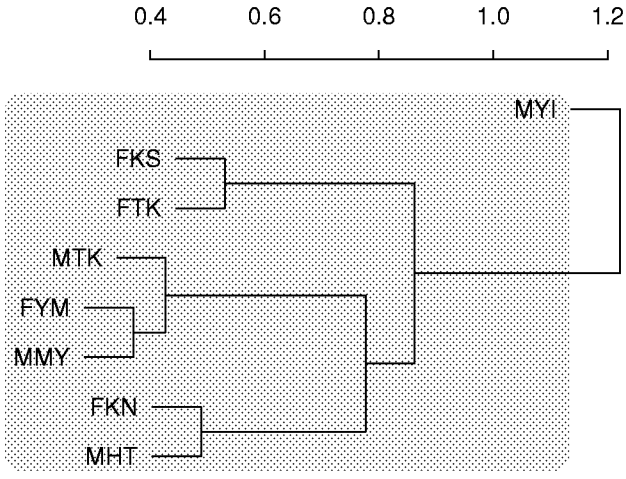

Figure 4: Tree structure corresponding with an HMnet for phoneme /p/

regarded as a typical path of the phoneme sub-HMnet.

$$d(M)=\left(\prod_{i\in P}^{n} m_i \times D(M,M_i)\right)^{\frac{1}{n}}$$

where, m_i indicates a number of training samples used in training of a path M_i, $D(M, N)$ indicates a distance between paths M and N using dynamic time-warping method.

2. We regard the distance between two typical paths as a distance between the two phonemes.

When we cannot choose a sub-HMnet of a phoneme, we used a speaker information of the tree corresponding with the nearest phoneme. Speaker information assigned to the chosen node of the tree is picked up, and new sub-HMnet is constructed from sub-HMnets corresponding with the speaker information. In this paper, this method is called "method 2".

3.2. Phoneme recognition experiment

To confirm effectiveness of our algorithm, we carried out a phoneme recognition experiment. Eight speakers (four male, four female) were used as an adaptation speaker, and one sentence per speaker is used for adaptation data. Other experimental conditions are the same as those used in the previous experiment.

Table 2: Phoneme recognition accuracy

	original	method 1	method 2
total	67.3%	68.8%	69.7%

Table 2 shows phoneme recognition accuracy. Total recognition accuracy of the original and method 1 is different from the previous experiments because the number of adaptation speakers is different from the previous experiment. Method 2 showed the highest performance of all. We can conclude that substitute for a sub-HMnet of a phoneme should be constructed using speaker information of similar phoneme.

In the experiments, all of phonemes not included in adaptation data are consonants. Consonants recognition accuracy was similar to that of the original when using method 1, but it was improved from that of the original when using method 2.

4. CONCLUTION

We propose a new high-speed speaker adaptation algorithm using phoneme-dependent tree-structured speaker clustering. This algorithm can consider the speaker difference for each phoneme independently. From the experimental results, amount of speaker difference is different dependent on kind of vowel. Totally, the new algorithm shows better performance than that of the original.

To improve the performance of the new algorithm, we define the distance between phonemes, and propose a construction algorithm of substitute for a sub-HMnet of a phoneme based on the distance between phonemes. It is effective to improve the performance of phoneme recognition system.

We should do the following works in near future:

1. Reexamine the definition of a distance between two HMnets.
2. Confirm effectiveness of our algorithm when various data uttered by a large number of speakers are given.

5. REFERENCES

1. T. KOSAKA and S. SAGAYAMA. "Tree-Structured Speaker Clustering for Fast Speaker Adaptation" Proc. of ICASSP94, 245-248, 1994.
2. M. SUZUKI, S. MAKINO, A. ITO, H. ASO, and H. SHIMODAIRA. "A New HMnet Construction Algorithm Requiring No Contextual Factors" IEICE *Trans. Inf.&Syst.*, E78-D, No. 6, 662-668, 1 995.
3. K. OHKURA, M. SUGIYAMA, and S. SAGAYAMA. "Speaker Adaptation Based on Transfer Vector Field Smoothing Method with Continuous Mixture Density HMMs" IEICE *Trans. Inf.&Syst.*, J76-D-II, No. 12, 2469-2476, 1993. (in Japanese)