論文 / 著書情報
Article / Book Information

Title	Speaker adaptation for demi-syllable based speech recognition using continuous HMM,
Author	K. Shinoda, K. Iso, T. Watanabe
Journal/Book name	Proc. of ICSLP-90, Vol. , No. , pp. 261-264
発行日 / Issue date	1990,

SPEAKER ADAPTATION FOR DEMI-SYLLABLE BASED SPEECH RECOGNITION USING CONTINUOUS HMM

Koichi Shinoda, Ken-ichi Iso and Takao Watanabe
C & C Information Technology Research Laboratories
NEC Corporation
4-1-1 Miyazaki, Miyamae-ku, Kawasaki 213, JAPAN

ABSTRACT

A novel speaker adaptation method, which is applied to the demi-syllable based speech recognition system using continuous density HMM, is proposed. In this method, mean vectors of HMM Gaussian pdfs for a standard speaker are adapted to those for a new speaker with a small amount of training data. Supervised speaker adaptation is first employed, and for the recognition units which are not adapted in the supervised adaptation, unsupervised speaker adaptation is performed.

The effectiveness of the proposed method was confirmed by large vocabulary word recognition experiments. Using 50 word utterances for speaker adaptation, the recognition rates were improved by 14.4 %, on an average.

1. INTRODUCTION

This paper presents a novel speaker adaptation method for speech recognition using continuous density Hidden Markov Model(HMM). HMM has been successfully used for large vocabulary speech recognition. It has the learning algorithm, called the Baum-Welch algorithm, and shows good recognition performance, when a sufficient amount of training data is available. However, it shows poor performance when the amount of training data is limited.

Recently, a demi-syllable based speech recognition system, using continuous density HMM, has been developed[1]. Using demi-syllables as recognition units, it can efficiently treat contextual variations caused by the co-articulation effect. In this system, a single Gaussian probability density function (pdf) is used as HMM output probability. With 250 word training data, it achieved high recognition performance. For practical use, however, the amount of training data is expected to be further largely reduced.

In order to reduce the amount of training data required, various kinds of speaker adaptation methods have been explored [2, 3, 4, 5, 6, 7]. They are classified into two categories. One is supervised adaptation, in which training words or sentences are known. The other is unsupervised adaptation, in which arbitrary utterances can be used. It was reported that, when the amount of training data was large and most recognition units have their training samples in the data, supervised adaptation outperformed unsupervised adaptation[7]. On the other hand, when the amount of training data is small and a lot of recognition units do not have their training samples, unsupervised adaptation seems to be more effective than supervised adaptation.

This paper proposes a novel speaker adaptation method, in which supervised speaker adaptation and unsupervised speaker adaptation are combined. It has the merits of both supervised and unsupervised adaptation; it shows as good performance as supervised adaptation when a sufficient amount of training data is available, and shows better performance than supervised adaptation when the amount of data is small. This method was applied to the demi-syllable based speech recognition system, and its effectiveness was experimentally verified.

The paper is organized as follows. In Section 2, the new speaker adaptation method is presented. In Section 3, several experimental results are described.

2. ADAPTATION METHOD

In the recognition system using demi-syllables as recognition units, each word of training data is presented by a concatenation of demi-syllable HMMs, which consists of several states. The case is considered in which the amount of training data is so small that some demi-syllable HMMs have no training samples in training data. HMM states were classified into two groups, Group A and Group B. Group A contains the states which have their training samples in training data. Group B contains the states which do not have their training samples in the data.

For the states of Group A, which have training samples in training data, supervised adaptation was accomplished. One

way to adapt these states would be to employ the Baum-Welch training algorithm. However, it was reported that training with an insufficient amount of data often decreased recognition accuracy to even less than that of speaker independent training[6]. It was also suggested that a training method, which changed only mean vectors and did not change the other two parameters, variances and transition probabilities, had better recognition performance than the training using Baum-Welch algorithm, especially when only a small amount of data was available[6].

