Speaker Adaptation for Context-Dependent HMM using Spatial Relation of Both Phoneme Context Hierarchy and Speakers

Yasuhiro KOMORI, Tetsuo KOSAKA, Masayuki YAMADA,
and Hiroki YAMAMOTO
Media Technology Laboratory, Canon Inc.
890-12 Kashimada, Saiwai-ku, Kawasaki-shi, Kanagawa 211 JAPAN,
Email:komori@cis.canon.co.jp

Abstract

To realize good speaker adaptation for context dependent HMM using small-size training data, reasonable adaptation of unseen models have to be realized using the relation of appeared models and the training data. In the paper, a new speaker adaptation method for context dependent HMM using two spatial constraints is proposed: 1) spatial relation of the phoneme context hierarchical models, and 2) spatial relation between speaker specific models and speaker independent models. Several implementations based on the idea are proposed and are evaluated under 520 word speech recognition. 25 words are used for adaptation par speaker. The best result improved 30% error rate showing the effectiveness of the proposed method.

1 Introduction

Although the performance of the speaker independent speech recognizer is improved by the context dependent HMM, there still remains a room for improvement compared to that of the speaker dependent one. Thus, a better speaker adaptation with small-size training data is required. To realize good speaker adaptation for context dependent HMM using small-size training data, reasonable adaptation of unseen models have to be realized using the relation of appeared models and the training data. And for good speaker adaptation with small-size data, it is important to reduce the number of the parameter for speaker adaptation using reasonable constraints such as acoustic similarity and/or acoustic phonetic knowledge. There are some researches [1, 2, 3] aiming at the problem solution and good results are reported. The methods [1, 2] utilize acoustically clustered tree structure but they do not utilize the acoustic phonetic knowledge such as phoneme contextual constraint. Also, neither of the method [1, 2, 3] utilizes the spatial relation of phoneme contexts and speakers.

The proposed speaker adaptation in this paper differs in the following two points:

1) The proposed method utilizes constraints of spatial relation of phoneme context hierarchical models. The smoothing area depends on the higher level of the phoneme context hierarchy.

2) The proposed method utilizes constraints of spatial relation between speaker specific model and speaker independent model.

First, the paper describes the proposed speaker adaptation methods with some implementations. Then speech recognition experiment using the proposed methods is reported.

2 Speaker Adaptation Method

2.1 Information for Adaptation

The proposed method utilizes two spatial relations and these relations are obtained from the following models:

a) phoneme context hierarchy:

SICI-HMM : Speaker Independent Context Independent HMM

SICD-HMM : Speaker Independent Context Dependent HMM

b) speakers' space:

SDCI-HMM : Speaker Dependent(specific) Context Independent HMM

SICI-HMM : Speaker Independent Context Independent HMM

The speaker independent HMMs (SICI-HMM, SICD-HMM) can be trained in advance and the speaker dependent context independent HMM (SDCI-HMM) is trained by small-size data for

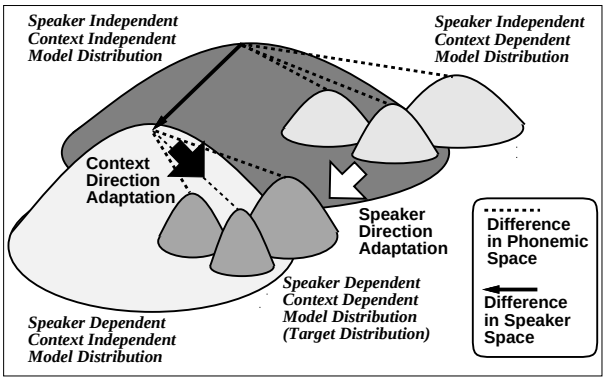

Figure 1: Relation of Speaker and Phoneme Context

speaker adaptation. Using these three models, finally the target model, the speaker dependent(adapted) context dependent HMM(SDCD-HMM) is created. Figure 1 shows the relations of speaker space and phoneme contextual space.

There are two different directions for speaker adaptation to obtain the SDCD-HMM using SICI-HMM, SICD-HMM and SDCI-HMM

1. speaker adaptation by mapping towards phoneme context direction (see figure 2)
2. speaker adaptation by mapping towards speaker difference direction (see figure 3).

All HMMs in the paper are three state models and are assumed that each state in the same phoneme class has close acoustical relation according to their sequence. The structure of SDCI-HMM is a single Gaussian model because small-size training data is required. The structure of SICI-HMM is also a single Gaussian model because of the facilitation to obtain the spatial relation of phoneme contexts and speakers. The structure of both SDCD-HMM(target model) and SICD-HMM is a multi-mixture model for better recognition performance.

