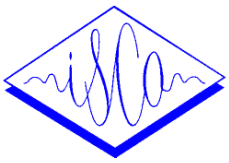


2nd International Conference on
Spoken Language Processing (ICSLP 92)
Banff, Alberta, Canada
October 12-16, 1992

SPEAKER ADAPTATION BY MODIFYING MIXTURE COEFFICIENTS OF SPEAKER-INDEPENDENT MIXTURE GAUSSIAN HMMS

Tatsuo Matsuoka† and Kiyohiro Shikano

NTT Human Interface Laboratories
3 - 9 - 11, Midori-cho, Musashino-shi, Tokyo, 180 JAPAN

Abstract

This paper proposes a new speaker adaptation method that uses speaker-independent HMMs as initial models, and it emphasizes the feature distribution of the target speaker's speech during adaptation training. A mixture Gaussian HMM with a large number of distributions attains good recognition accuracy for speaker-independent speech recognition when sufficient training speech is available. The proposed method uses such speaker-independent mixture Gaussian HMMs as initial models, and modifies mixture coefficients to maximize the likelihood for the target speaker. This method does not require phoneme segmentation and labeling of training speech, although it uses supervised training.

The adapted models using this method were evaluated by comparing with speaker-independent and speaker-dependent models. When the number of training words was less than 200, the speaker-adapted model achieved better recognition than either the speaker-independent or the speaker-dependent models.

10.21437/ICSLP.1992-107

1. INTRODUCTION

It is desirable that a speech recognition system can attain good recognition without preliminary adaptation training. A mixture Gaussian HMM with a large number of distributions performs well for speaker-independent speech recognition when a sufficient amount of training speech is available[1]. A recognizer using these kinds of speaker-independent HMMs as initial models does not require preliminary training, and can attain good recognition. It is widely agreed, however, that a speaker-dependent recognizer usually outperforms a speaker-independent recognizer, as long as a sufficient amount of training speech is available. Furthermore, if the amount of the target speaker's speech available is increasing, the recognizer should continue to improve performance by adapting model parameters for the target speaker. It is also desirable that speaker adaptation training does not require segmentation of speech and phoneme labeling because segmentation and labeling are very difficult and time consuming.

A number of speaker adaptation techniques have been reported in the literature[2]-[10]. A spectral mapping technique is one of the important techniques. For a recognizer based on discrete-density HMMs, the VQ-codebook mapping technique has been successfully applied[6]. For a recognizer based on continuous-density HMMs, the piecewise linear spectral mapping can be applied[7]. On the other hand, several speaker adaptation methods through speaker cluster selection have been proposed[8][9]. Using these speaker cluster selection adaptation, sharper and more appropriate HMMs can be generated. However, because each cluster has only a part of the training data, a large database is required.

A speaker adaptation technique that uses speaker-independent models as initial models is expected to be more robust than a technique that uses standard speaker's models. This paper proposes a speaker adaptation method that uses speaker-independent HMMs as initial models, and it emphasizes the feature distribution of the target speaker's speech during adaptation training. This method does not require utterances of predefined training words, nor does it require segmentation of training speech data.

2. SPEAKER ADAPTATION BY MIXTURE COEFFICIENTS

The method proposed here uses speaker-independent mixture Gaussian HMMs as initial models, and modifies mixture coefficients to maximize the likelihood for the target speaker's speech during adaptation training. Speaker-independent models represent many speakers' acoustic characteristics. The mean and variance values can be expected to contain speaker-common characteristics and various phoneme-environmental characteristics because the mean and variance are estimated from a large amount of speech. The speaker-independent HMMs probably should include

† Tatsuo Matsuoka is now staying at AT&T Bell Laboratories, Murray Hill, for his one-year visit as a visiting researcher.

