

SPEAKER ADAPTATION BASED ON TRANSFER VECTORS OF MULTIPLE REFERENCE SPEAKERS

Kazumi OHKURA Hiroki OHNISHI Masayuki IIDA

SANYO Electric Co., Ltd.

Hypermedia Research Center,

Speech Processing Laboratory

1-18-13, Hashiridani, Hirakata-City, Osaka 573

ABSTRACT

This paper addresses a speaker adaptation method for speaker-independent phoneme HMMs (SI-HMMs). The type of these HMMs is a Gaussian continuous mixture density HMM. It is important for speaker adaptation to make up for a deficiency of training data. In our proposed method, the *TRAnsfer VEctors* of multiple *Reference SpEakers* estimated from sufficient training data are used to make up for a deficiency of the training data. We call this method *TRAVErSE*. *TRAVErSE* was evaluated by 100 isolated words recognition experiment. The testing speakers were five males. The average recognition rate for the five speakers was 80% with SI-HMMs. Applying *TRAVErSE* to SI-HMMs, the recognition rates increased 86.1% and 90.5% with 1 and 5 isolated words for speaker adaptation.

I. INTRODUCTION

In the past few years many recognition methods which use continuous mixture density HMMs have been studied. Research has also been conducted on speaker adaptation methods using these HMMs [1]~[11]. These speaker adaptation methods agree in a basic thought which makes up for a deficiency of information obtained at the time of adaptation by information obtained beforehand. In our proposed method, transfer vectors are used as the information.

Speaker adaptation in HMMs is regarded as a kind of retraining problem, using a small amount of training data. The problem includes that (1) some phoneme HMMs are not trained and (2) errors in estimating HMM parameters can result from insufficient training data. To solve these problems, therefore, it is important for speaker adaptation to make up for a deficiency of training data. This retraining of models can also be viewed as a mapping from initial HMMs to retrained HMMs. This mapping is carried out by vectors obtained from differences between subspaces in initial HMMs and those in retrained HMMs, i.e. these vectors are equivalent to the mapping functions, we call these vectors *transfer vectors*. In our proposed method, the *TRAnsfer VEctors* of multiple *Reference SpEakers* estimated from sufficient training data are

used to make up for a deficiency of the training data. We call this method *TRAVErSE*.

Section II. gives details of the *TRAVErSE* algorithm. Section III. evaluates the *TRAVErSE* by experiments on recognizing 100 Japanese isolated words.

II. TRAVErSE ALGORITHM

TRAVErSE is carried out by the selection of reference speaker and the mapping of Gaussian mean vectors according to transfer vector. The concept of *TRAVErSE* is shown in Figure 1. The type of these HMMs is a Gaussian continuous mixture density HMM. The details of *TRAVErSE* algorithm are as follows:

1) Calculation of transfer vectors of reference speaker

In this step, transfer vectors of reference speaker are calculated. These transfer vectors are obtained from the difference between the mean vectors of the SI-HMMs and those created after retraining. The details of this step are shown below.

- (1) A set of SI-HMMs λ is used as the set of initial HMMs for the reference speaker.

$$\lambda = \{\lambda_1, \dots, \lambda_i, \dots, \lambda_I\},$$

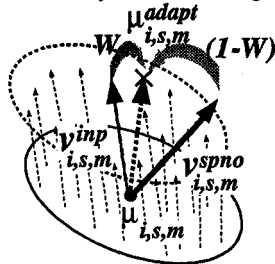
where I is the number of phoneme HMMs, e.g. in this paper we used 39 phoneme HMMs, thus $I = 39$. Each λ_i is a vector of HMM parameters completely characterizing the i 'th phoneme HMM, i.e.

$$\lambda_i = \{w_{i,s,m}, a_{i,s_j,s_k}, \mu_{i,s,m}, \sigma_{i,s,m}\},$$

where $w_{i,s,m}$, $\mu_{i,s,m}$ and $\sigma_{i,s,m}$ are branch probability, vector of Gaussian means and vector of Gaussian variances corresponding to the m 'th mixture in the s 'th state of the i 'th HMM respectively. a_{i,s_j,s_k} is the state transition probability from state s_j to state s_k of i 'th HMM. For example in this paper the order of feature parameters is 33, thus $\mu_{i,s,m}$ and $\sigma_{i,s,m}$ are 33 dimensional vectors.

- (2) $w_{i,s,m}$ and $\mu_{i,s,m}$ on λ_i are re-estimated using the n 'th reference speaker's utterances, by the

Sub-space of input speaker's models
restricted by k -nearest neighbors rule



Sub-space of speaker-independent models
restricted by k -nearest neighbors rule

Figure 3. Adaptation of mean vector.

