論文 / 著書情報
Article / Book Information

Title	Speaker Adaptation of Tied-Mixture-Based Phoneme Models for Text-Prompted Speaker Recognition
Author	Tomoko Matsui, Sadaoki Furui
Journal/Book name	IEEE ICASSP94, Vol. , No. I, pp. 125-128
発行日 / Issue date	1994, 4
権利情報 / Copyright	(c)1994 IEEE. Personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution to servers or lists, or to reuse any copyrighted component of this work in other works must be obtained from the IEEE.

Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing,
Adelaide, Australia, pp. I-125-128 (April 1994)

SPEAKER ADAPTATION OF TIED-MIXTURE-BASED PHONEME MODELS FOR TEXT-PROMPTED SPEAKER RECOGNITION

Tomoko Matsui *Sadaoki Furui*

NTT Human Interface Laboratories
3-9-11, Midori-Cho, Musashino-Shi, Tokyo, 180 Japan

ABSTRACT

Speaker adaptation methods for tied-mixture-based phoneme models are investigated for text-prompted speaker recognition. For this type of speaker recognition, speaker-specific phoneme models are essential for verifying both the key text and the speaker. This paper proposes a new method of creating speaker-specific phoneme models. This uses speaker-independent (universal) phoneme models consisting of tied-mixture HMMs and adapts the feature space of the tied-mixtures to that of the speaker through phoneme-dependent/independent iterative training. Therefore, it can adapt models of phonemes that have a small amount of training data to the speaker. The proposed method was tested using 15 speakers' voices recorded over 10 months and achieved a speaker and text verification rate of 99.4% even when both the voices of different speakers and different texts uttered by the true speaker were to be rejected.

1 INTRODUCTION

We recently reported very efficient text-prompted speaker recognition methods using only a limited number of training utterances for each speaker [1][2]. In text-prompted speaker recognition, an arbitrary key text can be used at each recognition, and the recognizer accepts the input utterance only when it decides that the true speaker correctly uttered the prompted sentence. Therefore, it is unnecessary to worry about the system being fooled by recordings of key words or sentences uttered by the registered speaker. In this type of speaker recognition, the speaker verification and text confirmation require models that accurately represent both speaker and phonetic information. Reference [2] compared methods that create speaker-specific phoneme models and showed the effectiveness of a method that was based on speaker-adaptation of universal phoneme models (3-state 4-mixture continuous HMMs) accomplished by estimating the mean values and the weighting factors of the mixtures.

By incorporating tied-mixture HMMs for universal phoneme models, this paper extends the previous work based on speaker adaptation of universal phoneme models for creating speaker-specific phoneme models. The tied-mixture HMM is better than the continuous HMM for the following two reasons. First, the tied-mixture HMM has a mixture of Gaussian components in which mean and variance values are tied across all the phoneme models, and it can be used to estimate parameters efficiently. This works very well when the amount of training data is small [3]. Second, the speaker information accumulated over all phonemes, which is used in text-independent speaker recognition [4], can be represented by the tied-mixtures. Therefore, speaker adaptation using tied-mixture HMMs is expected to be better than that using continuous HMMs for text-prompted speaker recognition.

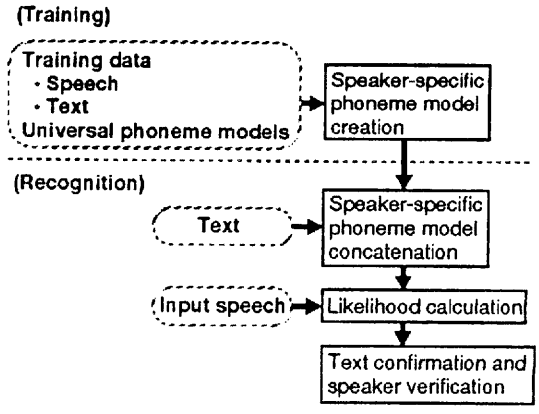

Figure 1. Main procedure of text-prompted speaker recognition.

In our previous work, we used a speaker adaptation method that was originally developed for speech recognition [5], and it proved to be inadequate for speaker recognition. The speaker-adapted phoneme models did not have enough speaker information, and the performance of the phoneme models was not good enough. To solve this problem, we used the combination method [2] of speaker-adapted phoneme models and a phoneme-independent speaker model, which made up for the lack of speaker information. The combination method had the problem that likelihood values for speaker verification and text confirmation had to be measured separately. In our new method presented in this paper, the combination is unnecessary, and the likelihood values for both speaker verification and text confirmation are evaluated simultaneously.