Therefore, for supervised adaptation, the following two kinds of algorithm were employed. One is restricted Baum-Welch algorithm, which is different from the original Baum-Welch algorithm, in that it only changes mean vectors, and does not change the other two parameters. The other algorithm is Viterbi learning algorithm. In this algorithm, the training data segmentation is carried out by the Viterbi algorithm, and back-tracking is implemented. Then, a mean vector for each HMM state is replaced with the average of feature vectors, which are time-aligned with the state.

Thus, the mean vector of a state $j \in A$ of a standard speaker, μ_j^A, is replaced by the mean vector of a new speaker, $\hat{\mu}_j^A$, which is calculated by the supervised adaptation method described above.

In other supervised methods, such as reported by Shikano *et al.*[2], time alignment between utterances of a standard speaker and that of a new speaker is carried out. In these methods, standard speaker's utterances are needed for speaker adaptation. On the contrary, the proposed method does not need standard speaker's utterances. Therefore, in the proposed method, the training data vocabulary can be easily changed.

Following the supervised adaptation described above, an unsupervised adaptation method is employed for the Group B states. An adaptation vector for state i is defined as Δ_i, which is the difference between a mean vector after adaptation, $\hat{\mu}_i$ and that before adaptation, μ_i. To adapt mean vectors for the Group B states, interpolation among adaptation vectors for the Group A states was carried out. For the interpolation, distances between mean vectors were used. The adaptation vector for the state $i \in B$ is determined so as to be aligned with the adaptation vectors for the Group A states, whose mean vectors are in the vicinity of the mean vector for state i. Figure 1 illustrates the unsupervised adaptation principle.

The unsupervised adaptation algorithm is given in the following.

1. For a state $j \in A$, the mean vector after adaptation, $\hat{\mu}_j^A$, has already been obtained. An adaptation vector, Δ_j^A, is given as,

$$\Delta_j^A = \hat{\mu}_j^A - \mu_j^A, \tag{1}$$

where, A denotes that the state j belongs to Group A.

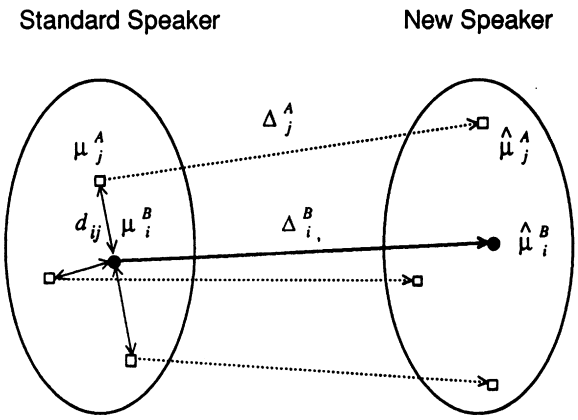

Figure 1: Speaker adaptation

The adaptation vector Δ_j^A is computed for every state of Group A.

2. For a state $i \in B$, an adaptation vector, Δ_i^B, is determined by interpolation among the adaptation vectors for the states $j \in A$,

$$\Delta_i^B = \sum_j w_{ij} \Delta_j^A. \tag{2}$$

Weight w_{ij} is defined as a function of the distance d_{ij} between μ_i^B and μ_j^A,

$$d_{ij} = \| \mu_i^B - \mu_j^A \|, \tag{3}$$

$$w_{ij} = \frac{d_{ij}^{-m}}{\sum_{j'} d_{ij'}^{-m}}, \tag{4}$$

where m is the parameter which determines the w_{ij} dependence on distance d_{ij}. Adaptation vector Δ_i^B is computed for every Group B state.

3. A mean vector for the state i for the new speaker, $\hat{\mu}_i^B$, is given as,

$$\hat{\mu}_i^B = \mu_i^B + \Delta_i^B, \tag{5}$$

where μ_i^B is the mean vector for the standard speaker's HMM.

3. EXPERIMENTS

3.1. Experimental Conditions

In the experiments, two 250 word utterance data sets were used; one was for speaker adaptation, and the other was for recognition experiments. Each data set consisted of 250 phonetically balanced words. The vocabularies in the two data sets differed from each other. These data sets were uttered by three male speakers.