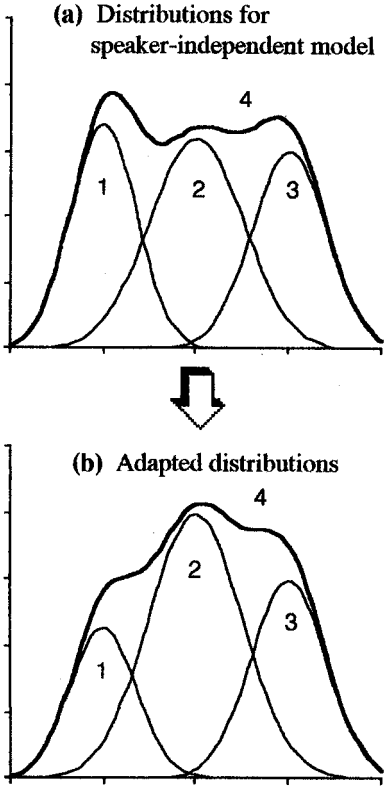


Figure 1 Adaptation of mixture Gaussian distribution

target speaker's characteristics, and adaptation training focuses on the target speaker's feature distribution so as to improve the recognition performance.

Figure 1 illustrates the process that this method adapts the models to the target speaker's speech. A speaker-independent mixture Gaussian HMM is expected to represent each speaker's or each speaker cluster's feature distribution as shown in Figure 1-(a). In Figure 1-(a), the curves 1 to 3 indicate distributions for a state in a speaker-independent mixture Gaussian HMM. The curve 4 is the weighted linear combination of the curves 1 to 3. For example, if a certain phoneme of a speaker is represented by the distribution 2, the distribution 4 is too broad to model the acoustic characteristics of that phoneme of the speaker. When the distribution 1 or the distribution 3 overlaps different speaker's different phoneme's distribution, that will cause recognition errors. Improvement of recognition will be possible by adapting a weighting coefficient for each distribution so as to maximize the likelihood value for the target speaker. Figure 1-(b) shows the adapted distributions. The distribution 2 was emphasized and the distributions 1 and 3 were weakened.

The adaptation training does not change mean and covariance values and state transition probabilities, but it only modifies weighting coefficients of the distributions. This is because adaptation using only a small amount of the target speaker's speech is desirable and it is difficult to reestimate all HMM parameters when the amount of training speech is limited.

To enable the adaptation without segmentation of speech and labeling, the adaptation training is carried out using the *"concatenated training"*[11]. For the *concatenated training*, phoneme HMMs are concatenated according to the words or sentences uttered, and the Forward-Backward algorithm is applied for the whole sequence of HMMs.

3. EXPERIMENTS

3.1 Experimental setup

This adaptation method was evaluated for Japanese vowel recognition and phoneme recognition. The speaker-adapted model was compared with speaker-independent and speaker-dependent models.

The speaker-dependent HMMs and the speaker-adapted HMMs were trained by the *concatenated training*, since the training speech was supposed to be neither segmented into phonemes nor labeled. The phonemes that had sufficient occurrence for training had two models, for the beginning of words and the middle of words. Sixty five models were used in total. The models are listed in Table 1. Each model had four states including a termination state, and three loops. All models were continuous-density HMMs with diagonal covariance matrices.

Table 2 shows the database and feature parameters used in the experiments. The even-numbered words in ATR 5240-word database were used for training, and the odd-numbered words were used for testing.

The speaker-independent model, used as an initial

Table 1 Phoneme models

/ a, i, u, e, o, b, d, g, m, n, p, t, k, s, sh, h, f, ch, ts, w, r, y, z, j /	Context-dependent • For the beginning of words • For the middle of words
/ by, gy, my, ny, py, ky, hy, ry, aa, ii, uu, ee, oo, ei, ou, N, Q /	Context-independent

Table 2 Speech data

Speech data	• ATR 5240 Japanese important words • 216 phoneme balanced words • 20 speakers (male: 10, female: 10) • For training: 16 speakers For testing: 4 speakers
Sampling	• 12 kHz, 16 bits
Analysis	• Hamming window • Frame length: 32 ms • Frame shift: 8 ms
Feature parameter	• Cepstrum (16) • Δ-Cepstrum (16) • Δ-Power

model, was a mixture Gaussian HMM with diagonal covariance matrices. The number of distributions in each state was set to 16 so that the model could represent the feature distributions of many speakers. The speaker-independent model was trained using training speech from 8 male speakers and 8 female speakers, about 300 words each. The number of distributions in each state of the speaker-adapted model was also 16. The number of distributions for the speaker-dependent model was set to one, because this model was trained using a small amount of speech. In the training of speaker-dependent model, a single Gaussian HMM trained by 16 training speakers was used as an initial model. Therefore, the difference between the training for the speaker-adapted HMM and that for the speaker-dependent HMM was the parameters to be modified. Only weighting coefficients were re-trained for the speaker-adaptive model, while all HMM parameters were re-trained for the speaker-dependent model.

