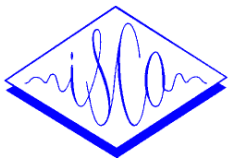

First International Conference on
Spoken Language Processing (ICSLP 90)
Kobe, Japan
November 18-22, 1990

SPEAKER ADAPTATION OF CONTINUOUS PARAMETER HMM

Yoshimitsu Hirata and Seiichi Nakagawa

Toyohashi University of Technology
Tempaku-cho, Toyohashi, 441, Japan

ABSTRACT

As an speaker adaptation method of continuous parameter HMM, we adapted mean vectors which are a part of parameters of multi-dimensional normal distributions. We regard a set of mean vectors belonging to each HMM as a codebook. The unsupervised adaptation algorithm modifies the mean vectors by using vector-quantization error-vectors for a test speaker. For two persons, 23 Japanese phoneme recognition accuracy was improved from 62% of non-adaptation into 73% after unsupervised adaptation and into 82% after supervised adaptation, respectively. Also, we describe adaptation results through multi-speaker mode HMM and comparison between the improvement by speaker adaptation and the degradation by vector-quantization of input vectors on speaker dependent mode HMM.

1. INTRODUCTION

HMM (hidden Markov model) has been recently superseding DTW (dynamic time warping) as a speech recognition technique. HMM approach is suitable for a speaker independent system, because it can deal with a variety of features for contexts and speakers as statistic models. Since it is difficult to remove differences among individual speakers, it is realistic to adapting an unknown speaker's speech into a standard speaker's one. So speaker adaptation techniques with small amount of training data are important for speaker independent speech recognition.

We modified the speaker adaptation algorithm proposed by Matsumoto et al. [1] as an adaptation technique of continuous parameter HMM. The adaptation algorithm is as follows: First, the feature vectors of a standard speaker are divided into some sub-spaces, that is, the code vectors in a codebook are grouped into some classes. Secondly, training data of a test speaker are vector-quantized by the standard speaker's codebook. Finally, using quantization error-vectors, the code vectors of the standard speaker are adapted into the test speaker's one. Chin-Hui Lee, et al. have reported another adaptation algorithm of continuous parameter HMM which assumes no correlation among each dimension of feature vectors [2].

In this paper, we describe adaptation results through single-speaker mode HMM and through multi-speaker mode HMM for 23 Japanese phonemes recognition. For comparison with the improvement by speaker adaptation, we compared the adapted vectors with vectors of quantized test data in speaker dependent mode.

2. UNSUPERVISED SPEAKER ADAPTATION ALGORITHM

In this section, we describe an unsupervised speaker adaptation algorithm of continuous parameter HMM in detail. Defining $\{X_i\}$ as a set of code vectors of a standard speaker's codebook and $\{Y_j\}$ as training data of an unknown speaker, the unsupervised speaker adaptation algorithm using vector-quantization error-vectors is formulated as follows.

<unsupervised speaker adaptation algorithm>

①A feature vector space of standard speaker is separated into M sub-spaces $\Omega_k (k=1,2,\cdots,M)$. Each space has a representative vector V_k.

② Training data of an unknown speaker are vector-quantized by a standard speaker's codebook, that is,

$$Y_j \rightarrow X_{i'}$$
$$i' = \underset{i}{arg\min}\{d(X_i, Y_j)\}$$

where $d(X,Y)$ denotes the Euclidian distance between X and Y.

③Average error vector Δ_k in each sub-space is computed by an equation

$$\Delta_k = \frac{1}{N_k} \sum_{j:Y_j, X_{i'} \in \Omega_k} (Y_j - X_{i'})$$

where N_k is a total number of training vectors that belong to k-th sub-space.