Table 1: Number of states in Group A, S_A, for each word number, W

W	1	10	50	100	250	∞
S_A	42	268	613	701	771	919
S_A/S_T (%)	4.6	29.2	66.7	76.3	83.9	100.0

For a recognition dictionary, a 5000 word set was prepared. In recognition experiments, the following method was employed to reduce computational amount. First, a similar word set was prepared for each word in the data set for recognition experiments. The similar word set, which consisted of 100 words, was selected from 5000 words, using manually designed phoneme symbol confusion matrix. Then, instead of using all the 5000 words, these similar sets were employed for recognition dictionaries. The computational amount was reduced by a factor of 50. It was confirmed that the recognition results, using the similar word sets, were almost the same as the results, using all the 5000 words for the dictionary[8].

The utterances were digitized at a 16 kHz sampling rate, and analyzed by a 10 msec frame period. The feature used is a vector of 21 components, consisting of a power derivative, 10 mel-scaled cepstral coefficients, and 10 corresponding mel-scaled cepstral time derivatives.

There were 241 demi-syllables. Most demi-syllable HMMs have four states, excepting that those for long vowels and silences have one state. The ratio of the number of states which have training samples in the training data(S_A), to that of total states(S_T), with various training word set sizes, is given in Table 1. With 10 words, 30 % of all the states appeared. With 50 words, 67 % appeared.

In order to obtain optimal conditions for the speaker adaptation method, several experiments were carried out. In these experiments, speaker A was chosen as a standard speaker, and speaker B was chosen as a new speaker. The training data amount was 50 words. The changed conditions were as follows;

- The number of iterations in the supervised adaptation
- The algorithm for the supervised adaptation
- The value of m in (4), which determines the w_{ij} dependence on distance d_{ij}.

The changes in recognition rates, due to the changes in these conditions, were quite small. The number of iterations was set to 1, the Viterbi learning algorithm was chosen, and the m value was set to be 1.0.

3.2. Comparison With Other Methods

The proposed speaker adaptation method was compared with the other two methods. One was speaker dependent training, using the original Baum-Welch algorithm, in which the training data was identical with that for speaker adaptation(Baum-Welch). In this experiment, HMM parameters for the standard speaker A were used as the initial HMM parameters. The other was the method which used only the supervised adaptation for the Group A states, and did not employ unsupervised adaptation for Group B states (Supervised adaptation). Speaker A was chosen as a standard speaker, and speaker B was chosen as a new speaker. The recognition results are shown is in Table 2, in which the training data amounts were 10 words and 50 words.

Table 2: Adaptation methods comparison

	Word Number	
	10	50
Baum-Welch	58.4 %	66.8 %
Supervised adaptation	62.4 %	73.6 %
Proposed Method	71.6 %	77.2 %

With only the supervised adaptation method, the recognition rate was 4.0 % higher than that for the Baum-Welch training, when the training data consisted of 10 words. It was proved that the supervised adaptation was more effective than the Baum-Welch training, when the data amount was small. In the Baum-Welch training, the estimation on transition probabilities and variance with an insufficient amount of training data seems to degrade the recognition performance. Furthermore, the recognition rate was improved as much as 9.2 %, when unsupervised adaptation was introduced. This result indicates that unsupervised adaptation is especially effective when the training data amount is small, that is, when the number of HMM states for Group B is much larger than the number of HMM states for Group A.

In addition to these experimental results, it would be profitable to refer to the recognition result of a speaker dependent training experiment with a sufficient amount of data, and that of a cross speaker experiment. In the speaker dependent training experiment, the original Baum-Welch algorithm was employed, and the data amount was 250 words. In this experiment, HMM parameters for the standard speaker A were used as the initial HMM parameters. In the cross speaker experiment, HMM parameters for the standard speaker A were used directly in the recognition experiment. The recognition rate for the speaker dependent training experiment was 84.4 %, and that for the cross speaker experiment was 66.0 %.

