論文 / 著書情報
Article / Book Information

Title	Unsupervised speaker adaptation for speech recognition using demi-syllable HMM
Author	K. Shinoda, T. Watanabe
Journal/Book name	Proc. ICSLP-94, Vol. , No. , pp. 435-438
発行日 / Issue date	1994,

UNSUPERVISED SPEAKER ADAPTATION FOR SPEECH RECOGNITION USING DEMI-SYLLABLE HMM

Koichi Shinoda and Takao Watanabe
Information Technology Research Laboratories
NEC Corporation
4-1-1 Miyazaki, Miyamae-ku, Kawasaki-shi 216, JAPAN

ABSTRACT

An unsupervised speaker adaptation method is proposed for application to a speaker independent recognition system which uses a demi-syllable based, continuous mixture-density HMM. The spectral interpolation technique, which has been used in supervised adaptation, and a scheme of utilizing recognition outputs of recognition system are employed. In an experimental application employing results obtained from a 5000-word Japanese vocabulary set recognition task, the error rate was reduced with the method from 15.5 % to 8.7 %, which is comparable to the error rate obtained by supervised adaptation. The method has also shown promise in the utilization of results obtained from a connected syllable recognition task in which the Japanese vocabulary set was unlimited.

1. INTRODUCTION

Remarkable progress has been achieved recently in the development of speaker-independent speech recognition systems which utilize Hidden Markov Models (HMMs) and in which the training data consists of the utterances of a large number of speakers. While such systems have the advantage of requiring no utterances from new users for the training of HMM parameters, and while their performance has shown some promise, recognition rates are still significantly lower than those of speaker-dependent systems, and rates may be extremely low for certain individual speakers.

To address this problem, the speaker adaptation technique using spectral interpolation has been recently developed[1, 2]. It is one method of supervised adaptation, and effective not only for recognition units for which there are training samples available, but also for recognition units for which there are no training samples, since the parameters for those units are estimated by an interpolation technique. Tests have indicated a high performance even when only a small amount of training data is available. For user-oriented applications, however, unsupervised adaptation is ordinary more preferable than supervised adaptation, because it does not require a special training session. In this paper, we propose an unsupervised adaptation method using the spectral interpolation technique. For this method, the scheme of utilizing recognition outputs of a recognition system, which has been used in several studies(e.g. [3, 4, 5, 6]), are employed.

We considered two different situations in which unsupervised adaptation might be most likely to be used; that in which a limited vocabulary set for recognition is used for adaptation, and that in which there is no such vocabulary limit. In the second case, connected-syllable recognition is employed as a recognition task. In this study, we have evaluated our proposed speaker adaptation method for both situations. We have designed it to be used with a demi-syllable based speech recognition system[8]. Demi-syllable systems can deal efficiently with the contextual variation caused by the coarticulation effect, can achieve high recognition rates with large vocabulary tasks.

Our paper is organized as follows: in Section 2, we briefly describe the demi-syllable based speech recognition system; in Section 3, we discuss the use of spectral interpolation in adaptation; in Section 4 we describe our proposed unsupervised adaptation method; and in Section 5 we report the results of experiments.

2. DEMI-SYLLABLE HMM

Demi-syllables are units which are made by dividing each syllable at its steady part of the vowel. Demi-syllables are helpful in dealing with the variations in phonetic characteristics caused by the co-articulation effect since they contain information regarding the transitional portions of syllables, information very important to phoneme recognition. The number of demi-syllables required for the recognition of Japanese speech is 241. In our modeling, we used a single left-to-right HMM for each demi-syllable. All of our demi-syllable HMMs have four states, with the exception of o those for long vowels and silences, which have one state. We represent the observation distribution in each state with a multivariate Gaussian mixture density, and use a diagonal covariance matrix for each mixture component.

3. SPECTRAL INTERPOLATION ADAPTATION METHOD

A continuous density HMM, in which a Gaussian mixture distribution is used as the observation distribution, has three kinds of parameters to be adapted: mean vectors, variances, and transition probabilities. Some previous studies have reported that when the amount of training data was small, variances and transition probabilities were wrongly estimated (e.g., [9]). With this in mind, we have chosen to limit our adaptation to mean vectors alone. For the discussion which follows, we first consider the case in which a

single Gaussian density function is used in each state. Then, we proceed to consider a Gaussian mixture density function which has more than one mixture component.

