論文 / 著書情報
Article / Book Information

Title	Concatenated Phoneme Models for Text-variable Speaker Recognition
Author	Tomoko Matsui, Sadaoki Furui
Journal/Book name	IEEE ICASSP1993, Vol. , No. 2, pp. 391-394
発行日 / Issue date	1993, 4
権利情報 / Copyright	(c)1993 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.

Reprinted from ICASSP-93 — 1993 IEEE INTERNATIONAL
CONFERENCE ON ACOUSTICS, SPEECH, AND SIGNAL
PROCESSING, Minneapolis, Minnesota, April 27-30, 1993

CONCATENATED PHONEME MODELS FOR TEXT-VARIABLE SPEAKER RECOGNITION

Tomoko Matsui   *Sadaoki Furui*
NTT Human Interface Laboratories
9-11, Midori-Cho 3-Chome
Musashino-Shi, Tokyo, 180 Japan

ABSTRACT

This paper investigates methods that create models to specify both speaker and phonetic information accurately by using only a small amount of training data for each speaker. For a text-dependent speaker recognition method, in which arbitrary key texts are prompted from the recognizer, speaker-specific phoneme models are necessary to identify the key text and recognize the speaker. Two methods of making speaker-specific phoneme models are discussed: phoneme-adaptation of a phoneme-independent speaker model and speaker-adaptation of universal phoneme models. Moreover, we also investigate supplementing these methods by adding a phoneme-independent speaker model to make up for the lack of speaker information. This combination achieves a rejection rate as high as 98.5% for speech that differs from the key text and a speaker verification rate of 100.0%.

1 INTRODUCTION

In text-dependent speaker recognition, the key text is usually fixed. The speaker recognition key, however, can easily be cracked by recording the registered speaker's voice uttering the key text. It would be better, therefore, if the key text could be changed every time the recognizer was used and the voice was accepted only when the true speaker utters the prompted text.

Some studies [1]-[3] have reported speaker recognition methods using sequences that consist of keywords such as digits and some fixed words for recognition. The sequences can be changed every time the recognizer is used. Modern digital recorders, however, can play back an arbitrary sequence of keywords.

Our recent studies [4][5] reported a speaker recognition method in which an arbitrary key text can be used at each recognition. The recognition system accepts the input utterance only when it decides that the true speaker correctly uttered the prompted sentence. Reference [4] reported three basic structures for implementing the method, and [5] reported some experimental results for one of them. That method [5] used speaker-specific phoneme HMMs (hidden Markov models), made by using only training utterances for each speaker, as basic acoustic units. As only a limited number of training utterances were used for each speaker, the number of phoneme models was limited by the size of training utterances to 25. The rejection rate for speech uttered by the true speaker that differs from the key sentence was 48.5% and the verification rate was 96.7% without likelihood normalization methods. With normalization, the rejection rate was 80.7% and the verification rate was 99.9%. Therefore, phonetic information could not be represented sufficiently in that method.

In this paper, two methods (I and II) of making speaker-specific phoneme HMMs are discussed. Method I is based on phoneme-adaptation of a phoneme-independent speaker HMM, whereas Method II is based on speaker-adaptation of universal phoneme HMMs. Universal phoneme HMMs are also used in Method I to make up for the small amount of training data. Moreover, we also investigate supplementing these methods by adding a phoneme-independent speaker HMM to make up for the lack of speaker information.

2 METHODS

2.1 Main Procedure

The main procedure is shown in Figure 1. The system creates speaker-specific phoneme models for each reference speaker. The following sections show two methods of making speaker-specific phoneme models.

In the speaker verification procedure, the phoneme-concatenation model corresponding to the key text is made, and the accumulated likelihood of the HMM for input speech frames is used to confirm the text and to accept or reject the speaker. The thresholds of the likelihood value are set for the text confirmation and the speaker verification.