2.2 Algorithm

First, two basic speaker adaptation realizations based on the idea are explained, then some modified implementations are presented.

The notations of the HMM parameters used for the explanation are shown in table 1.

Table 1: Notations of Each HMM

models		mean	variance
for adaptation	SDCI-HMM	μ_{CI}^{SD}	$\sigma^{2\,SD}_{CI}$
pre-trained	SICI-HMM	μ_{CI}^{SI}	$\sigma^{2\,SI}_{CI}$
pre-trained	SICD-HMM	μ_{CD}^{SI}	$\sigma^{2\,SI}_{CD}$
target	SDCD-HMM	μ_{CD}^{SD}	$\sigma^{2\,SD}_{CD}$

Method I-a

Locate the target SDCD-HMM at the relational position of SDCI-HMM using the difference of SICD-HMM and SICI-HMM considering the space of SDCI-HMM. This mapping performs to the phoneme context direction. The adaptation is realized by the next equation.

$$\mu_{CD}^{SD} = \mu_{CI}^{SD} + (\mu_{CD}^{SI} - \mu_{CI}^{SI})(\sigma_{CI}^{SD}/\sigma_{CI}^{SI})$$

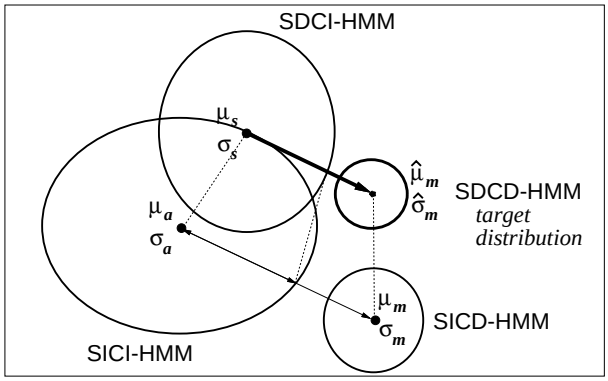

Figure 2: Speaker Adaptation Method I-a

Method II-a

Locate the target SDCD-HMM at the relational position of SICD-HMM using the difference of SDCI-HMM and SICI-HMM considering the space of SICD-HMM. This is a kind of speaker difference vector adaptation by separating the mapping space into each phoneme HMM state. The adaptation is realized by the next equation.

$$\mu_{CD}^{SD} = \mu_{CD}^{SI} + (\mu_{CI}^{SD} - \mu_{CI}^{SI})(\sigma_{CD}^{SI}/\sigma_{CI}^{SI})$$

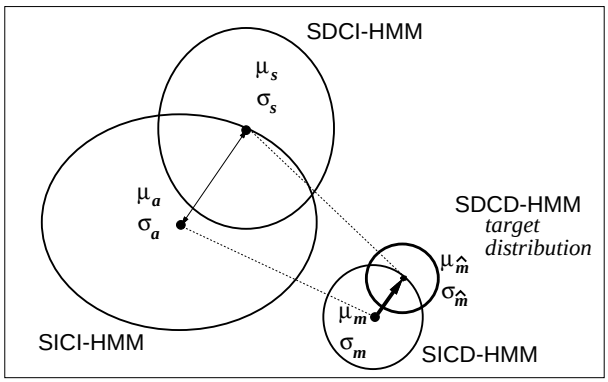

Figure 3: Speaker Adaptation Method II-a

Modified Methods

Other implementations using spatial information are realized by introducing the mixture composition

of SICD-HMM state. The mixture composition is realized by the next equation.

$$\mu_c = \sum_{k \in I} w_k \mu_k$$

$$\sigma_c^2 = \sum_{k \in I} w_k \sigma_k^2 + \sum_{k \in I} w_k (\mu_k - \mu_c)^2$$

where μ_c, σ_c^2 are the composite mixture, μ_k, σ_k^2 are the element in the state, w_k is the weight.

Method I-b

This method is a modification of method I-a using the state mixture composition. The adaptation is realized by the next equation.

$$\mu_{CD}^{SD} = \mu_{CI}^{SD} + (\mu_c - \mu_{CI}^{SI})(\sigma_{CI}^{SD}/\sigma_{CI}^{SI}) + (\mu_{CD}^{SI} - \mu_c)$$