The adaptation procedure is divided into two stages. In the first stage, adaptation of recognition units which appear among the training data (referred to here as Group A) is carried out. Training data for a new speaker are segmented by using Viterbi Algorithm. Speaker-independent HMMs are employed for this segmentation. The mean vector for each HMM state is replaced with the average of those feature vectors which have been time-aligned to the state.

In the second stage, interpolation is carried out for those units not appearing among the training data (referred to as Group B). First, 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 Group B states, interpolation among adaptation vectors for Group A states is carried out. Figure 1 illustrates the interpolation principle. The algorithm of the interpolation procedure may be stated as follows:

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 state j belongs to Group A. The adaptation vector Δ_j^A is computed for every state in 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} for Δ_j^A is defined as a function of the distance d_{ij} between μ_i^B and μ_j^A. For example, w_{ij} is defined as follows:

$$w_{ij} = \frac{d_{ij}^{-m}}{\sum_{j'} d_{ij'}^{-m}}, \tag{3}$$

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

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

$$\hat{\mu}_i^B = \mu_i^B + \Delta_i^B, \tag{4}$$

where μ_i^B is the mean vector for the speaker-independent HMM.

4. This 2-3 procedure is implemented for all the Group B states.

By modifying the the above procedure, one can apply the proposed method to HMMs whose output probability is modeled by a multivariate Gaussian mixture distribution. In the first stage of such a case, Viterbi segmentation is further carried out so as to select the mixture component which has the highest probability for the feature vector in each state. The components of the mixture distribution are now placed into Groups A and B. The second stage of adaptation is almost the same as the one for a single Gaussian distribution, except that interpolation is carried out for the mean vector of each mixture component. If the number of mixture components in each state becomes large, there would be a lot of mixture components which appear only few times among the training samples. Therefore, instead of the weight w_{ij} in (3), we use a weight w' which is defined as $w'_{ij} = n_j \cdot w_{ij} / \sum n_j$, where i and j are the indices for mixture components, and n_j was the number of training samples used for estimating $\hat{\mu}_j^A$. If the appearance number of mixture component j is large, the contribution of μ_j^A to estimating $\hat{\mu}_i^B$ becomes also large.

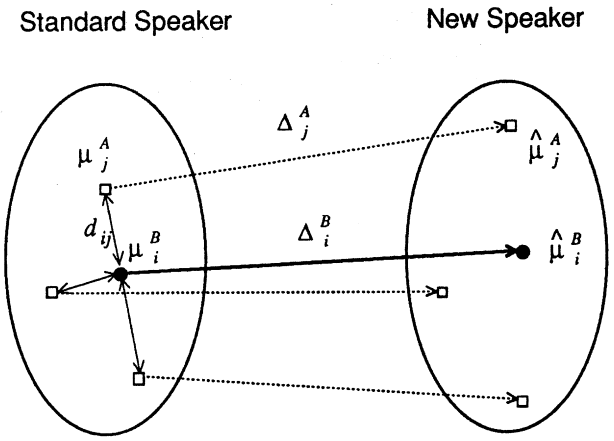

Figure 1: Interpolation principle

4. UNSUPERVISED ADAPTATION SCHEME

In our unsupervised adaptation scheme, outputs of recognition system is utilized as supervising signals. The outputs are a recognized text, its likelihoods, and likelihoods of the other texts. Among them, we use a recognized text as a supervising signal for adaptation.

This scheme has one problem that there is a limit to the size of vocabulary users can utter. However, we considered two different situations in which the limit of vocabulary does not degrade the comfortability of users: the one in which a limited vocabulary set for recognition is used for adaptation, and the other in which there is substantially no such vocabulary limit. In the first situation, recognition rates may be expected to be high, and so, consequently, may be the effectiveness of adaptation, but the rejection of a large number of out-of-vocabulary utterances might also be expected to be a problem. In the second situation, the recognition task that accepts any utterances is prepared. As such task, we employ connected-syllable recognition, in which finite state automata is employed as a grammar for recognition(Figure 2). Although the recognition rates are expected to be low,