4) Adaptation of mean vectors and branch probabilities.

$v_{i,s,m}^{spno}$ is modified by $v_{i,s,m}^{inp}$, and then the adapted $\mu_{i,s,m}^{adapt}$ is obtained, i.e.

$$\forall_{i,s,m} \in \Omega \quad \mu_{i,s,m}^{adapt} = W v_{i,s,m}^{inp} + (1.0 - W) v_{i,s,m}^{spno} + \mu_{i,s,m}$$

where W is a parameter representing the reliability of $v_{i,s,m}^{inp}$.

Finally, the branch probability of selected reference speaker $w_{i,s,m}^{spno}$ is used as that of adapted phoneme HMM λ_i^{adapt} , we get λ^{adapt} as the adapted HMM for the input speaker.

$$\lambda^{adapt} = \{ \lambda_1^{adapt}, \dots, \lambda_i^{adapt}, \dots, \lambda_I^{adapt} \},$$

$$\lambda_i^{adapt} = \{ w_{i,s,m}^{spno}, a_{i,s_j,s_k}, \mu_{i,s,m}^{adapt}, \sigma_{i,s,m} \}.$$

The concept of adaptation for $\mu_{i,s,m}^{adapt}$ is shown in Figure 3.

III. EVALUATIONS

III.1 Word recognition experiments

SI-HMMs λ was trained using fluently spoken utterances uttered by 30 males. The training utterances totaled 161 minutes and included Japanese 2554 sentences. 10 reference speaker's HMMs (λ^n ; $n = 1, \dots, 10$) were used as base models for speaker adaptation. These reference speaker's HMMs were trained using utterances of 10 male speakers which were included in the training data of SI-HMMs. For testing data, sets of 100 Japanese city name uttered once by 5 male speakers were used. This data was included JEIDA database (JS-WRD-89-0031). A silent period of 50 ms was added before and after each utterances. For speaker adaptation, 1 ~ 20 utterances in this testing data set were used. For word recognition, testing data set but the utterances for speaker adaptation were also applied. The evaluations were accomplished 2 times by changing words for speaker adaptation. The feature vector was a 33 dimensional vector consisting of 16 cepstral coefficients, 16 Δ cepstral coefficients and Δ logarithmic power. Analysis conditions are listed in Table 1.

Table 1. Analysis conditions.

pre-emphasis	$1 - 0.90z^{-1}$
window length	21.3 ms (Hamming window)
window shift	5 ms
LPC analysis order	16
LPC cepstrum order	16
Δ window length	100 ms

The HMMs themselves were continuous Gaussian-mixture density models, with no explicit duration modeling. The number of mixture components per state is 4. Each of the mixture components had a diagonal covariance matrix. The total number of mixture components in the 39 HMMs is 468. All models were left-to-right, 4-state and 3-loop models, with no skips.

In this experiment, K , f and W were 60, 2.0 and 0.5 respectively, and the function $g(w_{i,s,m}^n)$ was constant value 1.0. In this experiments, the adapted HMM λ^{adapt} obtained by *TRVERSE* was compared the following four models. Figure 4 shows the average recognition rates.

- (1) Speaker-independent model λ .
(Speaker-independent.)
- (2) Input speaker's model λ^{inp} re-trained by concatenation training technique using utterances for speaker adaptation.
(Concatenation training.)
- (3) Reference speaker's model λ^{spno} selected by a speaker selection method shown at II.
(Selected reference speaker.)
- (4) Reference speaker's model showed the highest recognition performance without adaptation.
(Best reference speaker.)

As can be seen in Figure 4, *TRVERSE* showed the highest recognition performance. These results show that *TRVERSE* is effective for speaker adaptation with insufficient training data.

III.2 Dependency of recognition performance and reference speaker's model

To evaluate dependency of recognition performance and reference speaker's model, the following experiments were performed.

Speaker adaptation was carried out using each reference speaker's model (λ^n ; $n = 1, \dots, 10$) and 10 λ^{adapt} were created, then word recognition experiment is accomplished by using these 10 λ^{adapt} . The best and worst recognition rates using these λ^{adapt} are shown to Figure 5. In Figure 5, the recognition rate of reference speaker's model selected by the reference speaker selection method that showed at II. is also shown. In

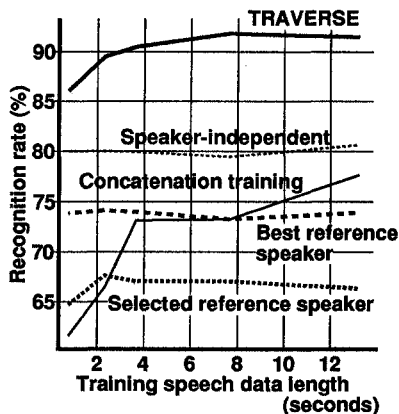


Figure 4. Word recognition results.

this experiment, K , f and W were 60, 2.0 and 0.5, respectively.

