A Codebook Adaptation Algorithm for SCHMM Using Formant Distribution

Tae Young Yang[1] *Won Ho Shin*[1] *Weon Goo Kim*[2] *Dae Hee Youn*[1]

[1]ASSP Lab. Dept. of Electronics Eng. Yonsei Univ.
134 Shinchon-dong Sudaemun-ku
Seoul, 120-749, Korea
e-mail) young@caas.yonsei.ac.kr
[2]Dept. of Electronical Eng. Kunsan National Univ.
68 Miryong-dong Kunsan City
Chonbook, 573-360, Korea

ABSTRACT

This paper describes a codebook adaptation process improving the performance of speaker adaptation. The proposed method is performed prior to Bayesian speaker adaptation method using the formant distribution of adaptation data. The reference codebook is adapted to represent the formant distribution of a new speaker.
The average recognition rate of Bayesian adaptation is improved from 91.4% to 95.1% using the proposed method. The proposed method is effective particularly when there exists a large mismatch between the reference codebook and a target speaker in feature space. In this cases the average recognition rate is 95.0% while 89.9% is obtained when only Bayesian adaptation is performed.

1. INTRODUCTION

We consider speaker adaptation process for a semi-continuous hidden Markov model (SCHMM) based speech recognition system, in which speaker adaptation is performed in two stages. One is a vector quantization (VQ) stage using reference codebook, the other is a decoding stage using SCHMM parameters.

There are several approaches for adaptation in VQ stage such as spectral transformation [1], spectral mapping [2][3], and unsupervised hierarchical clustering scheme [4]. To adapt HMM parameters, most approaches have concentrated on supervised probabilistic reestimation of parameters which compensates for the differences between reference codebook and target speaker [5][6]. Particularly, this paper focuses on the speaker adaptation of the reference codebook in VQ stage.

Recently, Baysian adaptation method has been proposed for speaker adaptation and successfully adopted in several applications [7]. However, the performance of Bayesian adaptation method depends on the characteristics of a target speaker and it could be degraded when the features of the target speaker are severely different from those in the reference codebook. To solve such a problem, the proposed method uses the distribution of formant frequencies extracted from the cepstral coefficients of the reference codebook and adaptation data. The distribution of formant frequencies characterizes the reference codebook and the target speaker in feature space. In the beginning of the proposed method, the formant frequencies of cepstral coefficients in the reference codebook are adapted to represent the global distribution of the formant frequencies of the target speaker. It reduces the mismatch between the two acoustic characteristics. Then the formant frequencies of each codeword is precisely adapted to formant frequencies of the target speaker. The resulting *pre-adapted* codebook is used for Bayesian adaptation as an initial codebook.

2. CODEBOOK ADAPTATION USING FORMANT DISTRIBUTION

This section presents a new technique for codebook adaptation utilizing the distribution of formant frequencies. The goal of codebook adaptation is to seek accurate correspondence between the reference codebook and the target speaker in feature space so as to minimize the mislabelings and the distortions that can occur when the target speaker is quantized with the reference codebook. The reference codebook should be changed to cover the feature characteristic of the target speaker.

The distribution of formant frequencies is one clue that can determine the feature characteristics. In the proposed method, the formant frequencies of each codeword in reference codebook are distributed according to the formant distribution of the adaptation data from the speaker. Then, Bayesian adaptation is performed using the resulting codebook as an initial codebook. The detail procedure is as follows:

step-1> Voice/Unvoice Detection

Only the formants of voiced speech are used. The voiced speech frames can be detected by *clipping auto-correlation function* [8]. To detect the voiced codeword of the reference codebook, we first detect the

voiced frames of the training speech data from which the reference codebook is generated. Then, These voiced frames are quantized back with the reference codebook. The selected codewords are considered as codewords of voiced speech frame.

step-2> Formant Extraction

The formant frequencies are measured by *peak detection method* [9] applied to the spectrum. The strategy is to sequentially examine each value of log-scaled spectrum $S(n)$. A peak is defined to exist when $S(n) \geq S(n-1)$ and $S(n) \geq S(n+1)$. The log-scaled spectrum is obtained by taking FFT of the cepstral coefficients. In experiments, we used three formants.

step-3> Formant Adaptation

Formant adaptation is performed by following iterative procedure.