④ Standard speaker's codebook $\{X_i\}$ is adapted into test speaker's codebook $\{X_i'\}$ using Δ_k, that is,

$$X_i' = X_i + \sum_{k=1}^{M} W_{ik} \Delta_k$$

$$W_{ik} = \frac{d(X_i, V_k)^{-1}}{\sum_{k=1}^{M} d(X_i, V_k)^{-1}}$$

In the case of using multi-dimensional normal distributions as an output probability density of continuous parameter HMM, each distribution has the mean vector and covariance matrix. We consider a set of mean vectors belonging to every HMMs for all categories as a codebook, the number of mean vectors as the codebook size, and the number of category as a number of sub-space, respectively. A representative vector

is a mean vector of code vectors belonging each sub-space.

3. EXPERIMENTS

3.1 Speech Corpus

Two separate groups of ATR speech database [3] spoken by professional announcers were used. One is 5240 common Japanese words. Even-numbered words of them are for training HMM and odd-numbered words are for testing. The other set is 216 phoneme balancing words for only speaker adaptation usage. For single-speaker mode HMMs, male speaker (MAU) was regarded as a standard speaker. We also used multi-speaker mode HMMs made from five male speakers (M1 ~ M5). Test speakers were 2 male speakers: MHT and MNM. All of data were sampled at 10 kHz and pre-emphasized. The 14-th order linear prediction analysis with a 256 points Hamming window was performed every 3 ms frame shift. The features used were 10th order LPC mel-scaled cepstral coefficients and their dynamics (regression coefficients [4] over 45 ms). The actual phonetic tokens were extracted from the words. The 23 Japanese phonemes used for experiments are follows :

/b/, /d/, /g/ : voiced plosives
/p/, /t/, /k/ : unvoiced plosives
/m/, /n/, /N/ : nasals
/s/, /sh/, /h/, /z/ : fricatives
/ch/, /ts/ : affricates
/r/, /w/, /y/ : liquid and semivowels
/a/, /i/, /u/, /e/, /o/ : vowels

3.2 HMM Structure

Fig.1 illustrates the structure of HMM with continuous mixture density probabilities. When dynamic feature parameters were added in parameters, we assumed the independence between static parameters and dynamic parameters, because the dimension of combined feature parameters became too high in comparison with the amount of training samples.

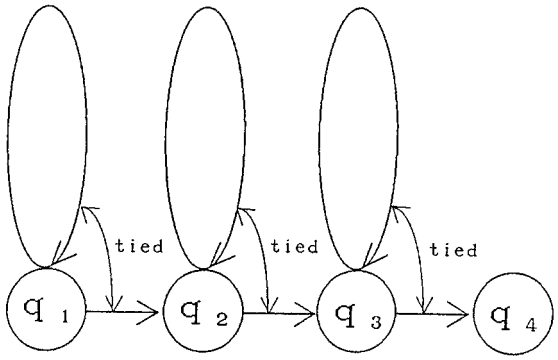

Fig.1 HMM with mixture continuous distributions

3.3 Adaptation of /b/,/d/,/g/

Using the model illustrated in Fig.1, we applied the unsupervised speaker adaptation algorithm described above to /b,d,g/ recognition task. We tried the adaptation from the standard speaker MAU into two unknown speakers MHT and MNM, respectively.

As adaptation training data, 10 or 50 tokens extracted from 5240 words were used for each phoneme. The supervised adaptation algorithm means the re-estimation of only mean vectors of distributions using Baum-Welch algorithm with three iteration. Tables 1 and 2 show recognition results.

From these tables, it found that the unsupervised adaptation algorithm is effective in using mixture distributions for the speaker (MHT) who got high recognition results without adaptation, and in using no-mixture distributions for the speaker (MNM) who didn't get enough recognition results without adaptation. From the fact there was a little difference between results of speaker dependent mode and these of supervised speaker adaptation mode with full training data, the covariance matrices can be used in common with no dependence upon speakers. This has been reported in a literature recently [2].

In this section, we described adaptations from a single speaker. Adaptations from multi speakers will be described in the later section 3.5.

