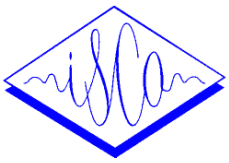


2nd International Conference on
Spoken Language Processing (ICSLP 92)
Banff, Alberta, Canada
October 12-16, 1992

SPEAKER ADAPTATION BASED ON TRANSFER VECTOR FIELD SMOOTHING WITH CONTINUOUS MIXTURE DENSITY HMMs

Kazumi Ohkura, Masahide Sugiyama and Shigeki Sagayama

ATR Interpreting Telephony Research Laboratories
2-2, Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-02, Japan

ABSTRACT

This paper describes a method of speaker adaptation for continuous mixture density HMMs (CDHMMs). Speaker adaptation in CDHMMs is regarded as a kind of retraining problem where a small amount of training data is available. The **"Vector Field Smoothing method (VFS)"** is used to deal with the problem of retraining with insufficient training data. "VFS" is applied simultaneously to inter-speaker and speaking-style adaptation. In this paper, the standard speaker is a male and the unknown speakers for adaptation are both one male and one female. When 11 sentences are uttered for adaptation phrase-by-phrase instead of word-by-word, the 23 phoneme recognition rate is 87.4% (none adaptation: 47.3%). The phrase recognition rate for HMM-LR is 85.1% (none adaptation: 21.5%).

10.21437/ICSLP.1992-106

1 INTRODUCTION

In the past few years many recognition methods which use continuous mixture density HMMs (CDHMMs) have been studied. Research has also been conducted on speaker adaptation methods using these CDHMMs. These methods are based on speaker-independent models or standard/reference speaker models [1, 2, 4, 5, 6, 7, 8, 18]. Both speaker adaptation methods can be regarded as retraining problems, where a small amount of training data is used. To retrain the HMM, the following two problems must be addressed:

Problem 1 : Not every phoneme is included in the typically small amount of training data available. Therefore, some phoneme CDHMMs are not trained.

Problem 2 : Errors in estimating CDHMM parameters can result from insufficient training data.

The **"Vector Field Smoothing method (VFS)"** is applied in order to deal with these general retraining problems. In "VFS" algorithm, the mapping from initial HMMs to retrained HMMs, i.e. from the reference speaker to a new speaker, is carried out according to a transfer vector field. "VFS" has three steps, i.e. **"Concatenation Training"**, **"Interpolation"** and **"Smoothing"**. First, the mean vectors of the Gaussian distributions in phoneme CDHMMs are trained by concatenation training. The transfer vector field is composed by transfer vectors calculated from the difference between the mean vectors of the initial CDHMMs and those created after retraining. The shifting of the mean vectors caused by the concatenation training is regarded as a transfer along the transfer vector. Second, the transfer vectors of untrained mean vectors are interpolated. **"Problem 1"** is solved in this step. However, since the training data does not represent the true distribution of each phoneme. To solve **"Problem 2"**, **"smoothing"** is carried out. The vector field is smoothed in accordance with the restored directions of the transfer vectors. There has been much research on speaking style to compensate for the difference between the speaking styles of the training data and the recognition data[8]-[12]. In this paper, "VFS" is applied to speaking-style adaptation.

Section 2 give details of the "VFS" algorithm. Section 3, evaluates the "VFS" algorithm using experiments on recognizing 23 Japanese phonemes. The potential of the "smoothing technique" is highlighted here. In section 4, "VFS" is evaluated by phrase recognition experiments using an HMM-LR with speaker- and speaking-style adaptation.

2 VFS ALGORITHM

For speaker adaptation with CDHMMs, the output probabilities (i.e. mean vectors and variances of the Gaussian distributions), branch factors and transition probabilities can be adapted. However, re-estimation of variances is critical with a small amount of data. As it is easier and more effective to adapt the mean vector than to adapt other parameters, the mean vector was chosen for adaptation. "VFS" has the following three steps:

1. Concatenation training

   In this step, the mean vectors of the Gaussian distribution in the CDHMMs are trained by concatenation training.