3.3. Training Data Amount Evaluation

The word recognition rates for one new speaker(B) with the data amount changed, is given in Figure 2. In this experiment, the standard speaker was A. In Figure 2, the ratio of the number of the states in Group A (S_A) to the number of total states(S_T) is also given. The recognition rate increased

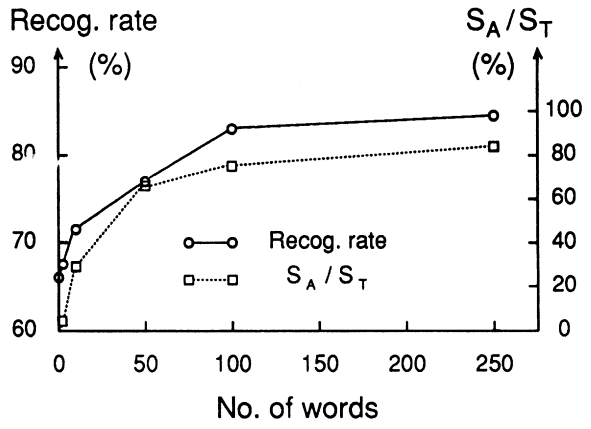

Figure 2: Recognition results vs. Number of words used for speaker adaptation

as the ratio S_A/S_T increased. The recognition rate reached as high as that of speaker dependent training, when the data amount was 100 words.

3.4. Speaker Difference

The word recognition rates for three speakers, A, B and C are given in Table 3. In each experiment, one speaker was a standard speaker, while the other two were new speakers. In Table 3, the recognition rates for cross-speaker experiments, and those for the experiments, in which speaker dependent training using Baum-Welch algorithm was employed, are also shown. Fifty words were used for speaker adaptation. In speaker dependent training by the Baum-Welch algorithm, 250 training words were used. Using the speaker adaptation method, the recognition rates were improved by 14.4 %, on an average.

Table 3: Recognition Performance for three speakers

Standard speaker	A		B		C	
Test speaker	B	C	A	C	A	B
Cross speaker	66.0	63.2	74.0	56.8	50.0	51.2
Speaker adaptation	77.2	75.2	79.2	74.0	76.4	72.8
Speaker dependent	84.4	86.8	89.6	86.8	89.6	84.4

4. CONCLUSION

A new speaker adaptation method for demi-syllable based speech recognition, using continuous density HMM, was proposed. A supervised adaptation method and an unsupervised adaptation method are combined, and demi-syllables, which do not appear in training data, are also adapted to a new speaker. The proposed method effectiveness was examined by 5000 word recognition experiments. The recognition rate was considerably improved, using a small amount of training data.

ACKNOWLEDGMENTS

The authors wish to thank members of the Media Technology Research Laboratory for their continuous support.

REFERENCES

[1] K.Yoshida, T.Watanabe, and S.Koga, "Large Vocabulary Word Recognition Based on Demi-Syllable Hidden Markov Model Using Small Amount of Training Data," *Proc. ICASSP-89*, pp.1-4, Glasgow, 1989.

[2] K.Shikano, K.-F.Lee, and R.Reddy, "Speaker Adaptation Through Vector Quantization," *Proc. ICASSP-86*, pp.2643-2646, Tokyo, 1986.

[3] R.M.Schwartz, Y.L.Chow, and F.Kubala, "Rapid Speaker Adaptation Using a Probabilistic Spectral Mapping," *Proc. ICASSP-87*, pp.633-636, Dallas, 1987.

[4] S.Furui, "Unsupervised Speaker Adaptation Method Based on Hierarchical Spectral Clustering," *Proc. ICASSP-89*, pp.286-289, Glasgow, 1989.

[5] H.Matsumoto, and Y.Nakato, "A Study on A Sequential Speaker Adaptation Method for Speech Recognition," *Research Report of PASL,* October 1989(in Japanese).

[6] C.-H.Lee, C.-H.Lin, and B.-H.Juang, "A Study on Speaker Adaptation of Continuous Density HMM parameters," *Proc. ICASSP-90*, pp.145-148, Albuquerque, 1990.

[7] Y.Hirata and S.Nakagawa, "A Study of Speaker Adaptation of Continuous Parameter HMM on Japanese Phoneme Recognition," *Committee of Speech Research, Acoust.Soc.Japan*, SP90-16, 1990(in Japanese).

[8] S.Koga, K.Yoshida, and T.Watanabe, "Evaluation of Large Vocabulary Speech Recognition Based on Demi-Syllable HMM," *Proc. of ASJ Autumn Meeting,* October 1989(in Japanese).