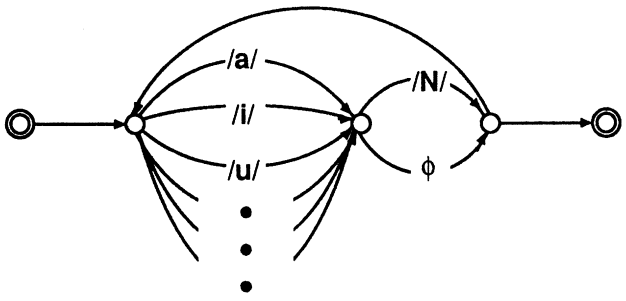

Figure 2: Japanese connected-syllable recognition

the recognition results are expected to be useful when they are used as supervised signals for unsupervised adaptation.

5. EXPERIMENTS

5.1. Experimental Conditions

Experiments were carried out under the demi-syllable based recognition system in which continuous density HMMs were used. The number of components of the mixture distribution was two for all HMM states. The effectiveness of the proposed unsupervised speaker adaptation method was evaluated in 5000 words recognition.

Three data sets were used for these experiments. One data set (DB1) was a multi-speaker database which consisted of 46 male speakers and 39 female speakers. The next data set (DB2) consisted of 3 male speakers (M1, M2, M3) and 4 female speakers (F1, F2, F3, F4). These 7 speakers were not included in the multi-speaker database, DB1. The last data set (DB3) consisted of the same 7 speakers. For each data set, a vocabulary of phonetically-balanced 250 words was selected from a Japanese lexicon. The vocabulary used in each data set was different from each other. In each data set, each word was uttered once by each speaker. For a recognition dictionary, a vocabulary of 5000 words was selected from a Japanese lexicon.

In recognition experiments, the following method was employed to reduce the required computational amount. First, a phoneme symbol distance matrix which assigned the distances between phonemes was manually defined. Next, one calculated respective distances from each word for recognition to the other words in the recognition dictionary, by accumulating the phoneme distances using the distance matrix [10]. Then, 100 similar words were selected for each word. Finally, for each word for recognition, the recognition experiments were carried out using these 100 similar words as a recognition dictionary, instead of using all the 5000 words. It was confirmed that the recognition rate when these similar words were used for a recognition dictionary was almost the same as the recognition rate achieved when all the 5000 words were used [10]. In the speaker adaptation, The parameter m in (3) was 1.0, and the squared distance weighted by variance was used as d.

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

The ratio of the number of HMM states which appeared among the DB2 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 in the training data. With 50 words, 67% appeared.

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

In all experiments, a speaker-independent model, which was trained using the multi-speaker database DB1, was used as the initial model for adaptation.

5.2. Evaluation under 5000 word recognition

First, the effectiveness of the proposed method was evaluated when 5000 word recognition was used for adaptation. This experiment corresponds to the first case in Section 4. The word set for the adaptation was chosen from the DB2 database. Table 2 shows the results of 5000-word recognition experiments when the amounts of training data for SA were 50, 100, 150 and 250 words. Table 2 also shows the results of the other two experiments, speaker-independent recognition (SI) and speaker-dependent recognition (SD). In the SD experiments, Baum-Welch Algorithm was employed and whole 250 words were used for training. The recognition rates were increased when unsupervised adaptation was used. The rates were less than those of supervised adaptation, but the difference between them was very small. It was confirmed that the proposed unsupervised adaptation was effective when large-vocabulary recognition was used for the adaptation.

Next, Japanese connected-syllable recognition was used for the adaptation. The number of syllable used was 101, which was sufficient for representing all Japanese utterances. Frame synchronous search algorithm using the Bundle Search[11] were employed for recognition. The recognition accuracies were improved when 250 words were used for adaptation.