   (1) The CDHMMs of the reference speaker are used as the initial
   CDHMMs of the new speaker.

   (2) The mean vectors of the Gaussian distribution C^I of these initial CDHMMs are re-estimated using new speaker utterances, by the concatenation training technique. ($C^I = (c_1^I, \ldots, c_K^I)$, where c_K^I = the k-th mean vector in the initial CDHMMs and K = the number of all mean vectors in the initial CDHMMs.)

2. Interpolation

   In this step, the untrained mean vectors are transferred to the new speaker's voice space by using an interpolated transfer vector. This transfer vector is interpolated from the differences between mean vectors in CDHMMs before and after training [1, 2, 8, 13, 14, 18].

   In step 1, some of the mean vectors in the initial CDHMMs are not retrained, because not every phoneme is included in the typically small amount of training data available. In this paper, mean vectors are divided into two groups K_1 and K_2, where K_1 represents the group of retrained mean vectors and K_2 represents the untrained mean vectors. The retrained mean vectors are characterized by c_k^R ($k \in K_1$) while the untrained mean vectors are characterized by c_n^I ($n \in K_2$). The concept of "interpolation" is shown in Fig. 1.

   (1) Calculation of transfer vectors
   The differential vector v_k is calculated from the difference between the mean vectors of the initial CDHMMs ($c_k^I, k \in$

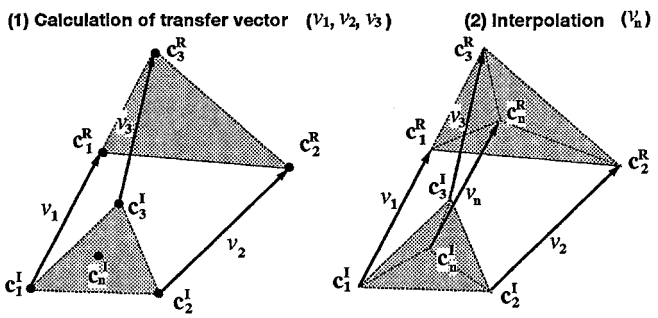


Figure 1: The concept of interpolation

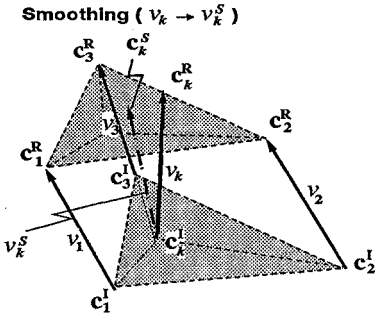


Figure 2: The concept of smoothing

K_1) and those created after retraining ($c_k^R, k \in K_1$). In this algorithm, v_k is called the "transfer vector", where

$$v_k = c_k^R - c_k^I \quad (k \in K_1).$$

The shifting from c_k^I to c_k^R caused by the concatenation training can also be regarded as a transfer along v_k.

(2) Interpolation

c_n^I ($n \in K_2$) is transferred to c_n^R using the transfer vector v_k, fuzzy membership function $\mu_{n,k}(f)$, and c_n^I, i.e.

$$\begin{aligned} v_n &= \sum_{k \in K_1} \mu_{n,k}(f) v_k, \\ c_n^R &= c_n^I + v_n, \\ \mu_{n,k}(f) &= 1/[\sum_{j \in K_1} (d_{n,k}/d_{n,j})^{1/(f-1)}], \end{aligned}$$

where $d_{n,k}$ is the distance between c_n^I and c_k^I. c_k^I is chosen using the k-nearest neighbors rule. $\mu_{n,k}(f)$ is the weight of the transfer vector v_k, and f is a weight control parameter called "fuzziness".

3. Smoothing of transfer vectors

In this step, each transfer vector is modified in accordance with the other transfer vectors. The concept of "smoothing" is shown in Fig. 2. The vector field is smoothed using these modified transfer vectors.

(1) Calculate the fuzzy membership function $\mu_{k,m}(f)$ ($k \neq m$, $m \in N(k)$), where $N(k)$ is the group of the mean vectors nearest to c_k^I and c_k^I. m is the index of the mean vector c_m^I.