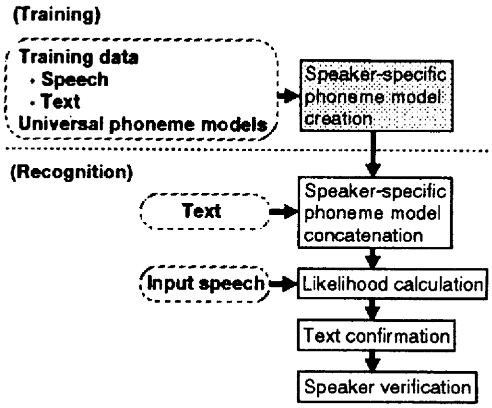

Figure 1. Block diagram of main procedure.

0-7803-0946-4/93 $3.00 © 1993 IEEE

2.2 Method I

Method I is based on phoneme-adaptation of a phoneme-independent speaker HMM. In this method, a phoneme-independent speaker HMM is trained to give it phonetic information by using training data of each speaker and universal phoneme HMMs.

Figure 2 shows a block diagram of this method. A phoneme-independent speaker model is first created for each speaker as a 1-state 64-mixture Gaussian HMM. Then each state in universal (speaker-independent) phoneme models, which are made as 3-state 4-mixture Gaussian HMMs using a large amount of speech data uttered by many speakers, is represented using a phoneme-independent speaker HMM. A set of data is artificially created so as to satisfy the 4-mixture Gaussian distribution of each state in universal phoneme HMMs. The data set is applied to phoneme-independent speaker 1-state 64-mixture HMMs and the mixture weighting factors are adapted (estimated) for the distribution of each state in universal phoneme HMMs. (The multiple-speaker data can be applied to phoneme-independent speaker 1-state 64-mixture HMMs, but the amount of calculation becomes enormous.) Then initial models for speaker-specific phoneme HMMs are made as 3-state 64-mixture HMMs using the 1-state 64-mixture HMMs and transition probabilities of universal phoneme HMMs.

The phoneme HMMs are concatenated in the sequence of phonemes in the training text. The training speech data is applied to the phoneme-concatenation HMM and then the mixture weighting factors are estimated for each phoneme [6]. The mean and covariance values of the mixtures are fixed to the initial values. Finally, the speaker-specific phoneme models are created as 3-state 64-mixture HMMs.

In the above, only the mixture weighting factors are adapted in order to keep speaker information in a phoneme-independent speaker HMM.

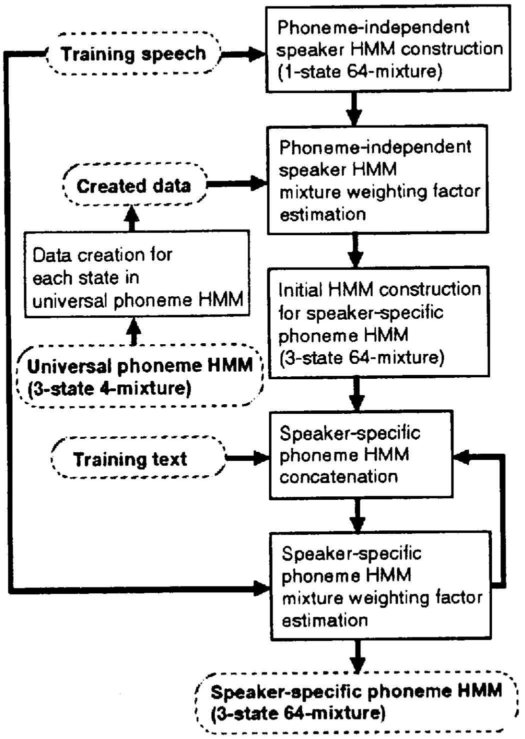
Training speech
Phoneme-independent speaker HMM construction (1-state 64-mixture)
Phoneme-independent speaker HMM mixture weighting factor estimation
Created data
Data creation for each state in universal phoneme HMM
Universal phoneme HMM (3-state 4-mixture)
Initial HMM construction for speaker-specific phoneme HMM (3-state 64-mixture)
Training text
Speaker-specific phoneme HMM concatenation
Speaker-specific phoneme HMM mixture weighting factor estimation
Speaker-specific phoneme HMM (3-state 64-mixture)

Figure 2. Block diagram of Method I.