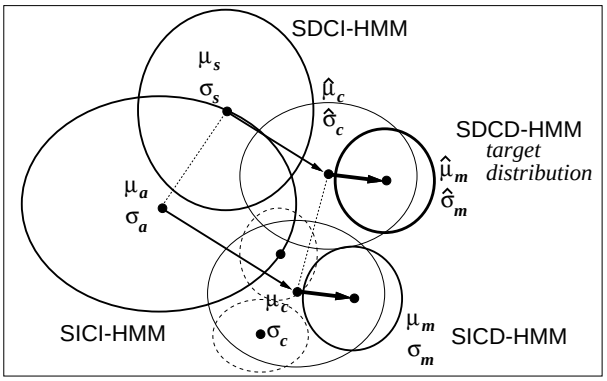

Figure 4: Speaker Adaptation Method I-b

Method II-b

This method is a modification of method II-a using the state mixture composition. The adaptation is realized by the next equation.

$$\mu_{CD}^{SD} = \mu_{CD}^{SI} + (\mu_{CI}^{SD} - \mu_{CI}^{SI})(\sigma_c/\sigma_{CI}^{SI})$$

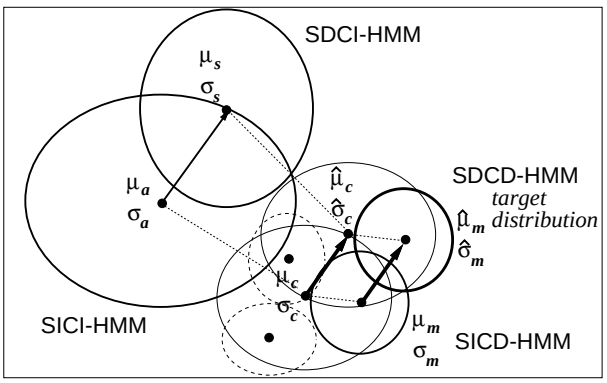

Figure 5: Speaker Adaptation Method II-b

Method II-c

This method is a modification of method II-a using the state mixture composition. The difference of method II-c to method II-b is the consideration of the area of the speaker independent space and the speaker dependent space ($\sigma_{CI}^{SD}/\sigma_{CI}^{SI}$). The adaptation is realized by the next equation.

$$\mu_{CD}^{SD} = \mu_c + (\mu_{CI}^{SD} - \mu_{CI}^{SI})(\sigma_c/\sigma_{CI}^{SI}) + (\mu_{CD}^{SI} - \mu_c)(\sigma_{CI}^{SD}/\sigma_{CI}^{SI})$$

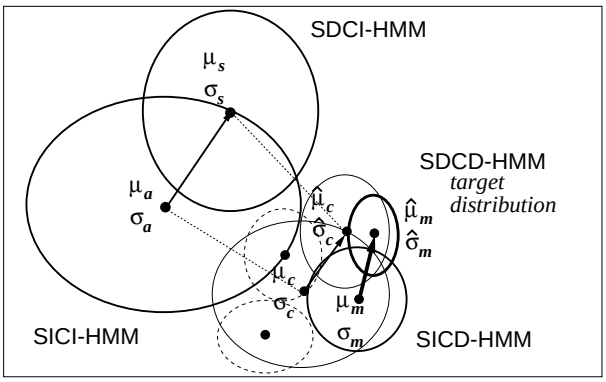

Figure 6: Speaker Adaptation Method II-c

Adaptation Ratio and Variance

Practically, we introduce an speaker adaptation ratio α (experimentally determined). The final adapted model is computed as:

$$\hat{\mu}_{CD}^{SD} = \alpha \mu_{CD}^{SD} + (1 - \alpha)\mu_{CD}^{SI}$$

The variance of SDCD-HMM is not adapted in this paper and is equal to that of SICD-HMM as:

$$\hat{\sigma^2}_{CD}^{SD} = \sigma^2{}_{CD}^{SI}$$

3 Experiments

3.1 Conditions

We performed a 520 word speech recognition experiment of 20 speakers. The step size of the speaker adaptation ratio is 0.1 ($0.0 \leq \alpha \leq 1.0$). 25 words are used for speaker adaptation to create the SDCI-HMM. SICI-HMM are pre-trained by the same 25 words of 60 speakers. Experimental conditions are briefly shown in table 3.

3.2 Results

The performance of speaker adaptation methods is shown in figure 7 along with the adaptation ratio (α). Details of the best result for α in each method are shown in table 2 with the accuracy of the best

improving speaker, the accuracy of the worst degradation speaker, and the number of speakers who improve and who get worse. Improvements and degradations of each speaker when method I-a gave the best result are also shown in the figure 8.

The **method I-a,b** resulted the best and improved form 95.6% to 97.0% on the average. The method I-a gave best improved speaker, +9.7% improvement. In this case, two speakers got worse while their degradation were only -1.9% and -1.0%.