The speaker-dependent and the speaker-adapted models were trained using the target speaker's speech, changing the number of training words from 50 to 550. For adaptation training, 216 phoneme balanced words and a part of the even-numbered words from ATR 5240-word database were used.

3.2 Results

Figure 2 shows the results for Japanese vowel recognition. The dashed line indicates the recognition error rate for the speaker-independent model. (This speaker-independent model was also used as an initial model for the speaker-adapted model.) The filled dots and the filled squares indicate the results for the speaker-dependent models and the speaker-adapted models, respectively.

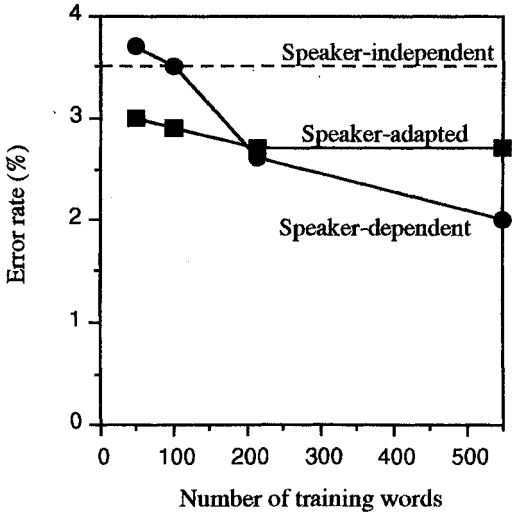
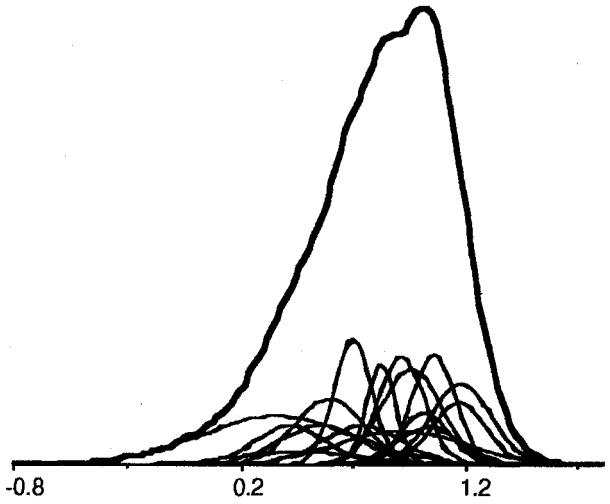


Figure 2 Results for vowel recognition

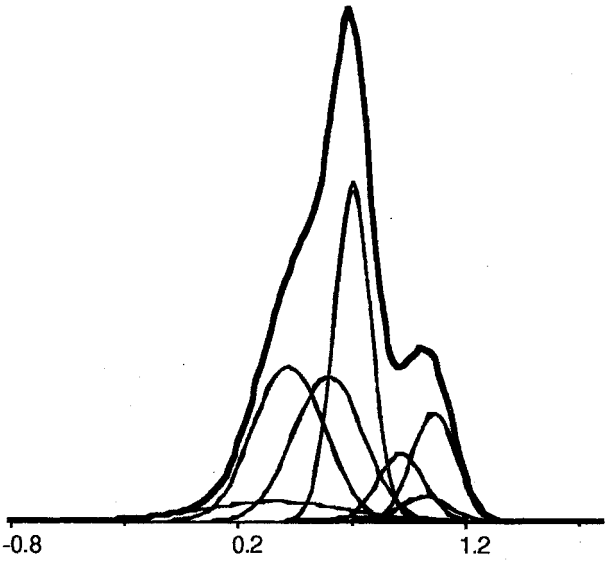
When the number of training words was less than 216, the speaker-adapted model achieved better recognition than either the speaker-independent or the speaker-dependent models. When the number of training words was more than 216, the speaker-dependent model was the best.