2.3 Method II

Method II is based on universal phoneme HMMs and adapts them to the training speech of each speaker. Figure 3 shows a block diagram of Method II. Universal phoneme HMMs are used as the initial models for each speaker. Universal phoneme HMMs are concatenated in the sequence of phonemes in the training text. The training speech data is applied to the phoneme-concatenation HMM, and then the mixture weighting factors and the mean values of the mixtures are adapted (estimated) for each phoneme [6]. Finally, the speaker-specific phoneme models are created as 3-state 4-mixture HMMs.

In the above, the covariance values of the mixture are not estimated because of the small amount of training data.

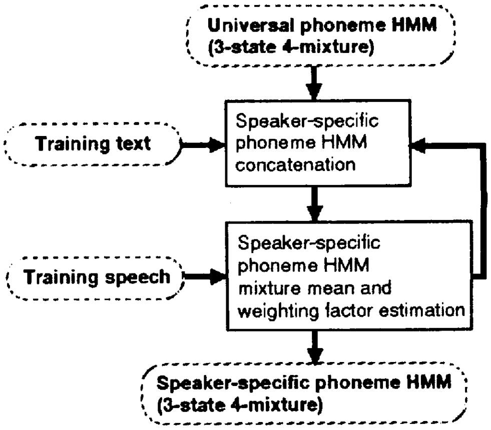
Universal phoneme HMM (3-state 4-mixture)
Training text
Speaker-specific phoneme HMM concatenation
Training speech
Speaker-specific phoneme HMM mixture mean and weighting factor estimation
Speaker-specific phoneme HMM (3-state 4-mixture)

Figure 3. Block diagram of Method II.

2.4 Combination of Speaker Models

We investigated supplementing the main procedure in Figure 1 by adding a phoneme-independent speaker HMM to make up for the lack of speaker information. The weighted summed likelihood L_{sum}, which is used for speaker decision, is given by

$$L_{sum} = w \times L_{SP} + (1 - w) \times L_{PI} \qquad 0 \leq w \leq 1,$$

where L_{SP} is the likelihood of speaker-specific phoneme HMMs and L_{PI} is the likelihood of a phoneme-independent speaker HMM. The value of weight w is set experimentally.

3 EXPERIMENTAL CONDITIONS

3.1 Database

The database consisted of sentence data uttered by 10 male and 5 female talkers. This database was recorded on three sessions (A, B, and C) over six months. Cepstral coefficients were calculated by LPC analysis with the order of 16, a frame period of 8 ms, and a frame length of 32 ms.

Ten sentences from session A were used for training, and five sentences from session B or C were used for testing. 150 utterances (15 people × 5 sentences × 2 sessions) were used for evaluation. The duration of each sentence was about 4 s.

3.2 Evaluation

The performance of our method was evaluated by the following two measures. One is the speaker verification rate. The threshold was set a posteriori to equalize the probability of false acceptance and false rejection. In these experiments, speech data of key texts uttered by each speaker correctly was used.

The other measure is the false acceptance rate for speech uttered by the true speaker that differs from the key text. The threshold for rejecting speech was set a posteriori so as not to reject any speech of any correct key texts uttered by the true speaker.