Table 1 Phoneme recognition results for /b,d,g/ (no-mixture, codebook size : 9)

Method	Number of sub-space	Number of data	MHT (%)	MNM (%)	Aver. (%)
Non-adaptation	———	———	86.8	67.2	76.9
Unsupervised speaker adaptation	3	10	87.6	76.6	82.0
		50	86.1	77.1	81.5
Unsupervised speaker adaptation	9	10	81.3	81.0	81.1
		50	82.3	79.8	81.0
Supervised speaker adaptation	———	10	97.3	84.8	91.0
		50	98.1	88.7	93.3
		all	99.7	94.7	97.2
Speaker dependent	———	all	99.7	94.6	97.1

Table 2 Phoneme recognition results for /b,d,g/ (3-mixtures, codebook size : 27)

Method	Number of sub-space	Number of data	MHT (%)	MNM (%)	Aver. (%)
Non-adaptation	———	———	86.3	71.7	78.9
Unsupervised speaker adaptation	3	10	88.7	73.3	80.9
		50	87.6	74.3	80.9
Unsupervised speaker adaptation	9	10	90.5	71.6	80.9
		50	90.4	70.8	80.5
Supervised speaker adaptation	———	10	97.1	86.9	91.9
		50	97.8	91.0	94.4
		all	97.8	92.2	95.0
Speaker dependent	———	all	99.8	96.7	98.2

3.4 Adaptation of 23 Japanese Phonemes

For 23 Japanese phonemes, we adapted HMMs with no mixture distribution. Expanding categories of phonemes tests whether the adaptation technique is universal. The first 100 of 216 words were also used for adaptation. Table 3 shows average recognition results of 23 phonemes in using both static and dynamic features. Because there were little differences between recognition results of unsupervised speaker

adaptation with 23 sub-spaces and 69 sub-spaces, the table shows in the case of 23 sub-spaces only.

In the case of supervised adaptation, increasing the number of training tokens from 10 to 50 per each phoneme improved the recognition accuracy. On the other hand it was no influence on unsupervised adaptation. There were some degradations in using 100 words. The reasons for these are that the number of tokens among the phonemes was unbalance and the environment of utterances was different between 100 words and 5240 words. The performance of unsupervised speaker adaptation is around medium between non-adaptation and supervised adaptation.

Table 3 Phoneme recognition results for 23 Japanese phonemes
(no-mixture, codebook size : 69, sub-space : 23)

Training data	Method	MHT (%)	MNM (%)	Aver. (%)
Non-adaptation	————	66.0	58.8	62.4
10 tokens per phoneme	Unsupervised	75.3	72.4	73.8
	Supervised	83.5	82.1	82.8
50 tokens per phoneme	Unsupervised	75.1	72.4	73.8
	Supervised	87.1	85.2	86.2
100 words	Unsupervised	74.6	70.6	72.6
	Supervised	84.3	78.7	81.5
Speaker dependent	————	95.3	93.1	94.2

Another experiments within smaller categories divided 23 phonemes into 7 groups were done. As results, the speaker MHT got high recognition accuracy for /b,d,g/ and /s,sh,h,z/ (92.4% and 96.0%, respectively) even without speaker adaptation, and got no improvement after unsupervised speaker adaptation. While the speaker MNM got lower recognition accuracy for /b,d,g/ (72.8%) and /s,sh,h,z/ (93.9%) without adaptation, they were improved into 83.8% and 95.9% after unsupervised adaptation. For vowel groups (/a,i,u,e,o/) of the both speakers, the recognition result didn't rise by unsupervised adaptation because of high recognition accuracy without speaker adaptation.

3.5 Adaptation from Multi-Speaker Mode HMM

Multi-speaker mode HMM reflects variations of speakers and articulation. We made multi-speaker HMMs from five male speakers (M1～M5). Limiting number of tokens has reduced the variety of phoneme contexts comparing with full data described previously.

Table 4 shows 23 phoneme recognition results for the adaptation from multi-speaker HMM with parameters of dynamic features. Comparing with the results of single-speaker HMM (see Table 3), the average phoneme recognition rate without adaptation was improved around 17%, and the performance was equivalent to one of unsupervised adaptation with 10 or 50 training data per each phoneme. This implies an advantage of multi-speaker HMM in speaker independent mode.