2 TEXT-PROMPTED SPEAKER RECOGNITION

The main procedure of text-prompted speaker recognition is shown in Figure 1. The system creates speaker-specific phoneme models for each reference speaker in the training phase. In the speaker and text verification, the phoneme-concatenation model corresponding to the key text is made, and the accumulated likelihood of the input speech frames for the model is compared with a threshold to decide whether to accept or reject the speaker.

Since the likelihood has a wide range for different input speech data, it is difficult to set stable thresholds for speaker and text verification using speech recorded at several sessions using different texts. To set stable thresholds, the likelihood value is normalized using a posteriori probability [2]. This normalization method was experimentally compared with a similar method [6] based on the likelihood ratio and shown to be more effective.

The a posteriori probability, which is used in the normalization method, is given by

$$p(s_c, t_c|x) = \frac{p(x|s_c, t_c) \times p(s_c, t_c)}{\sum_i \sum_j \{p(x|s_i, t_j) \times p(s_i, t_j)\}} \approx \frac{p(x|s_c, t_c)}{\sum_i \sum_j p(x|s_i, t_j)},$$

where s_i is a speaker, and t_j is a text; in particular s_c is the claimed speaker and t_c is the prompted text. The $p(s_i, t_j)$ is the simultaneous probability for speaker i and text j, and is assumed to be a constant for all combinations of speakers and texts. The $p(x|s_c, t_c)$ is the probability of the claimed speaker's HMM for the prompted text. $\sum_i \sum_j p(x|s_i, t_j)$ is approximated by the average of the n highest likelihoods by using parallel phoneme HMM networks for all registered speakers including the claimed speaker.

3 PHONEME DEPENDENT/INDEPENDENT ITERATIVE SPEAKER ADAPTATION: PDI

3.1 Principle

In speaker recognition, a method that needs a large amount of training data is unrealistic. However, when using only a small amount of training data, there can be some infrequent or even zero-frequency phonemes (phonemes that are not included in the training data), so it is difficult to adapt the model parameters to the speaker. Although it can be considered that the feature spaces for infrequent phonemes are estimated from those for frequent phonemes by using some assumptions about the geometric structure of the feature spaces, the geometric structure differs from speaker to speaker, so it will be very difficult to find efficient assumptions for each speaker. On the other hand, if universal phoneme models are directly used for infrequent phonemes, input speech that includes such phonemes may be mistakenly rejected since the speaker's voice does not necessarily fit the universal phoneme models.

Here we investigate a new method, in which the model parameters of even infrequent phonemes can be adapted to a new speaker. We assume that the feature parameter space for each speaker can be approximately represented by the distribution of feature parameters accumulated over several sentences. This assumption is based on our previous work [4] which indicated that such distribution represented by Gaussian mixtures or a VQ-codebook could be effectively used for text-independent speaker recognition. In the next section we introduce the procedure of our new method based on this principle.

3.2 Procedure

Phoneme dependent/independent iterative speaker adaptation (PDI) is a speaker adaptation method using tied-mixture HMMs for phoneme models. In this method, the feature space of the tied-mixtures is adapted to that of the speaker by applying both phoneme-dependent and independent training in series. There can be several types of series. As one of the robust determination techniques of HMM parameters when training data are insufficient, the technique of deleted interpolation is well-known [7]. In deleted interpolation, two (or more) models are trained separately, and the model parameters are determined by interpolating those models. However, it is very difficult to interpolate phoneme models and phoneme-independent speaker models. For PDI, tied-mixtures are used as a common component, and two different types of training are applied to them sequentially. The phoneme-dependent training corresponds to conventional speaker adaptation applied to each phoneme model. In the phoneme-independent training, speaker information across all phonemes is used for adapting the tied-mixtures, and the feature spaces of each phoneme are shifted according to the distribution of feature parameters across all the phonemes in the training data. Therefore, even phoneme models for which the amount of training data is small can be adapted to the speaker.

In the phoneme-dependent training, universal phoneme HMMs are concatenated according to the sequence of phonemes in the training text. The training speech data is applied to the phoneme-concatenation HMM, and the mean values and the weighting factors of the tied-mixtures are estimated for each phoneme [5]. In this training, models for phonemes that are not included in the training data are not explicitly adapted.

In the phoneme-independent training, a 1-state HMM is created using the tied-mixtures for each speaker. The training speech data is used for estimating both the mean values and the weighting factors of the mixtures characterizing the 1-state HMM for each speaker. (The initial values of the weighing factors are set to a common value for all the mixtures.) The mean values of the tied-mixtures are then replaced by those of the 1-state HMM for each speaker. Since phoneme information is not used in this training, it is not assumed that phonetic discrimination power is improved. However, it has the advantage that even a mixture used in modeling a phoneme not included in the training data can be shifted to the feature space of the speaker.