4 RESULTS

4.1 Speaker Verification

Figure4 shows speaker verification error rates. PI, $M0$, $M1$, $M2$, $M1+PI$ and $M2+PI$ have the following meanings:

PI: method using a phoneme-independent speaker HMM (1-state 64-mixture HMM, 1 phoneme model)

$M0$: previous method [5] (1-state 64-mixture HMM, 25 phoneme models)

$M1$: Method I (3-state 64-mixture HMM, 65 phoneme models)

$M2$: Method II (3-state 4-mixture HMM, 65 phoneme models)

$M1+PI$: combined method of Method I and a phoneme-independent speaker HMM (weight $w = 0.5$)

$M2+PI$: combined method of Method II and a phoneme-independent speaker HMM (weight $w = 0.625$)

Method I achieves a lower error rate for speaker verification than the phoneme-independent speaker HMM. This rate is slightly lower than the previous method and significantly lower than Method II. Thus, Method I keeps speaker information better than Method II, because it uses a phoneme-independent speaker HMM with sufficient speaker information as an initial model. The error rate for the method combining Method I (or Method II) and a phoneme-independent speaker HMM was almost one-tenth that using a phoneme-independent speaker HMM. Thus, the combination method is very effective for speaker verification.

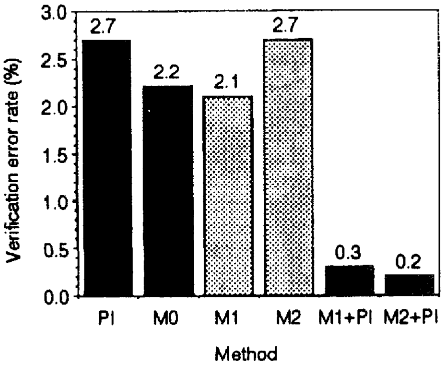

Figure 4. Verification error rates.

Figure 5 shows speaker verification error rates using the combination method with different values of weighting factor w. The value is not very sensitive and can be set to approximately 0.5.

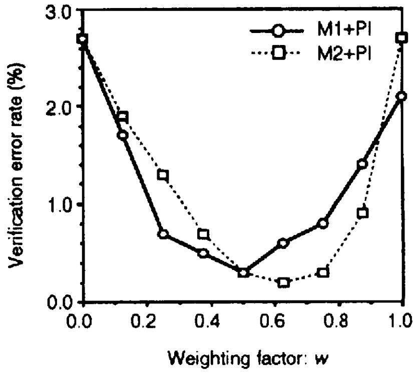

Figure 5. Verification error rates as functions of the values of weighting factor w

4.2 Rejection of Incorrect Speech

Figure 6 shows the false acceptance rates for speech that differs from the key text. The rate using Method I was roughly half that using the previous method [5]. The rate using Method II was only 3% of that using the previous method. This result indicates that speaker-specific phoneme HMMs in Method II have sufficient phonetic information, because Method II uses universal phoneme HMMs with sufficient phoneme information as initial models.

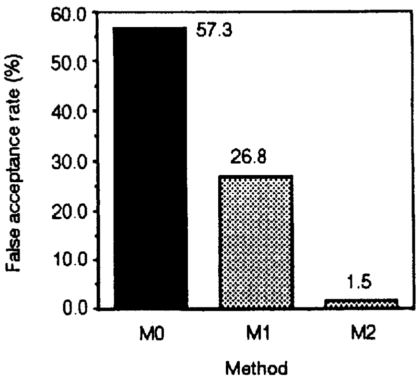

Figure 6. False acceptance rates for incorrect speech.

5 LIKELIHOOD NORMALIZATION

As the likelihood has a wide range for different input speech data, it is difficult to set stable thresholds for speaker verification and for rejection of incorrect speech using speech recorded on several sessions that have different texts. We investigated the effects of using likelihood normalization methods for speaker verification and for rejection of incorrect speech.

5.1 Likelihood Normalization for Speaker Verification

Higgins et al.[3] have reported a normalization method for similarity values (corresponding to the likelihood of HMMs in this paper) that uses the similarity values between input speech and models of other reference speakers. On the other hand, in our method, the likelihood value is normalized by subtracting the average value of the n highest likelihoods [5]. Figure 7 shows speaker verification error rates using Method I, Method II, and the two combined methods. For each method, the smallest error rate ($n = 3$) was 0%. These results confirm the effectiveness of this normalization method.