(2) Calculate the transfer vector v_m.

(3) Calculate the smoothed transfer vector v_k^S using v_m and $\lambda_{k,m}(f)$, i.e.

$$v_k^S = \sum_{m \in N(k)} \alpha_m \lambda_{k,m}(f) v_m / \sum_{m \in N(k)} \alpha_m \lambda_{k,m}(f),$$

where α_m is a parameter representing the reliability of v_m. The degree of smoothing is controlled by f, and $\lambda_{k,m}(f)$ is defined by the following equation:

$$\lambda_{k,m}(f) = \begin{cases} 1 & \text{if } k = m \\ \mu_{k,m}(f) & \text{otherwise.} \end{cases}$$

(4) c_k^I is transferred to c_k^S using v_k^S and c_k^I, i.e.

$$c_k^S = c_k^I + v_k^S,$$

where c_k^S is the mean vector of the Gaussian distribution after smoothing.

In this paper, $\alpha_m = 1$ ($m \in K_1$), $\alpha_m = 0$ ($m \in K_2$). $\mu_{k,m}$ is calculated using all c_m^I ($m \in K_1$).

Table 1: Analysis conditions

pre-emphasis	$1 - 0.98z^{-1}$
window length	20.0 ms (Hamming window)
window shift	5 ms
LPC analysis order	16
LPC cepstrum order	16
Δ window length	50 ms

3 RECOGNITION OF 23 PHONEMES

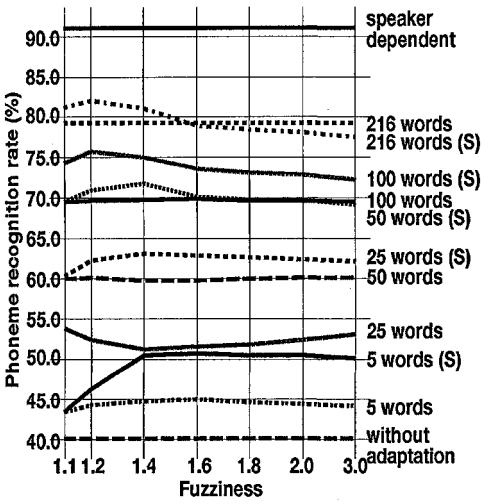
3.1 Experimental conditions

This experiment used the ATR speech database. The reference speaker was a male. The unknown speakers were one male and one female. The feature was a 34 dimensional vector consisting of 16 cepstral coefficients, 16 Δcepstral coefficients, logarithmic power and Δlogarithmic power. Analysis conditions are listed in Table 1. The number of mixture components per state depends on the phoneme HMM (2 $\sim$ 15 mixtures). Each of the mixture components had a diagonal covariance matrix. The total number of mixture components in the 49 HMMs is 1152. HMMs for consonants are 4-state and 3-loop models. HMMs for vowels, syllabic nasal sounds and silent periods are 2-state models. The phoneme HMMs of the reference speaker were trained using about 5500 isolated words.

3.2 Evaluation of "Smoothing"

In this section, the first 5, 25, 50, 100 or 216 words of the phonetically balanced 216 isolated words uttered by an unknown speaker were used as the adaptation data. The testing data was comprised of 23 Japanese phonemes extracted from the phrase data uttered by each unknown speaker. The test samples totaled about 1700 per person. The details of the utterances used for adaptation are shown in Table 2(a). A silent period of 100 ms was added before and after each word and used for adaptation.