1) Select one frame randomly from the adaptation data.

2) Find K-neighborhood codewords which have similar formants to the selected frame of adaptation data. K represents the number of neighborhood codeword and is a function of iteration. In the beginning of iteration, K is sufficiently large so that the formant distribution of codewords in reference codebook can represent the formant distribution of the adaptation data globally. As iteration proceeds, K is diminished to only a single codeword, which provides each codeword can be precisely adapted to the adaptation data.

3) Calculate the jth adapted formant $\hat{f_{ij}^c}$ of the ith neighborhood codeword using the jth formant f_j^a of the selected adaptation data frame and the jth formant f_{ij}^c of the ith neighborhood codeword.

$$\hat{f_{ij}^c} = f_{ij}^c + \alpha(f_j^a - f_{ij}^c),\ 1 \leq i \leq K, 1 \leq j \leq F \quad (1)$$

where K is the number of codeword, F is the number of formant, and α is a learning coefficient, $0 < \alpha < 1$.

4) Go to 1) until the final iteration reaches.

step-4> Spectrum Mapping

The adapted spectrum of each codeword is estimated from the adapted formants obtained in step-3. The spectral envelope between two formants of each reference codeword is linearly mapped into between the adapted formants.

step-5> Feature Extraction

From the spectrum in step-4, the codewords of the adapted codebook are obtained by taking IFFT.

step-6> Bayesian Adaptation

Bayesian adaptation is performed with the codebook from step-5 as an initial codebook.

3. SCHMM PARAMETER ADAPTATION

The SCHMM model λ consists of three parameters; state transition probability $\mathbf{A}$, observation probability $\mathbf{B}$, initial state probability $\mathbf{\Pi}$. The adapted SCHMM is achieved by modifying the observation probability $\mathbf{B}$. For a given adaptation data sequence y_t of the target speaker at time t, the observation probability $b_j^y(l)$ of the lth codeword in state j is obtained.

$$b_j^y(l) = \frac{\sum_{t=1}^{T} \delta(s_t = j)\delta(y_t \cong l)}{\sum_{t=1}^{T} \delta(s_t = j)},\ 1 \leq j \leq N, 1 \leq l \leq L \quad (2)$$

where,

$$\delta(s_t = j) = \begin{cases} 1, & \text{if the state } s_t \text{ of } \lambda \text{ is } j \text{ at } t \\ 0, & \text{otherwise} \end{cases}$$

$$\delta(y_t \cong l) = \begin{cases} 1, & \text{if } l\text{th codeword is selected for } y_t \\ 0, & \text{otherwise} \end{cases}$$

Then, the adapted observation probability of the lth codeword in state j is measured as follows;

$$\hat{b_j}(l) = \beta b_j^y(l) + (1-\beta)b_j(l),\ 1 \leq j \leq N, 1 \leq l \leq L \quad (3)$$

where β is a learning coefficient, $0 < \beta < 1$.

4. EXPERIMENTS AND RESULTS

To confirm the effectiveness of the proposed technique, isolated-word recognition experiments are carried out on the 61-word set [10] which covers all of the Korean phonemes. Three data sets are needed to train the SCHMM recognizer. The first and second sets consist of 10 utterances of the 61 words from one male and one female speakers to obtain male-trained speaker-dependent model and female-trained speaker-dependent model, respectively. The third one consists of 3 utterances from 12 speakers (7 males and 5 females) to obtain speaker-independent model. For evaluation, we use the set of 3 utterances of each word from two male speakers and two female speakers. The two speaker-dependent models are adapted to the four test speakers in two adaptation schemes and compared with speaker-independent model.