This adaptation method was also evaluated for Japanese phoneme recognition. In this case, when the number of training words was 216, the error rates were 9.7%, 10.0% and 12.8% for the speaker-adapted model, for the speaker-independent model, and for the speaker-dependent model, respectively. Though the speaker-adapted model was better than the speaker-dependent model, the difference between the speaker-adapted model and speaker-independent model was small.

Figure 3 shows the actual distributions of the first cepstral coefficients for the phoneme /a/. Figure 3-(a) shows the distributions for the speaker-independent model. There are 16 distributions inside and the biggest curve is the weighted linear combination of the distributions. Figure 3-(b) shows the distributions for the speaker-adapted model. In Figure 3-(b), only 7 distributions of the 16 distributions can be seen inside. This is because the distributions that were not effective for the target speaker was eliminated during adaptation training. The linear combination of the distributions is sharper than that of the speaker-independent model. This sharpness prevents confusion of phonemes.



(a) Distributions for the speaker-independent model



(b) Distributions for the speaker-adapted model

Figure 3 Distributions of the first cepstral coefficient for the phoneme /a/

4. SUMMARY

A new speaker adaptation method was proposed. The method uses speaker-independent mixture Gaussian HMMs as initial models, and it modifies mixture coefficients to maximize the likelihood for the target speaker's speech during adaptation training. This method does not require phoneme segmentation and labeling of speech, although it uses supervised training.

The proposed method was evaluated for Japanese vowel recognition. When the number of training words was less than 200, the speaker-adapted model achieved better recognition than either the speaker-independent or the speaker-dependent models.

If the amount of target speaker's speech available is increasing, it would be helpful to have a model that gradually changes from speaker-independent to speaker-dependent. The proposed method is expected to serve as the basis for a method that automatically interpolates speaker-independent and speaker-dependent models according to the amount of available speech data.

Acknowledgment

The authors wish to thank Dr. Furui and other members of Furui Research Laboratory, NTT Human Interface Laboratories as well as the members of NTT Basic Research Laboratories for their helpful discussions.

References

[1] C. H. Lee, L. R. Rabiner, R. Pieraccini and J. G. Wilpon, "Acoustic Modeling for Large Vocabulary Speech Recognition," Computer Speech and Language, Vol. 4, pp. 127-165, 1990

[2] K. Shikano, K. -F. Lee, and R. Reddy, "Speaker Adaptation Through Vector Quantization," Proc. ICASSP86, 49.5, pp. 2643-2646, 1986

[3] R. Schwartz, Y.-L. Chow and F. Kubala, "Rapid Speaker Adaptation Using a Probabilistic Spectral Mapping," Proc. ICASSP87, 15.3, pp. 633-636, 1987

[4] Y. Shiraki and M. Honda, "Speaker Adaptation Algorithms based on Piecewise Moving Adaptive Segment Quantization Method," Proc. ICASSP90, S12.5, 1990

[5] S. Furui, "Unsupervised Speaker Adaptation based on Hierarchical Spectral Clustering," Proc. ICASSP89, S6.9, pp. 286-289, 1989

[6] S. Nakamura and K. Shikano, "Speaker Adaptation Applied to HMM and Neural Networks," Proc. ICASSP89, S3.3, pp. 89-92, 1989

[7] H. Matsumoto and H. Inoue, "A Piecewise Linear Spectral Mapping for Supervised Speaker Adaptation," Proc. ICASSP92, pp. I449-452, 1992

[8] A. Imamura, "Speaker-Adaptive HMM-based Speech Recognition with a Stochastic Speaker Classifier," Proc. ICASSP91, S13.3, pp. 841-844, 1991

[9] H. Hattori, "Speaker Adaptation based on Markov Modeling of Speakers in Speaker-Independent Speech Recognition," Proc. ICASSP91, S13.4, pp. 845-848, 1991

[10] J. -L. Gauvain and C. H. Lee, "Bayesian Learning of Gaussian Mixture Densities for Hidden Markov Models," Proc. DARPA Speech and Natural Language Workshop, pp. 272-277, Feb. 1991

[11] Y. Minami, T. Matsuoka and K. Shikano, "Evaluation of Concatenated Training HMMs Using a Corpus for Speaker-Independent Continuous Speech Recognition," IEICE, Vol. SP91-113, 1992