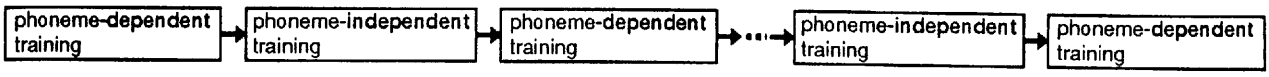

Figure 2. Phoneme dependent/independent iterative speaker adaptation (PDI) procedure.

Table 1. Verification error rates (%). [ ]: the error rate using the likelihood normalization method.

session	*continuous*	*tied-mixture*	*PDI*
T1	2.1 [2.2]	1.1 [1.6]	0.8 [0.9]
T2	2.5 [1.8]	3.6 [1.0]	3.0 [0.3]
T3	4.7 [2.3]	5.0 [1.7]	4.4 [1.0]
T4	1.5 [4.6]	1.5 [1.6]	1.1 [0.9]
Average	2.7 [2.7]	2.8 [1.3]	2.3 [0.7]

4 EXPERIMENTAL CONDITIONS

The database consists of sentence data uttered by 10 male and 5 female speakers. It was recorded in five sessions (T0-4) over ten months. Cepstral coefficients were calculated by LPC analysis with an order of 16, a frame period of 8 ms, and a frame length of 32 ms. Ten sentences from session T0 were used for training, and five sentences from sessions T1, T2, T3, or T4 were individually used for testing. In the ten sentences for training, the texts of half of them were the same for all speakers and all sessions, and the other half differed from speaker to speaker and from session to session. The sentences for testing were different from those for training, and were the same for all speakers and all recording sessions. 300 utterances (15 people × 5 sentences × 4 sessions) were used for evaluation. The average duration of each sentence was 4.2 s. The number of phoneme models was 41. The tied-mixture HMMs were 3-state 256-mixture HMMs.

For PDI, phoneme-dependent/independent training was carried as in Figure 2. (This series showed the best performance under these experimental conditions.) For the purpose of comparison, 3-state 4-mixture HMMs were used as continuous HMMs. The HMM parameters were estimated by using the Baum-Welch algorithm.

The performance of our method was evaluated by the speaker and text verification error rate. The threshold was set a posteriori to equalize the probability of false acceptance and false rejection. In these experiments, we also used speech data of texts that differed from the prompted texts but were uttered by the true speaker as data that should be rejected. The likelihood values were calculated using the trellis algorithm.

5 RESULTS

Table 1 lists the speaker and text verification error rates. Here, "continuous" and "tied-mixture" means methods

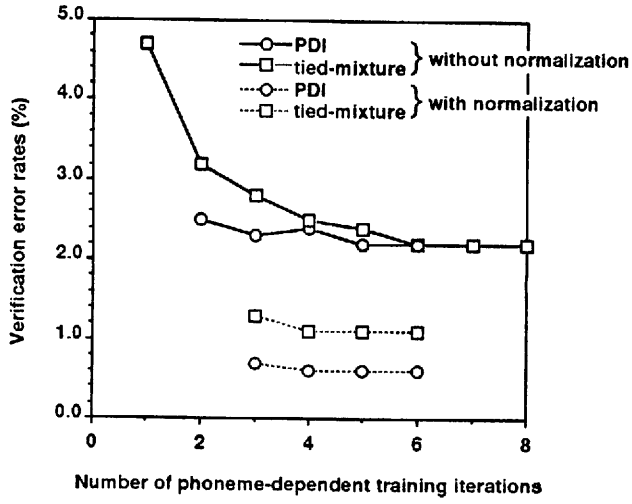

Figure 3. PDI vs. tied-mixture.

that create speaker-specific phoneme models by the conventional phoneme-dependent training algorithm for phoneme models using either continuous or tied-mixture HMMs. The [ ] indicates the error rate using the likelihood normalization method. For all methods, phoneme-dependent training was iterated three times. (For PDI, phoneme-dependent training was applied alternately, so the total number of training iterations in the series was five.) These results show that PDI is more effective than "continuous" and "tied-mixture," and that the likelihood normalization method is very effective.

Figure 3 compares performances of PDI and "tied-mixture" for different numbers of phoneme-dependent training iterations. For PDI, the total number of training iterations in the series was twice the number of phoneme-dependent training iterations minus one. These results show that PDI is more effective than "tied-mixture", especially when using the likelihood normalization method.

6 DISCUSSION

6.1 Phoneme-dependent/independent training series

For PDI, additional experiments for several types of phoneme-dependent/independent training series were carried out. Table 2 lists the speaker and text verification

error rates. "D^n" means that phoneme-dependent training was iterated n times. "ID^n" means that phoneme-independent training was applied first and then phoneme-dependent training was iterated n times. "$(ID)^n$" means that the series of phoneme-independent and dependent training was iterated n times. "$D(ID)^n$" is the same series as in Figure 2. These results indicate that when the number of iterations for phoneme-dependent training is fixed, the error rate is decreased by applying phoneme-independent training for any series. The best accuracy, 99.4%, was obtained for $D(ID)^4$.