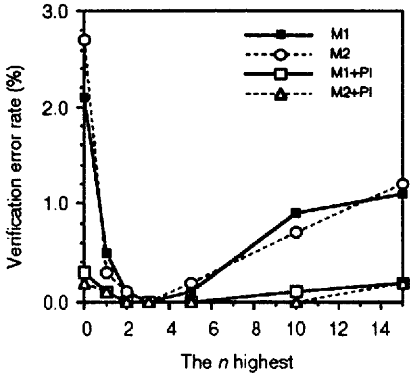

Figure 7. Likelihood normalization for speaker verification.

5.2 Likelihood Normalization for Rejection of Incorrect Speech

The likelihood value for rejection of incorrect speech was normalized by subtracting the likelihood value calculated using a text-independent model for the speaker. Here a phoneme-independent speaker HMM (1-state 64-mixture) for the speaker was used as the text-independent model. Figure 8 shows the results for the false acceptance rates using or not using the normalization method. For the previous method [5], the rate with the normalization method was roughly half that without the normalization method. For Method I, the rate with the normalization method was roughly 70% of that without the normalization method. For Method II, the rate with the normalization method is larger than that without the normalization method. This is probably because a phoneme-independent speaker HMM used as a text-independent model for the normalization method has a different structure from the universal phoneme HMMs.

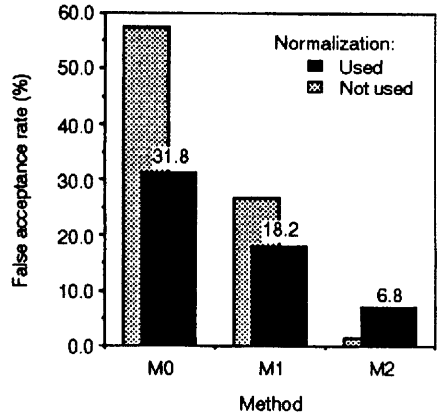

Figure 8. False acceptance rates for incorrect speech.

6 CONCLUSION

This paper investigated two methods for making speaker-specific phoneme models: one is based on phoneme-adaptation of a phoneme-independent speaker model (Method I) and the other is based on speaker-adaptation of universal phoneme models (Method II). For speaker verification, Method I was more efficient than Method II, and for rejection of incorrect speech, the reverse was true. Moreover the combination of either of these methods and a phoneme-independent speaker model was very effective for speaker verification. To set stable thresholds for speaker verification and speech rejection, normalization methods for the likelihood values were investigated. When combining the speaker-adaptive phoneme models and a phoneme-independent speaker model, and normalizing the likelihood values, the rejection rate was as high as 98.5% for speech uttered by the true speaker that differs from the key text and the speaker verification rate was 100.0%.

We are still investigating methods for making speaker-specific phoneme models and are studying methods for setting thresholds of speaker verification and speech rejection beforehand.

REFERENCES

[1] A. E. Rosenberg and F. K. Soong, "Recent Research in Automatic Speaker Recognition," *in Advances in Speech Signal Processing, S. Furui and M. M. Sondhi (Ed.), Marcel Dekker, New York, pp.701-738 (1992)*

[2] J. M. Naik, "Speaker Verification: A Tutorial," *IEEE Communications Magazine, 28, 1, pp.42-48 (1990)*

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

[4] S. Furui and T. Matsui, "Free-Text Speaker Recognition Methods Using Phoneme Class Models," *IEEE Workshop on Interactive Voice Technology for Telecommunications Applications, NJ (1992)*

[5] T. Matsui and S. Furui, "Speaker Recognition Using Concatenated Phoneme HMMs," *Proc. Int. Conf. Spoken Language Processing, Banff, Th.sAM.4.3 (1992)*

[6] Kai-Fu Lee, "Automatic Speech Recognition - The Development of the SPHINX System," *Kluwer Academic Publishers, Boston (1989)*