STUDIA PHONOLOGICA XXV (1991)

Improved Speaker Markov Modelling for Unsupervised Speaker Normalization

Pascale FUNG, Tatsuya KAWAHARA, Shuji DOSHITA
and Martine ADDA

ABSTRACT

We propose new methods of improved speech recognition with speaker-variable information. Hidden Markov Model-based recognizers which are trained by reference speaker(s) (RS) are normalized by our two different approaches to give a better speaker-independent recognition rate. Our normalization methods are based on the same principle of inter-speaker Markov mapping. This mapping gives inter-speaker parameters which are used differently in our two approaches. The first Speaker Markov Model Converter (SMMC) converts new speaker spectral data into label data similar to that of the reference speaker utterance, which is passed directly to the recognizer. In the second Integrated Markov Model (IMM) approach, inter-speaker emission probabilities (ISE) are integrated as weights to the HMM emission probabilities. The recognizer in this case is modified according to inter-speaker variable information whereas the normalization is done in context. The inter-speaker mapping in both cases are unsupervised to save new speaker (NS) effort. HMM score thresholding, template matching and DP thresholding techniques are applied to select suitable data for unsupervised mapping of NS and RS data. This mapping is done in parallel to the recognition process. Iterations are performed to improve the unsupervised mapping.

1. INTRODUCTION

In automatic speech recognition (ASR) tasks, we strive to recognize sequences of phonemes, words, or sentences given spectral input from signal analysis front-end. The main goals are to achieve a reeognition rate as high as possible while keeping the system as robust, time-efficient and user-friendly as possible. To reach an optimal solution for such ASR systems, we need to deal with actual environmental and human variabilities.

One variability that affects the system robustness to a great extent lies in the differences between speakers. All ASR systems go through a training process by one or multiple reference speakers (RS) and are modelled after the utterances of the RS. When a new speaker (NS) makes input utterances, the system tries to recognize them by assuming that they come from the RS. Such assumption leads to rec-

Pascale FUNG, Tatsuya KAWAHARA (河原達也), Shuji DOSHITA (堂下修司): Department of Information Science, Faculty of Engineering, Kyoto University (Ms. Fung was a visiting researcher at LIMSI, CNRS, France from Apr. to Sep. 1991, and currently at BBN Inc. in Cambridge, Mass., USA)
Martine ADDA: LIMSI, CNRS, France

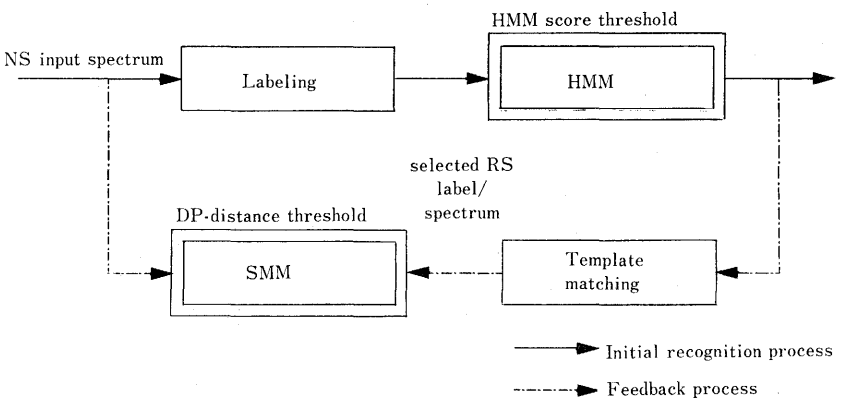

Fig. 1. Feedback process

ognition errors and lowers the system performance.

In speaker-independent recognition systems, speaker variabilities are often treated by having multiple RS training. Such RS aim to cover a wide range of speaker individualities [4]. However, there are always inevitable distortions.