4 Conclusion

The paper proposed a new speaker adaptation method for context dependent HMM using spatial relation of both phoneme context hierarchy and speakers. Several implementations are proposed. They are evaluated through 520 word speech recognition. 25 words are used for adaptation par speaker. The best result improved 30% error rate showing the effectiveness of the proposed method.

Acknowledgment

The authors wish to thank Dr. Hideyuki TAMURA, Head of the Media Technology Laboratory at Canon Inc., Mr. Minoru FUJITA, the manager of the Intelligent Media Division, Mr. Tsuyoshi YAGISAWA, the manager of our deportment, for giving us the opportunity of this study.

References

[1] P.C.Woodland, et al.: The 1994 HTK Large Vocabulary Speech Recognition System, Proc. ICASSP95, Vol.1, pp73-76, 1995.

[2] K.Shinoda, et al.: Speaker Adaptation with Autonomous Model Complexity Control By MDL Principle, ICASSP96, vol.2, pp.717-720, 1996-5.

[3] J.Ishii, et al.: A Study on Vector Field Smoothing Using Successive State Splitting Process: ASJ fall, 3-2-14, pp.133-134, 1995-9. (in Japanese)

Table 2: Details of the Results

	α	rate (%)	best (%)	worst (%)	get better	get worse
none	—	95.6	—	—	—	—
I-a	0.5	97.0	+9.7	-1.9	13	2
I-b	0.5	97.0	+8.7	-1.0	13	3
II-a	0.5	96.7	+4.8	-1.9	14	4
II-b	0.6	96.8	+4.8	-1.0	14	3
II-c	0.5	96.6	+5.8	-1.9	13	5

Table 3: Experiment Condition

Acoustic Analysis (25 dim.)	hamming: 25.6ms, pre-emphasis: 0.97, 8kHz sampling, frame: 10ms, LPC-Mel-Cep, ΔCep, Δpower
Eval. Data	ATR speech database(20 speakers) 520 words (104 words/speaker)
SICD-HMM	262 right context models, 3state6mix 200 speakers,72,000 utterances (ASJ+ATR+CANON speech data)
SICI-HMM (phone)	24 phone models, 3state1mix 25 words × 60 speakers
SDCI-HMM (phone)	24 phone models, 3state1mix 25 words of a specific speaker

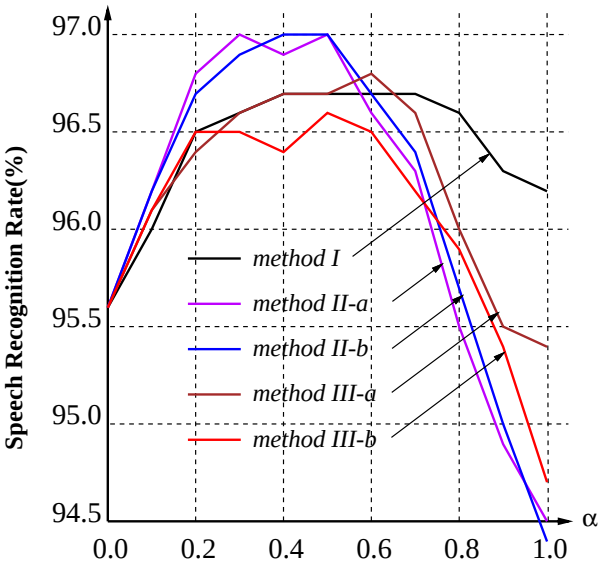

Figure 7: Recognition Rate of Each Method

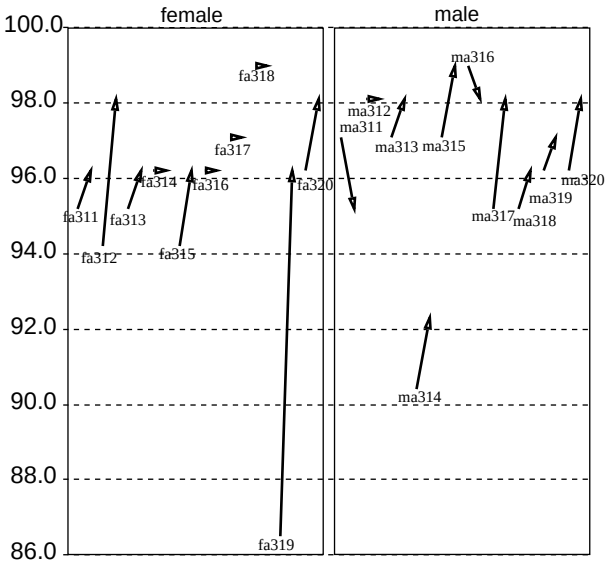

Figure 8: Each Speaker Improvement(Method1-a)