All of the speech signals are low-pass filtered with the cutoff frequency of 5 kHz, sampled at 10 kHz, pre-emphasized by $1 - 0.95z^{-1}$, and analyzed at every 10 ms with 20 ms Hamming window. The feature used is a vector of 14 elements cepstral coefficients obtained from

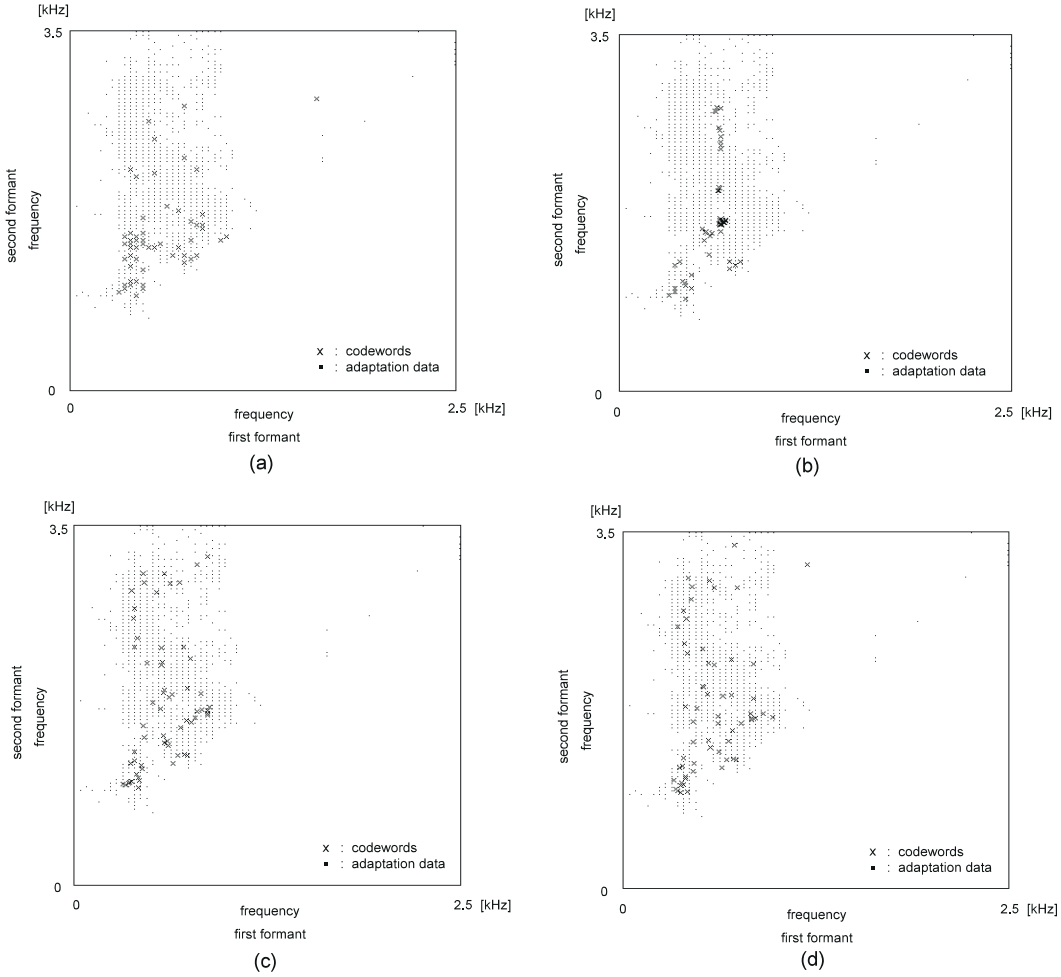

Figure 1: An example of formant adaptation process: (a) iteration=0, K=30 (b) iteration=200, K=10 (c) iteration=1000, K=5 (d) iteration=3000, K=1

auto-correlation and LPC analysis.

The recognition rates of 98.4% and 97.8% were obtained from male and female speaker-dependent models, respectively. Table1 shows the speaker-independent recognition rate. The average recognition rate is 83.2% for the four test speakers. The speaker-adaptive models are obtained from the two speaker-dependent models using one utterance of each words from four test speakers.

Table 1: The recognition rates of speaker-independent recognizer. M1 and M2 denote male speakers and F1 and F2 denote female speakers. [%]

	M1	M2	F1	F2	mean
SI	78.7	86.3	80.3	87.4	83.2

Table 2 shows the speaker-adaptive recognition results, where SAB means speaker-adaptive model obtained from conventional Bayesian speaker adaptation process and SAF means speaker-adaptive model from Bayesian speaker adaptation with proposed codebook adaptation process using the formant information. Recognizing the four test speakers without speaker adaptation process, the average recognition rates are 57.9% and 58.7% for speaker-dependent models trained by a male reference speaker and by a female reference speaker, respectively. For SAB model the average rates of the four test speakers are improved to 91.8% and 91.0%. After applying the proposed codebook adaptation process, for SAF model, the average rates are significantly improved to 94.3% and 95.8%.