Speaker adaptation is used to eliminate or minimize inter-speaker distortions. It performs a mapping from RS to NS (adaptation) or vice versa (normalization). Our approach uses a spectrum to code mapping where input NS spectrum is mapped to RS spectral and label data to give a Markov model between speakers. In the first Speaker Markov Model Converter case, input NS spectral data is converted into RS-similar code sequences which go directly through the recognition system. In other words, given NS data, the SMMC system predicts how the RS would have made the same utternce in the place of NS. In the IMM case, input NS data is not converted explicitly. The system takes in inter-speaker parameters as weights in the recognition process. And the weighed recognition system recognizes spectral data directly into dictionary templates.

Inter-speaker mappings done by most other research works [2, 3, 5, 9, 12, 13] require to be supervised, that is we have to map two sets of same utterances of NS and RS, therefore require special training data from NS. Yet we have learned that, in real situations, even a minimal training utterance requirement causes some inconvenience in NS. It therefore becomes our chief concern that adaptation process be unsupervised so that no training data is required from NS. In both approaches that we propose in this paper, the mapping is unsupervised.

2. UNSUPERVISED SPEAKER NORMALIZATION

In order to eliminate training data and time of the new speaker, the inter-speaker mapping of our Converter as well as the IMM system are obtained unsupervised in parallel with the recognition process. This unsupervised inter-speaker mapping process relies on the combined efforts of recognition score thresholding, template matching, and DP thresholding.

The unsupervised mapping is based on a feedback process as shown below and as illustrated in Figure 1

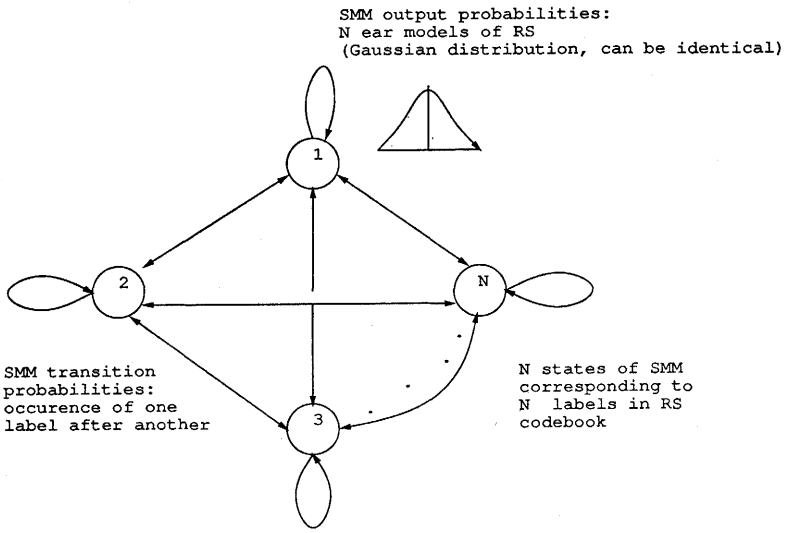

Fig. 2. Markov modelling of inter-speaker variations

1: Initial NS input spectral data are labelled by the RS codebook.
2: The HMM recognition system outputs recognized samples together with their scores.
3: The output phoneme sequence is screened by a recognition score threshold and then goes through template selection from the RS training corpus to give plauble RS spectral/label sequence and the corresponding NS spectrum.
4: This selected RS spectrum and label sequence is fed back to do the mapping with the corresponding NS spectrum.

These steps are iterated until the final recognition system is well established.

In step 3, each of the m NS recognized samples is matched by the same template present in the training data of every RS (n in number) given out $n \times m$ pairs of matching templates. In template matching, we assume that recognized templates which cannot find their match in RS training data are incorrect ones and are skipped. HMM score threshold and template-selection eliminate some incorrectly recognized samples. This process sholud be such that more incorrect ones are eliminated than correct ones, and the resulting ratio of correct to incorrect templates should be higher than that of the recognition result in step 1. Here we assume that incorrect ones tend to get lower scores. The HMM score threshold is chosen according to this consideration. The mapping is done by the Dynamic Programming process which also contribute to the elimination of mismatched NS/RS template pairs as described in the next section.