Fig. 3 shows the average recognition rates of two unknown speakers as a function of fuzziness, the parameter to control "smoothing", and as a function of the amount of adaptation data. In Fig. 3 the following two results are observed: 1) As the number of training words increases, the optimal value of "fuzziness" decreases. 2) When "smoothing" is applied with optimal "fuzziness" the recognition rates are higher than those without "smoothing". The recognition rates are 69.9% and 79.3% for 100 and 216 isolated words respectively, when speaker adaptation without "smoothing" is applied. With "smoothing", these recognition rates increase to 75.8% and 82.1%. "Smoothing" reduces recognition error rates by 19.6% and 13.5% for these sets. These results show that "smoothing" is effective for HMM training with insufficient training data and that it can be controlled by "fuzziness". When 50 words are used for adaptation, "smoothing" reduces the recognition error rate by 29.3%. As shown in Table 2(a), almost all phonemes are included in the 50 word utterances. However, the amount of data in 50 words is insufficient for training each HMM. In



(S): with smoothing

Figure 3: Results of recognizing 23 phonemes

this case, "smoothing" technique can help to re-estimate the mean vectors.

3.3 Speaker and speaking-style adaptation

In the above experiments, the speaking style of utterances for speaker adaptation was different from that in the test data. This section describes two experiments that show "VFS" can carry out speaker and speaking-style adaptation simultaneously.

1. Speaker adaptation using phrase-by-phrase utterances

This experiment uses phrase-by-phrase utterances that have the same speaking style as the test data for speaker adaptation. The number of phrases used for adaptation and the recognition rates are shown in Table 2(b). Comparing Tables 2(a) and (b) shows that using phrase data for speaker adaptation gives better recognition performance was achieved with less adaptation data. For example, the recognition rate using 216 words (total speech length: 162 sec) for adaptation was 82.1%, while the rate using 11 sentences (total speech length: 66.3 sec) was 84.5%. The total speech length does not include a silent period of 100 ms added before and after each utterance.

2. Speaker adaptation using phrase and continuous speech

The above experiment showed the potential of speaking-style adaptation, by comparing two speaking styles, i.e. words and phrases. However, the differences between word data and phrase data used for adaptation depends not only on the speaking style, but also on the phoneme context. Next experiments were carried out to investigate the potential of speaking-style adaptation by using phrases and continuous utterances from the same context. As can be seen in Table 3, when the speaking style of the adaptation data is the same as that of the test data, the recognition rates were higher. This result shows that HMMs adapted using "VFS" contain the features of both the speaking style and the speaker.

This shows that speaker and speaking-style adaptation are carried out simultaneously using "VFS".

4 PHRASE RECOGNITION

In this section, VFS is evaluated for phrase recognition experiments using HMM-LR[11]. HMM-LR is a phrase recognition method based on HMMs and LR parsing. For the test data, 279 phrases were uttered by unknown speakers. The speaker adaptation used phonetically balanced 216 isolated words and phrase data. This phrase

Table 2: Recognition rates for 23 phonemes

(a) Using word utterances for adaptation

Number of words	5	25	50	100	216
Speech length (sec)	3.6	18.7	37.3	75.8	162.6
Recognition rates	50.7%	63.1%	71.8%	75.8%	82.1%
Number of retrained HMMs (total phonemes in training data)	11 (32)	38 (156)	43 (319)	46 (640)	49 (1359)

(b) Using phrase-by-phrase utterances for adaptation

Number of sentences (Number of phrases)	1 (7)	2 (13)	3 (20)	5 (47)	11 (102)
Speech length (sec)	4.0	8.3	13.1	29.5	66.3
Recognition rates	59.9%	63.2%	68.8%	78.4%	84.5%
Number of retrained HMMs (Total phonemes in training data)	22 (49)	27 (100)	30 (157)	38 (334)	42 (734)

Table 3: 23 phoneme recognition results on speaker- and speaking-style adaptation for words, phrases or continuous speech

	Test data: 23 phonemes extracted from phrase utterances	Test data: 23 phonemes extracted from continuous utterances
216 word adaptation	82.1%	76.3%
Speech length: 162.6 sec		
Without smoothing	79.3%	73.3%
102 phrase adaptation (11 sentences)	84.5%	77.1%
Speech length: 66.3 sec		
Without smoothing	83.6%	74.1%
11 sentence adaptation	76.6%	80.0%
Speech length: 51.4 sec		
Without smoothing	73.9%	77.5%
Speaker dependent (HMMs are trained using words)	91.0%	83.5%

The underlined figures show that the speaking styles of the adaptation data and test data data are identical.

data is different from the test phrase data. The basic HMM-LR system uses two duration control techniques. One is phoneme duration control for each HMM and the other is state duration control for each state. State durations are controlled using state duration penalties. These durations are calculated from word utterances uttered by the reference speaker and are modified using the duration of phrase utterances[15, 16]. However, the duration estimated from utterances of the reference speaker does not fit the duration of the unknown speaker. Therefore, these experiments evaluated the HMM-LR without state duration control.