Fig. 2 illustrates average phoneme recognition results vs. training data for unsupervised speaker adaptation from single-speaker mode HMM and multi-speaker mode HMM. The multi-speaker mode HMM was also superior to the single-speaker mode HMM in the case of unsupervised adaptation, but there was a little effect for the adaptation. On the other hand, the multi-speaker mode HMM didn't have the advantage in supervised adaptation in comparison with single-speaker mode HMM. The reason for this is considered as follows: As described in the section 3.3, covariance matrices can be used in common with no dependence on speakers if mean vectors are well adapted with enough training data. So even if a single speaker (i.e. MAU) is used as a standard, covariance matrices have little effects to recognition performance when mean vectors are suitable for an unknown speaker through supervised adaptation. (The main difference between the single-speaker mode HMM and multi-speaker mode HMM is covariance matrices.) On the other hand, when the adaptation of mean vectors is not perfect such as unsupervised adaptation, multi-speaker mode HMM which has the advantage of more precise covariance matrices becomes effective.

Table 4 Adaptation from multi-speaker mode HMM
(no-mixture, codebook size : 69, sub-space : 23)

Training data	Method	MHT (%)	MNM (%)	Aver. (%)
Non-adaptation	————	75.9	81.9	78.9
10 tokens per phoneme	Unsupervised	81.7	78.2	80.0
	Supervised	85.2	82.6	83.9
50 tokens per phoneme	Unsupervised	78.7	81.6	80.2
	Supervised	85.5	87.3	86.4
100 words	Unsupervised	81.6	76.1	78.9
	Supervised	83.8	79.1	81.5
Speaker dependent	————	95.3	93.1	94.2

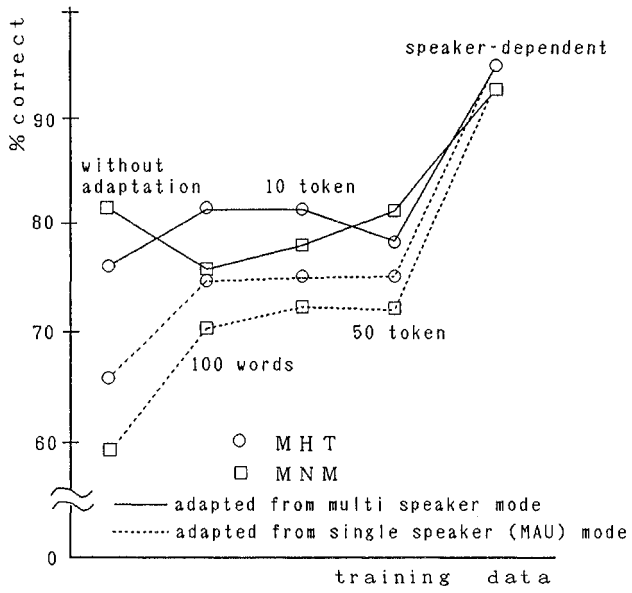

Fig. 2 23 Japanese phoneme recognition results after unsupervised speaker adaptation
(no-mixture, codebook size : 69, CEP+ΔCEP)

3.6 Vector Quantization of Input Speech Vectors

Test data was quantized by a standard speaker's codebook to show which level of vector quantization the accuracy of the spectrums estimated by speaker adaptation corresponds to.

The codebooks were made from the 100 words, and the codebooks of cepstrums and their dynamics were made separately.

Table 5 shows average phoneme recognition results among 18 consonants and 5 vowels, respectively. The recognition was done in speaker dependent mode and feature vectors with dynamic features of the mel-warped cepstral coefficients. As the table shows, the vector-quantization of input speech led to the degradation of recognition accuracy, especially for consonants. To compare the improvement by speaker adaptation using 100 word training data (see Table 3) with the degradation by VQ, recognition results among consonants and vowels were separately summarized as Table 6. From Tables 5 and 6, it found that the spectral accuracy of both consonants and vowels were inferior to that of 64 level VQ in recognition accuracy without the adaptation. But the accuracy with unsupervised adaptation were equivalent to that of 64 level VQ. By supervised adaptation, the recognition accuracy of consonants rose furthermore as well as 256 level VQ. As above, the estimated spectra by the adaptation technique described in this paper are equivalent to them obtained by 64 ~ 256 level vector quantization of input data in speaker dependent mode.