After each iteration, the system performance is evaluated by the recognition rate of NS input data through it. We chose to stop feedback after a fixed number of iterations which used the NS database comparable to data size needed for retraining HMM recognition system as in the speaker-dependent case. However, if only eliminating NS training data is of concern, then the limit can be set according to improvement rate, training time desired, or final improved recognition rate threshold.

Overall computation time comprises of several initial recognition-mapping-final recognition cycles. After the system is well established or system improvement reaches certain limit, however, only a straight forward final recognition process is carried out. The difference between SMMC and the IMM system lies in how the inter-speaker parameters are used in the final adaptation-recognition process. In the SMMC, final recognition has two steps - conversion of NS spectral data to RS labels, then recognition. In the IMM, final recognition takes in NS spectral data and performs a weighing in the recognition process.

3. Markov Modelling of Inter-Speaker Variation

We choose to use a spectrum-to-code mapping to model the inter-speaker variation since the speaker spectral data reflects speaker features directly. Meanwhile, we want to model such variation by parameters so that it can be easily monitered. Moreover, we know that speaker dynamic features are also important in normalization. Hidden Markov Models in speech recognition have been shown to be efficient in capturing such features and are parametric. Therefore, we develop an improved form of the Speaker Markov Models proposed in [12].

The feedback process gives a stream of recognized samples which are screened by the HMM score threshold. This label stream is used to template-match with the training data of each of the n RS. Its original spectral data are also kept with the label data. This NS spectrum and corresponding RS spectrum and label sequence obtained from template-matching are selected to perform inter-speaker mapping for the Speaker Markov Models.

1: The transition probabilities of the Speaker Markov Model (SMM) are calculated by counting how often one symbol follows another one in the RS label stream.

2: The initial distribution of SMM is found by counting how often each symbol occurs at the beginning of RS label stream.

3: The NS spectrum and RS spectrum are time aligned by Dynamic Programming (DP). For each symbol in RS label stream, there is a vector in the corresponding NS time window. The mean values and covariance matrices of the inter-speaker emission probabilities (ISE) of each label are thus found. Our ISE parameter is common for all labels in the RS codebook. This is the only parameter that will be used in both SMMC and IMM approach. In the IMM system, no inter-speaker transitional or initial probabilities are used. Instead, the modelling of RS behavior relies completely on the HMM models from the original RS training.

The DP distance threshold can be used to eliminate further mismatching between NS and RS data. NS/RS couples with DP distance exceeding certain threshold are not considered in the calculation of inter-speaker parameters. Again, we assume that couples with very big DP distance are usually mismatched.

4. Conversion by SMMC

The Viterbi algorithm is used to produce the optimal RS label sequence from NS

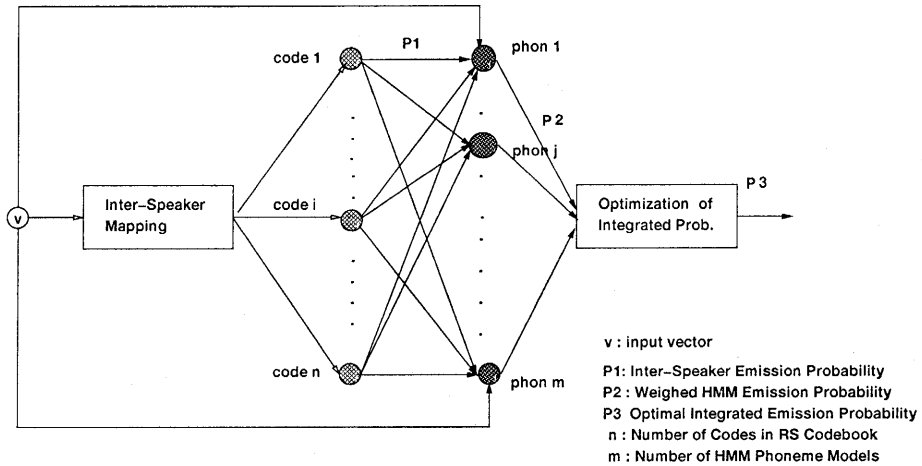

Fig. 3. Final Output Probabilities by Integrating Speaker Parameters

input speech data in SMMC. Given NS input spectral data, the Viterbi algorithm traces the plausible output state sequence according to the output probability, the transition probability and the initial probability of each label in RS obtained from SMMC training. The obtained state sequence is nothing other than the converted label sequence.