It understands from figure 5 that recognition rate changes by reference speaker biggest 10 % level in *TRVERSE* method. And, this reference speaker selection method does not select the best reference speaker's model in spite of selecting good reference speaker's model. If this reference speaker selection method is modified from this experimental result, although recognition rate will be thought to improve furthermore, this point would be continued to investigate.

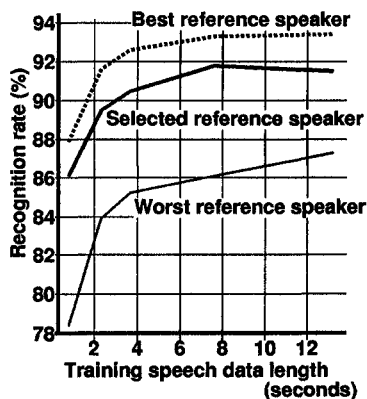


Figure 5. Dependency of recognition performance and reference speaker's model.

IV. CONCLUSIONS

This paper showed the effectiveness of a speaker adaptation method based on the *TRANSfer Vectors* of multiple *Reference SpEakers (TRVERSE)*.

Future plans include:

- (1) Evaluation of the case when increased a number of reference speakers and testing speakers.
- (2) Improvement of reference speaker selection method.

And, we would evaluate the case when made decrease model number to refer to by clustering of reference speaker's model furthermore.

REFERENCES

- [1] Y. Hirata and S. Nakagawa. "A Study of Speaker Adaptation of Continuous Parameter HMM on Japanese Phoneme Recognition," Tech. report of IEICE, SP90-16, (Jun. 1990).
- [2] T. Matsuoka and K. Shikano. "Speaker Adaptation by Modifying Mixture Coefficients of Speaker Independent Mixture Gaussian HMMs," Proc. of Meeting of Acoust. Soc. of Japan, 1-1-6, pp. 11-12 (Mar. 1992).
- [3] T. Iwahashi and K. Nakajima. "Continuous Mixture Densities Based Speaker Adaptation for Acoustic Phonetic Segment HMM," Proc. of Meeting of Acoust. Soc. of Japan, 1-5-21, pp. 45-46 (Mar. 1991).
- [4] C. H. Lee, C. H. Lin and B. H. Juang. "A Study on Speaker Adaptation of the Parameters of Continuous Density Hidden Markov Models," IEEE Trans. on SP, **39**, pp. 806-814 (1991-04).
- [5] T. Koshikawa and S. Nakagawa, "Speaker Adaptation for Syllable-Based HMM using Continuous Speech by Maximum a Posteriori Estimation," Proc. of Meeting of Acoust. Soc. of Japan, 1-1-9, pp. 17-18 (Mar. 1992).
- [6] T. Kosaka, S. Matsunaga and S. Sagayama, "Tree-structured Speaker Clustering for fast Speaker Adaptation," Proc. of Meeting of Acoust. Soc. of Japan, 2-7-14, pp. 97-98 (Oct. 1993).
- [7] T. Matsuoka and C. H. Lee, "A Study of Online Bayesian Adaptation for HMM-based Speech Recognition," Proc. of Meeting of Acoust. Soc. of Japan, 2-7-13, pp. 95-96 (Oct. 1993).
- [8] K. Ohkura, M. Sugiyama and S. Sagayama, "Speaker Adaptation Based on Transfer Vector Field Smoothing with Continuous Mixture Density HMMs," Proc. of ICSLP92, pp. 369-372 (Oct. 1992).
- [9] Y. Nakato and H. Matsumoto. "A Study on Unsupervised Speaker Adaptation of Continuous Parameter HMM," Tech. report of IEICE, SP90-67, pp. 79-86 (Dec. 1990).
- [10] K. Shinoda, K. Iso, and T. Watanabe: "Speaker Adaptation For Demi-Syllable Based Continuous Density HMM", Proc. ICASSP91, S13.7, pp. 857-860 (1991-05).
- [11] Y. Miyazawa and S. Sagayama, "An Examination for a Method of Standard Speaker Selection for Speaker Adaptation Based on Transfer Vector Field Smoothing Model," Proc. of Meeting of Acoust. Soc. of Japan, 2-5-2, pp. 121-122 (Oct. 1992).
- [12] H. Hattori and S. Sagayama. "Vector Field Smoothing Principle for Speaker Adaptation," Proc. of ICSLP92, pp. 381-384 (Oct. 1992).