Table 5 Average phoneme recognition results by VQ of input speech (Speaker : MHT)

	Quantization level					
	64	128	256	512	1024	∞
Consonants	70.1%	75.4%	79.7%	84.9%	87.3%	94.0%
Vowels	93.2%	96.7%	97.3%	98.2%	98.3%	98.7%
Average	76.6%	81.4%	84.7%	88.7%	90.4%	95.3%

Table 6 Average phoneme recognition results by speaker adaptation (Speaker : MHT)

Method	Consonants	Vowel	Average
Non-adaptation	57.4%	87.7%	66.0%
Unsupervised speaker adaptation	67.0%	93.7%	74.6%
Supervised speaker adaptation	79.5%	96.5%	84.3%

4. CONCLUSIONS

We recognized 23 Japanese phonemes to evaluate the speaker adaptation technique on continuous parameter HMM which adapted only mean vectors of multi-dimensional normal distributions. There were 10% improvement in unsupervised adaptation and 20% in supervised adaptation as the results of experiments that used 100 words for adaptation using both static and dynamic feature vectors. This implies that the speaker adaptation of mean vectors is effective on continuous parameter HMM without dealing with covariance matrices.

Using very same training and testing database, Nakamura et al. applied a speaker adaptation algorithm to discrete parameter HMM based on fuzzy vector quantization [5]. They used HMMs with multiple-codebooks for static features, dynamic features and power. Our scheme of speaker adaptation didn't deal with the power. In the case of their discrete HMM, the standard speaker's codebook was mapped into an unknown speaker's one. So the comparison between them is not rigorous. Table 7 shows the comparison of the recognition results obtained by the continuous parameter HMM and the discrete parameter HMM (the adaptation from speaker MAU to MNM by training data of 100 words). As the table shows, the continuous HMM is superior to the discrete HMM in all cases without adaptation, with adaptation and in speaker dependent mode.

Multi-speaker mode HMM was effective on recognition experiments of speaker independent mode (improvement of 10 ~ 20% than using single-speaker mode HMM). Also on unsupervised speaker adaptation, the multi-speaker mode HMM performed a little less than 5% improvement of recognition accuracy as compared with the single-speaker mode HMM. But no advantage was given by multi-speaker mode HMM on supervised speaker adaptation in comparison with the single-speaker mode HMM.

Table 7 23 Japanese phoneme recognition results

Method	Discrete HMM[5]	Continuous HMM
Without adaptation	62.1%	66.0%
Unsupervised adaptation	———	74.6%
Supervised adaptation	75.6%	84.3%
Speaker dependent	92.7%	95.3%

ACKNOWLEDGEMENTS

Authors would like to thank to Dr. Shigeki Sagayama and the parties concerned of ATR Interpreting Telephony Research Laboratories for giving us convenience of using computers and the speech database.

REFERENCES

[1]H.Matsumoto and Y.Yamashita : "Unsupervised Speaker Adaptation Based on Piecewise Average Errors in Vector Quantization", Trans. EIC, Vol.72A, No.5, pp869-872 (1989, in Japanese).

[2]Chin-Hui Lee, et al. : "A Study on Speaker Adaptation of Continuous Density HMM Parameters", Proc. ICASSP-90, pp.145-148 (1990).

[3]A.Waibel, et al. : "Phoneme Recognition Using Time-Delay Neural Networks", IEEE Trans. ASSP-37, No.3, pp.328-339 (1989).

[4]S.Furui : "Speaker-Independent Isolated Word Recognition using Dynamic Features of Speech Spectrum", IEEE Trans. ASSP-34, No.1, pp.52-59 (1986).

[5]S.Nakamura and K.Shikano : "Speaker Adaptation Applied to HMM and Neural Networks", Proc. ICASSP-89, pp.89-92 (1989).