5. IMM as Weights in Recognition

In this approach, we modify the SMMC system. We do not use SMM initial and transitional parameters. This is due to the belief that since such probabilities are not phoneme-dependent, they do not give us additional information of RS behavior. As a result, they are negligible in significance compared to normal HMM initial and transitional probabilities in describing the RS utterance patterns. On top of improving SMM parameters, we are most interested in impvoving the transformation itself. The one-to-one conversion as done by the Viterbi algorithm in SMMC has the disadvantage of making a conversion decision frame-by-frame which are subject to distortions. Therefore, it is better to use a less restrictive distributed weighing function of all probabilities in making the final transformation-adaptation. Moreover, SMM conversion was done directly on spectral vectors at each frame considering only the transition from the previous frame to the present one. We believe we can obtain a more reliable transformation if such process is "guided" by some higher-level syntactical information from the context. In both isolated-word or continuous speech context, this information comes from the lexical model. That is to say, the transformation process and the recognition process should be interactive.

To achieve this goal, we integrate the ISE into the recognition process as weights, (see Figure 3). The optimum integrated score for all RS phoneme models at state i are computed by combining the HMM phoneme output probability with ISE obtained from the Speaker Markov Model estimation for RS label b and NS spectral vector v:

$$P_{INT}(v|i) = max_b[P_{HMMout}(b|i) * P_{ISE}(v|b)]$$

Table 1 Condition of experiments

Daia set	DS 1	DS 2	DS 3
total number of phonemes (HMMs)	26	11	11
RS speakers	2 male	2 male	10 male
NS utterances	585	290	290
RS total utterances	1164	575	1142
HMM training samples (RS)	1164	575	1142
HMM testing samples (closed)	RS 1164	RS 575	RS 1142
HMM testing samples (open)	NS 585	NS 290	NS 290

In other words, given new speaker input spectral data, the system recognizes which template (word, sentence) it is by weighing the difference between this utterace and all the templates in the reference speaker training corpus. At the meantime, the description of such difference is aided by considering the syntactical context in isolated-word or continuous speech recognition:

$$P_{likelihood}(v|i)=max_{i-1}[P_{liklihood}(v_{prev}, i-1)]*P_{HMMtrans}(i|i-1)*P_{INT}(v|i)$$

Here, the SMM transition from the previous frame label to the next frame label is given by the HMM phoneme models. Since the ISE of each frame vector is dependent on the previous frame vector, it is therefore dependent on the sequence of HMM models contained in the word or sentence context. We can see that for each word or sentence utterance in the dictionary, there is a distinctive way by which inter-speaker difference is described.

6. EXPERIMENTAL RESULTS

Due to the collaboration effort of the authors, the experiments were done in two different laboratories in Japan and France, using Japanese database for the first approach and French database for the second.

6.1 *Speaker Markov Model Converter*

We evaluated our SMMC method in phoneme (consonant) recognition. Input NS data was a sequence of isolated Japanese words each containing one vowel-consonant-vowel phoneme sample spectrum extracted manually. The recognition system was a speaker-independent Hidden Markov Model phoneme recognition system, trained by multiple RS phoneme sequence contained in isolated words. There were 26 HMM phoneme models in one case and 11 in another experiment. The two different sets of samples were chosen to provide different initial rates for the recognition system so as to see how the Converter performance is affected. In addition, we also investigated the effect of having more RS on the Converter and recognition performance.

The condition of the experiments for SMMC is listed in Table 1. LPC analysis was used for the signals, sampled at 16 kHz, 16bits.