As can be seen in Table 4(a)(b), the phrase recognition rates without state duration control are lower than those with state duration control. The decrease in the recognition rate is due to the increased insertion errors in 2-state HMMs, i.e. vowels and syllabic nasal sounds. 4-state and 3-loop structures were used for HMMs corresponding to vowels and syllabic nasal sounds in order to eliminate insertion errors. In the preceding experiments 23 phonemes, the number of mixture components of vowels and syllabic nasal models was 15 per state. However that of the 4-state models is 12 per state, therefore the Gaussian distributions in 49 HMMs total 1383. The phrase recognition rates using 4-state models are shown in Table 4(c).

The recognition rates of 4-state models were higher than those of 2-state models with state duration control except when 47 phrases were adapted. Specifically, the recognition rate of the 4-state models

Table 4: Phrase recognition results

	(a) With duration control	(b) Without duration control	(c) Without duration control
	Vowel Syllabic Nasal 2 states	Vowel Syllabic Nasal 2 states	Vowel Syllabic Nasal 4 states
Speaker-dependent	87.1%	75.8%	88.3%
None Adapt.	28.7%	13.6%	21.5%
50 words adapt.	70.0%	51.2%	71.5%
100 words adapt.	77.4%	56.7%	81.7%
216 words adapt.	82.0%	63.6%	82.6%
47 phrases adapt.	75.1%	54.8%	72.9%
102 phrases adapt.	79.0%	68.4%	85.1%

when 102 phrases were used was 6.1% higher than the rate of the 2-state models with duration control. This is for the following two reasons: 1) The parameters of state duration control estimated from training word samples of the reference speaker do not fit the parameters of phrase utterances of unknown speakers. 2) As can be seen in Table 5, the phoneme recognition rate of the 4-state model was higher than that for the 2-state models.

The above results show that speaker and speaking-style adaptation is carried out simultaneously and that the results of applying "VFS" to phrase recognition (85.1%) are similar to the results of applying the speaker-dependent model (88.3%) when one-minute speeches were used for adaptation.

5 CONCLUSIONS

This paper showed the effectiveness of a speaker adaptation method based on the transfer vector field smoothing model "VFS"with continuous mixture density HMMs. "VFS" was evaluated for recognizing 23 Japanese phonemes and for 279 phrase recognition and produced the following results:

1. For 23 Japanese phonemes, "smoothing" helped to re-estimate the mean vectors when the amount of training data was insufficient.

2. In phrase recognition experiments, speaker and speaking-style adaptation was carried out simultaneously and the recognition rate of "VFS" in phrase recognition (85.1%) was close to that of the speaker-dependent model (88.3%) when approximately one-minute speeches were used for adaptation.

In this paper, the mean vectors of the Gaussian distribution in CDHMMs were adapted from the CDHMMs of the reference speaker. Future plans include: 1) adaptation of the variance of the Gaussian distribution, branch factor and transition probability to unknown speakers, and 2) applying "VFS" to adaptation from a speaker-independent model to an unknown speaker.

ACKNOWLEDGMENTS

The authors wish to thank Dr. A. Kuramatsu for his support of this work. We would also like to acknowledge H. Hattori and K. Yamaguchi for their useful advice.