6.2 Likelihood values of phoneme models

For PDI and "tied-mixture", the likelihood value of input speech for each phoneme model was examined as a function of the frequency of each phoneme in the training data. The phoneme for each input speech frame was determined by its alignment using the Viterbi search. Before the comparison between PDI and "tied-mixture", the log-likelihood values were normalized to the normal distribution $N(0,1)$ for each method. The averaged differences between the likelihood values for each phoneme model by PDI and "tied-mixture" were calculated and ordered according to the frequency of each phoneme in the training data. Table 3 lists the averaged differences when the phoneme frequency order was divided into four parts. The first part "1st-10th" represents phonemes that are from the lowest to the 10th-lowest frequency. These results show that for infrequent phonemes, the likelihood values of PDI are higher than those of "tied-mixture" and vice versa for frequent phonemes. This indicates that infrequent phonemes are more effectively adapted to the speaker using PDI than using "tied-mixture".

Table 2. Verification error rates for several types of phoneme-dependent/independent training series. D: phoneme-dependent training, I: phoneme-independent training.

No. of iterations		*I*				
		0	1	2	3	4
D	1	D 4.7	ID 3.1	-	-	-
	2	D^2 3.2	DID 2.5	$(ID)^2$ 2.7	-	-
	3	D^3 2.8 [1.3]	ID^3 2.7	$D(ID)^2$ 2.3 [0.7]	$(ID)^3$ 2.5 [1.3]	-
	4	D^4 2.5 [1.1]	-	-	$D(ID)^3$ 2.4 [0.6]	$(ID)^4$ 2.4
	5	D^5 2.4 [1.1]	-	-	-	$D(ID)^4$ 2.2 [0.6]

Table 3. Difference between the normalized log-likelihood values of PDI and "tied-mixture": $L(PDI) - L(tied-mixture)$.

session	*phoneme frequency order (low → high)*			
	1st-10th	11th-20th	21th-30th	31th-41th
T1	0.14	0.00	0.05	-0.07
T2	0.20	-0.04	0.03	-0.05
T3	0.23	-0.04	0.05	-0.06
T4	0.20	0.04	0.00	-0.07
Average	0.19	-0.01	0.03	-0.06

7 CONCLUSIONS

A new speaker adaptation method called PDI has been developed by using tied-mixture-based phoneme models for text-prompted speaker recognition. PDI is more effective than the conventional methods. It uses the advantage of the tied-mixture HMMs, which can represent speaker information accumulated over all phonemes in the tied-mixtures. The phoneme models made by PDI can, therefore, be used simultaneously for speaker and text verification. We also showed that the likelihood normalization method is very effective. PDI achieved a speaker and text verification rate of 99.4%.

For future work, we plan to investigate the geometric structure of the feature spaces for phonemes in order to estimate the feature spaces for infrequent phonemes from those for frequent phonemes.

8 ACKNOWLEDGMENT

The authors wish to thank the members of the Furui Research Laboratory of NTT Human Interface Laboratories for their valuable and stimulating discussions.

REFERENCES

[1] T. Matsui and S. Furui, "Speaker Recognition Using Concatenated Phoneme HMMs," *Proc. ICSLP, Banff, pp. I-603-606 (1992)*

[2] T. Matsui and S. Furui, "Concatenated Phoneme Models for Text-Variable Speaker Recognition," *Proc. ICASSP, Minneapolis, pp. II-391-394 (1993)*

[3] X. D. Huang, "Phoneme Classification Using Semi-continuous Hidden Markov Models," *IEEE Trans. ASSP. Vol. 40, No. 5, pp. 1062-1067 (1992)*

[4] T. Matsui and S. Furui, "A Text-Independent Speaker Recognition Method Robust Against Utterance Variations," *Proc. ICASSP, Toronto, pp. I-377-380 (1991)*

[5] K.-F. Lee, "Automatic Speech Recognition - The Development of the SPHINX System," *Kluwer Academic Publishers, Boston (1989)*

[6] A. Higgins, L. Bahler, and J. Porter, "Speaker verification using randomized phrase prompting," *Digital Signal Processing 1, pp. 89-106 (1991)*

[7] F. Jelinek and R.L. Mercer, "Interpolated estimation of Markov source parameters from sparse data," *in Pattern Recognition in Practice, E.S. Gelesma and L.N. Kanal, Eds. Amsterdam, The Netherlands: North-Holland, pp. 381-397 (1980)*