In DS 1, one of the RS uttered 581 words while the other one made 583 utteran-

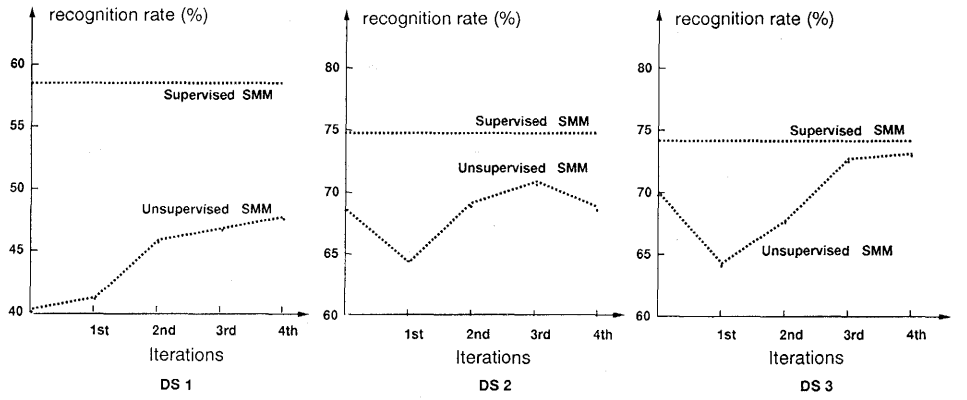

Fig. 4. Recognition results of the SMMC-HMM system

Table 2 Data size at different stages of SMMC

data	ite.	initial NS data	correct	HMM select.	correct	* temp. match.	correct	* DP match.	correct
DS 1	1	146	34.2%	115	38.3%	95	46.3%	75	52.0%
	2	292	39.4%	237	42.6%	199	50.6%	82	57.2%
	3	438	37.7%	357	40.6%	303	47.7%	128	53.5%
	4	585	40.3%	483	43.6%	415	50.4%	173	56.7%
DS 2	1	83	66.3%	73	68.5%	65	76.9%	55	81.7%
	2	172	68.0%	149	71.8%	135	78.9%	116	84.1%
	3	224	67.0%	199	70.4%	185	75.4%	160	81.2%
	4	290	68.6%	256	72.3%	241	76.6%	210	82.1%
DS 3	1	72	69.4%	67	71.6%	60	73.6%	53	82.4%
	2	145	69.0%	131	70.2%	119	77.4%	104	81.1%
	3	216	70.4%	196	72.4%	181	78.1%	161	82.3%
	4	290	70.0%	260	73.1%	241	77.8%	209	81.5%

** (virtual NS/RS template pairs)/(number of RS)*

ces. In DS 2, only those RS utterances from DS 1 which contained the 11 phonemes wanted were chosen as training data. In DS 3, 8 other speakers making a total utterance of 567 words were added to DS 2 to form a larger RS training sample group with 11 phonemes.

To test the SMMC performance, each NS data set was divided into four smaller groups to go through iterations. After each iteration, all the four groups of data were converted and recognized to evaluate Converter performance and improvement of the HMM recognition system. In addition, superivsed training was performed to compare with and evaluate unsupervised training performance. In supervised training, template-matching was supervised to ensure no mismatching. We performed supervised training using the entire data group.

Since the recognition accuracy for the five Japanese vowels reaches above 90% in general, we assumed vowels preceding and following the consonants were recognized and known. This information was used for both supervised and unsupervised template matching. The results are shown as in Figure 4.