Table 5: Comparison of the number of states of the Vowel and Syllabic Nasal HMMs on experiments on recognizing 23 phonemes

	Vowel and S. Nasal 2 states	Vowel and S. Nasal 4 states
Speaker dependent	91.0%	92.1%
None adapt.	40.1%	47.3%
50 words adapt.	71.8% (60.1%)	74.5% (66.1%)
100 words adapt.	75.8% (69.9%)	78.3% (72.9%)
216 words adapt.	82.1% (79.3%)	84.7% (81.1%)
47 phrases adapt.	78.4% (73.5%)	81.5% (78.6%)
102 phrases adapt.	84.5% (83.6%)	87.4% (85.6%)

( ): The recognition rate without smoothing

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] Y. Nakato and H. Matsumoto. "A Study on Unsupervized Speaker Adaptation of Continuous Parameter HMM," Tech. report of IEICE, SP90-67, pp. 79-86 (Dec. 1990).

[3] S. Mizuta and K. Nakajima. "An Optimal Discriminative Training Method for Speaker-Independent Word Recognition Using Continuous Mixture Density Acoustic-Phonetic Segment HMMs," Tech. report of IEICE, SP91-58, pp. 21-28 (Sep. 1991).

[4] 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).

[5] 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).

[6] 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, Vol.39, pp. 806-814 (Apr. 1991).

[7] F. Kubala, R. Schwartz and C. Barry. "Speaker Adaptation from a Speaker-Independent Training Courpus," Proc. of ICASSP90, pp. 137-140 (Apr. 1990).

[8] K. Ohkura, M. Sugiyama and S. Sagayama. "Speaker Adaptation Based on Transfer Vector Field Smoothing Model with Continuous Mixture Density HMMs," Proc. of Meeting of Acoust. Soc. of Japan, 2-Q-17, pp. 191-192 (Mar. 1992).

[9] T. Matsuoka and K. Shikano. "Robust HMM Phoneme Modeling for Different Speaking Styles," Proc. of ICASSP91, S5.4, pp. 265-268, (May 1991).

[10] T. Hanazawa and K. Nakajima. "Study of Speaker Independent Continuous Speech Recognition Using Speaking-Style Dependent HMM," Proc. of Meeting of Acoust. Soc. of Japan, 2-1-1, pp. 51-52 (Mar. 1992).

[11] K. Yamaguchi and S. Sagayama. "HMM-LR Continuous Speech Recognition Using Continuous Mixture HMMs," Proc. of Meeting of Acoust. Soc. of Japan, 1-P-5, pp. 113-114 (Mar. 1992).

[12] Y. Kato and M. Sugiyama. "Continuous Speech Recognition Using Neural Networks with Multiple Input-Output Units," Proc. of Meeting of Acoust. Soc. of Japan, 3-1-1, pp. 71-72 (Mar. 1992).

[13] H. Hattori and S. Sagayama. "Speaker Adaptation with Small Size Data Based on Codebook Mapping Algorithm," Proc. of Meeting of Acoust. Soc. of Japan, 1-5-23, pp. 49-50 (Mar. 1991).

[14] Y. Shiraki and M. Honda. "Speaker Adaptation Algorithm for Segment Vocoder," Tech. report of IEICE, SP87-67, pp. 49-56 (Oct. 1987).

[15] T. Hanazawa, T. Kawabata and K. Shikano. "Recognition of Japanese Voiced Stops Using Hidden Markov Models," J. Acoust. Soc. Japan., Vol. 45, No. 10, pp. 776-783 (Oct. 1989).

[16] T. Kawabata, T. Hanazawa and K. Shikano. "Word Spotting Method Based on HMM Phoneme Recognition," Proc. of Meeting of Acoust. Soc. of Japan, 3-P-5, pp. 237-238 (Mar. 1988).

[17] S. Nakamura, T. Hanazawa and K. Shikano. "Phoneme Recognition Evaluation of HMM Speaker Adaptation Based on Vector Quantization," Tech. report of IEICE, SP88-106, pp. 1-8 (Dec. 1988).

[18] K. Shinoda, K. Iso, and T. Watanabe. "Speaker Adaptation For Demi-Syllable Based Continuous Density HMM," Proc. of ICASSP91, S13.7, pp. 857-860 (May. 1991).