The performance improvements are more significant when there exists a serious mismatch between the reference codebook and the target speaker. Figure 1 shows an

Table 2: The recognition rates of speaker-adaptive recognizer: (a) SA is obtained from SD trained by a male reference speaker, (b) SA is obtained from SD trained by a female reference speaker. [%]

	SD trained by a male speaker				
	M1	M2	F1	F2	mean
no adapt	60.1	74.3	43.2	54.1	57.9
SAB	95.1	92.9	87.4	91.8	91.8
SAF	96.7	94.0	92.3	94.0	94.3

(a)

	SD trained by a female speaker				
	M1	M2	F1	F2	mean
no adapt	56.3	45.9	63.9	68.6	58.7
SAB	91.3	89.1	89.6	93.9	91.0
SAF	97.3	96.2	93.4	96.2	95.8

(b)

example of the formant adaptation process in the proposed method. The mismatch of the formant frequences between the reference codebook and the target speaker is reduced after the formant adaptation process. It can provide accurate correspondence between reference codebook and target speaker in feature space. In case that the recognizer trained by a male speaker is adapted to female speakers, the average recognition rate of SAF is improved by 3.6% compared to that of SAB. When the model of a female speaker is adapted to male speakers the average recognition rate of SAF is improved by 6.6% compared to that of SAB. These results verify the effectiveness of the proposed codebook adaptation method.

5. CONCLUSION

A codebook adaptation method using the distribution of formant frequencies is presented. The proposed method is applied to speakeradaptive recognizer with Bayesian adaptation method. The speaker-adaptive recognizer obtained from this adaptation scheme is compared with speaker-dependent, speaker-independent, and speaker-adaptive recognizer using Bayesian adaptation method alone.

Experimental results show that proposed method improves the recognition performance of the speaker-adaptive recognizer. The advantages of the proposed method are that it is relatively independent of HMM parameter adaptation stage and that it can use text-independent speech as adaptation data.

6. REFERENCES

1. F. Class, A. Kaltenmeier, and P. Regal, "Fast Speaker Adaptation Combined with Soft Vector Quantization in an HMM Speech Recognition System," *Proc. ICASSP*, pp. 461–464, Mar. 1992.
2. S. Nakamura and K. Shikano, "A Comparative Study of Spectral Mapping for Speaker Adaptation," *Proc. ICASSP*, pp. 157–160, Apr. 1990.
3. H. Matsukoto and H. Inoue, "A Piecewise Linear Spectral Mapping for Supervised Speaker Adaptation," *Proc. ICASSP*, pp. 449–452, Mar. 1992.
4. S. Furui, "Unsupervised Speaker Adaptation Method Based on Hierarchical Spectral Clustering," *Proc. ICASSP*, pp. 286–289, May 1989.
5. C.-H. Lee, C.-H. Lin, and B.-J. Huang, "A Study on Speaker Adaptation of Continuous Density HMM Parameters," *Proc. ICASSP*, pp. 145–148, Apr. 1990.
6. G. Rigoll, "Speaker Adaptation Using Improved Speaker Markov Models," *Proc. ICASSP*, pp. 566–569, Apr. 1993.
7. C.-H. Lee and J.-L. Gauvain, "Speaker Adaptation Based on MAP Estimation of HMM Parameters," *Proc. ICASSP*, pp. 558–561, Apr. 1993.
8. L. R. Rabiner, "On the Use of Autocorrelation Analysis for Pitch Detection," *IEEE Trans. ASSP*, Vol. 25, No.1, pp. 24–33, Feb. 1977.
9. F. M. Wang, P. Kabal, R. P. Pamachandran, and D. O'Shaughnessy, "Frequency Domain Adaptive Postfiltering for Enhancement of Noisy Speech," *Speech Communication*, Vol. 12, No. 1, pp. 41–56, Mar. 1993.
10. Korean Broadcasting Station, *A Korean Pronunciation Dictionary*, Aumoongack, Seoul, Korea, 1993.