Although 11 phoneme data set gives better initial rate of NS than 26 phoneme data set, the improvement rate of recognition is higher in the 26 phoneme set. We

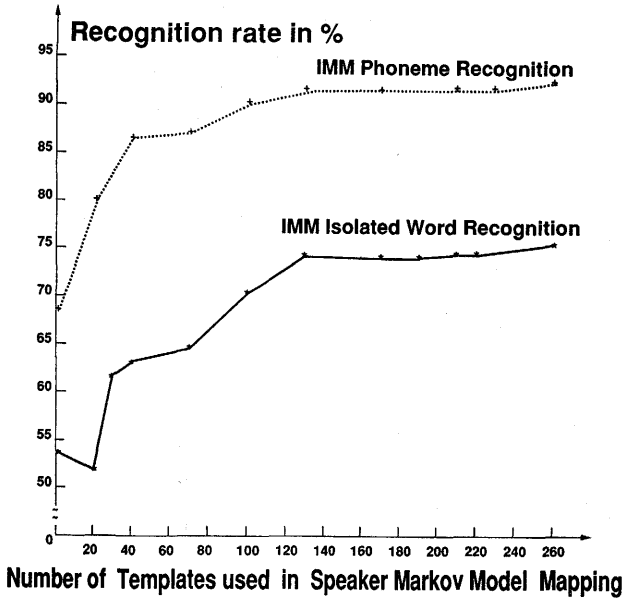

Fig. 5. Results of IMM Recognition System

can deduce from the supervised SMMC performance that improvement rate decreases as initial rate rises. The effect of having more RS is that it gives a better initial NS rate as expected. Also, having more RS training samples brings the unsupervised SMMC performance on recognition improvement closer to supervised SMMC performance when we compare DS3 with other data sets. Since supervised training on DS3 can only bring a 4% increase in NS recognition rate, the unsupervised process can be considered to have performed well with its 3% increase.

It is also evident that iteration improves SMMC performance, this is attributed to an increase in input NS sample that can be used to train SMMC. However, the improvement rates of all data groups are not linear. Although the rate is positive in general, there is a momentary negative rate in DS2 and a local maximum in DS3. This shows that there is a saturation point for SMMC improvement at least locally.

Furthermore, we observed the effectiveness of HMM score threshold, template-matching, and DP distance threshold in improving the rate of correctly matched NS/RS data couples to incorrect ones as shown below.

As we can see from Table 2, HMM score threshold eliminates more incorrectly recognized samples than correct ones. Moreover, template matching and DP distance threshold did the same on mismatching templates. These steps help to give more plausible matching between NS and RS since the ratio of correct matching to mismatching is increased after each step.

6.2 *Integrated Markov Model Recognition System*

For this approach, we have performed the experiments using one male speaker as the reference for training the HMM isolated-word recognition system, another male speaker as the new speaker. The database is 500 isolated French words for each

speaker. There are 2421 phonemes in the lexicon. Speaker-dependent recognition rates for teh reference speaker and the new speaker is 93% and 92% respectively. Speaker-independent rate is 53.6%.

Experiments have been done using new speaker data set, divided into many groups, and passed to the IMM system group by group to further evaluate the effect of unsupervised mapping. In the initial zero iteration, no speaker information is used. In the final iteration, all recognized words over the HMM score threshold are used for mapping. The performance of the system at each iteration point is evaluated by passing all the 500 NS words into the system and shown in Figure 5. The recognition result of the phonemes contained in these words is also presented in Figure 5.

By comparison, the IMM system has a more steady increase in performance over the iterations than SMMC and we do not observe a local maximum in its graph.

7. Discussion

The greatest advantage of our method is that no extra training utterance is required for NS. NS user only needs to speak what is to be recognized, and the system is modified at the same time. Compared with conventional SMM systems[12], our approach eliminates any amount of NS test data and any retraining of the recognition system. Both the SMMC and the IMM system switch automatically for different NS without any need for specification since it is dependent upon NS input data.

To realize unsupervised inter-speaker mapping, HMM score threshold, template selection and DP distance threshold have been proved to be effective in reducing NS/RS mismatching.

From the SMMC experimental results, we can also conclude that as closed recognition rate decreases with the increase of the number of reference speakers, the possible range for speaker normalization improvement is also reduced. When the difference between RS closed recognition rate and NS open recognition rate is big as with DS1, the range for SMMC performance is large as well. A perfect SMMC performance would bring the open recognition rate to the level of closed recognition rate. A poorer SMMC performance would still have more room to improve the open recognition rate. In general, it is easier to improve a recognition system that performs poorly on NS than one that performs well initially.