Although the recognition rates of connected-syllable recognition was low, the unsupervised adaptation using recognition results had certain effect. This fact can be explained as below. It can be assumed that the similarity between syllables in the acoustical feature vector space will be reserved before and after adaptation. The effectiveness of the supervised adaptation using spectral interpolation confirms this hypothesis. In the connected-syllable recognition, the misrecognized syllable is acoustically similar to the correct one. From the hypothesis above, it is considered that corresponding adaptations vectors of these two syllables are also similar. As a result, the adaptation is successful even

Table 2: Recognition result when the 5000 words recognition was used for adaptation(%)

	M1	M2	M3	F1	F2	F3	F4	ave.
SI	78.8	89.6	88.4	79.1	86.4	82.8	86.7	84.5
SD	86.0	94.0	92.4	90.8	88.8	90.4	90.0	90.3
50 words	84.0	86.0	90.0	81.5	84.4	79.5	85.2	84.4
	(82.4)	(90.4)	(90.0)	(83.9)	(84.4)	(83.9)	(84.8)	(85.7)
100 words	82.8	91.2	90.8	83.1	88.0	83.1	83.6	86.1
	(87.6)	(90.0)	(90.4)	(86.3)	(87.6)	(87.1)	(86.8)	(88.0)
150 words	87.6	91.6	92.8	87.1	88.8	85.1	88.4	88.8
	(92.4)	(92.8)	(92.4)	(89.2)	(89.2)	(90.0)	(90.0)	(90.9)
250 words	90.8	94.4	94.4	89.2	90.0	89.2	90.8	91.3
	(92.0)	(95.2)	(94.0)	(92.0)	(90.4)	(93.6)	(91.2)	(92.6)

The values in are the results of the supervised adaptation.

Table 3: Recognition result when the connected-syllable recognition was used for adaptation(%)

	M1	M2	M3	F1	F2	F3	F4	ave.
50 words	77.2	87.6	85.2	77.9	83.2	79.5	77.6	81.2
100 words	78.4	83.6	86.4	79.1	82.8	79.5	76.8	80.9
150 words	78.8	87.6	87.2	79.5	85.6	81.9	80.8	83.1
250 words	81.6	90.4	90.0	80.7	91.2	87.1	86.0	86.7

though the recognition rate of connected-syllable recognition is not so high. In this sense, this method is partly related to the unsupervised adaptation method in which the similarities among acoustical features are only used for adaptation[12].

6. CONCLUSION

An unsupervised speaker adaptation method for continuous density HMM has been proposed. The effectiveness of the proposed method was confirmed by 5000-word recognition experiments. It has as high performance as supervised adaptation. The adaptation using connected-syllable recognition, in which any utterance of a speaker is accepted, was also proved to be effective.

ACKNOWLEDGMENTS

The authors wish to thank members of the Human Language Research Laboratory for their continuous support.

REFERENCES

[1] K.Shinoda, K.Iso, and T.Watanabe, "Speaker Adaptation for Demi-Syllable Based Speech Recognition Using Continuous HMM," *Proc. of ICSLP-90,* pp.261-264, Kobe, 1990.

[2] K.Shinoda, K.Iso, and T.Watanabe, "Speaker Adaptation for Demi-Syllable Based Continuous Density HMM," *Proc. of ICASSP-91,* pp.857-860, Toronto, 1991.

[3] Y.Zhao, "Self-Learning Speaker Adaptation Based On Spectral Validation Source Decomposition," *Proc. of EuroSpeech93,* pp.359-362, Berlin, 1993.

[4] Y.Tsurumi and S.Nakagawa, "Unsupervised Speaker Adaptation for Continuous Parameter HMM by Maximum A Posteriori Probability Estimation," *Technical Report of IEICE,* SP93-104, pp.1-8, 1993.

[5] T.Matsuoka and C.H.Lee, "On-Line Speaker Adaptation Using Maximum A Posteriori Estimation," *Technical Report of IEICE,* SP93-133, pp.39-46, 1994.

[6] Y.Miyazawa, J.Takami, S.Sagayama, and S. Matsunaga, "All-Phoneme Ergodic Hidden Markov Network For Unsupervised Speaker Adaptation," *Proc of ICASSP94,* pp.249-252, Adelaide, 1994.

[7] 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.

[8] R.Isotani, K.Hatazaki, T.Watanabe, H.Ohtsubo, and M.Mizuno, "Speaker-Independent Speech Recognition Based on Demi-syllable Hidden Markov Models," *Proc. of ASJ Autumn Meeting,* October 1990(in Japanese).

[9] 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.

[10] 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).

[11] T.Watanabe, K.Yoshida, and K.Hatazaki, "High Speed Continuous Speech Recognition Using a Bundle Search Algorithm," *Journal of IEICE,,* Vol.J75-D-II, No.11, pp.1761-1769, 1992.

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