From the results of the IMM, we can say that this system is efficient in combining the acoustical phonetic information, higher-level syntacitcal information from the lexicon as well as the inter-speaker variable information to give an improved speaker-independent recognition rate. It performs better than the non-adapted speaker-independent system and is more robust than the SMMC system. It saves the conversion time needed for the SMMC. It also shows that HMM phonetic models are sufficient in describing RS behavior. Thus the SMM initial distribution and transitional porbabilities are not necessary for inter-speaker mapping. Moreover, we have observed the general trend of system improvement and conclude that if we take enough input data at each iteration point, the system is guaranteed to improve its performance. In other words, any set back in the system improve-

ment is only local and can be avoided by paying attention to the interval of iterations.

We have also observed that in isolated word recognition task, there is less mismatching between word templates as in vowel-consonant-vowel sets. As a result, DP-distance thresholding can be skipped in such a case when HMM score thresholding and template-matching have already eliminated many mismatched NS/RS template pairs.

It is our next step to compare SMMC and IMM using the same database to further confirm our above conclusions. We will also perform IMM experiments on different RS and NS to observe its general performance.

REFERENCES

[1] Lalit R. Bahl, Peter F. Brown, Peter V. de Souza and L. Mercer: Speech Recognition with Continuous-Parameter Hidden Markov Models. *Proc. IEEE-ICASSP, New York* pp. 40–43, 1988.

[2] H. Bonneau and J.L. Gauvain: Vector Quantization for Speaker Adaptation. *Proc. IEEE-ICASSP, Dallas* pp 34–37, 1987.

[3] M.W. Feng et al: Improved Speaker Adaptation Using Text Dependent Spectral Mappings. *Proc. IEEE-ICASSP, New York* pp 131–134, 1988.

[4] Sadaoki Furui: Digital Speech Processing, Synthesis, and Recognition. *Marcel Dekker Inc.* 1989.

[5] Sadaoki Furui: Unsupervised Speaker Adaptation Method Based on Hierarchical Spectral Clustering. *Proc. IEEE-ICASSP, Glasgow* pp 286–289, 1989.

[6] Hiroaki Hattori: Speaker Adapattion based on Markov Modeling of Speakers in Speaker-Independent Speech Recognition. *Proc. IEEE-ICASSP, Toronto* pp 845–848, 1991.

[7] Hiroaki Hattori, Satoshi Nakamura, Kiyohiro Shikano: Supplementation of HMM for Articulatory Variation in Speaker Adaptation. *Proc. ICSLP, Kobe* pp 153–156, 1990.

[8] Tatsuya Kawahara and Shuji Doshita: Speaker-Independent Consonant Recognition by Integrating Discriminant Analysis and HMM. *Proc. IEEE-ICASSP, Toronto* pp 557–560, 1991.

[9] Satoshi Nakamura and Kiyohiro Shikano: Speaker Adaptation Applied to HMM and Neural Networks. *Proc. IEEE-ICASSP, Glasgow* pp 89–92, 1989.

[10] Masafumi Nishimura and Kazuhide Sugawara: Speaker Adaptation Method for HMM-Based Speech Recognition. *Proc. IEEE-ICASSP, New York* pp 207–210, 1988.

[11] L.R. Rabiner, S.E. Levinson and M.M. Sondhi: On the Application of Vector Quantization and Hidden Markov Models to Speaker-Independent, Isolated Word Recognition. *The Bell System Technical Journal, Vol.* 62, *No.* 4 pp 1075–1105, 1983.

[12] Gerhard Rigoll: Speaker Adaptation for Large Vocabulary Speech Recognition Systems using 'Speaker Markov Models'. *Proc. IEEE-ICASSP, Glasgow* pp 05–08, 1989.

[13] Richard Schwartz, Yen-Lu Chow, Francis Kubala: Rapid Speaker Adaptation using a Probabilistic Spectral Mapping. *Proc. IEEE-ICASSP, Dallas* pp 33–36, 1987.