

論文 / 著書情報
Article / Book Information

Title	Speaker Recognition Using Concatenated Phoneme Models
Author	Tomoko Matsui, Sadaoki Furui
Journal/Book name	Proc. ICSLP'92, Vol. , No. , pp. 603-606
発行日 / Issue date	1992,

SPEAKER RECOGNITION USING CONCATENATED PHONEME MODELS

Tomoko Matsui Sadaoki Furui

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

ABSTRACT

This paper investigates a new text-dependent speaker recognition method in which the key texts can be changed every time the recognizer is used and the voice is accepted only when the true speaker utters the prompted text. This method solves the problem in conventional methods which use fixed texts and thus can be defeated by recording the true speaker's voice. Our method is based on the following techniques. First, a text-independent speaker spectral model is created for each speaker as a 1-state 64-mixture-Gaussian hidden Markov model (HMM) with diagonal covariance matrices. Next, the mixture weighting factors are estimated for each phoneme class using training utterances of each speaker. Then the likelihood of the phoneme-class-concatenation model corresponding to the prompted key text is used for the recognition decision. When the proposed method is used, the rejection rate for speech uttered by the true speaker that differs from the key text is as high as 80.7% and the verification rate is 99.9%.

1 INTRODUCTION

Text-dependent speaker recognition, using key words or sentences, is usually more reliable than text-independent recognition because it uses text-specific voice information. However, if the key text is fixed, the speaker recognition key can easily be cracked by recording the registered speaker's voice uttering the key text. It would be better, therefore, if the key texts could be changed every time the recognizer was used and the voice was accepted only when the true speaker utters the prompted text. In speaker recognition using non-fixed key texts, the text-specific voice information cannot be registered in the system beforehand. The text-specific voice information must be constructed using speaker-specific acoustic units according to the key text every time the text is changed.

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.

This paper investigates a new text-dependent speaker recognition method in which the key texts can be changed every time the recognizer is used, and the voice is accepted only when the true speaker utters the prompted text.

2 METHODS

The system uses speaker-dependent phoneme class models created for each reference speaker. Figure 1 illustrates the training procedure. In speaker recognition, a method that needs a large amount of training data is unrealistic. The serious problem here is how to make stable phoneme class models of each speaker using a small amount of data. A text-independent speaker spectral model is first created for each speaker as a 1-state 64-mixture Gaussian hidden Markov model with diagonal covariance matrices, and the model is used as an initial model of each phoneme class. Phoneme class models are concatenated in the sequence of phoneme classes in the training text. The training speech data is applied to the phoneme-class-concatenation model and then the mixture weighting factors are estimated for each phoneme class [4]. The mean values and covariance values of the mixtures are fixed to the initial values.

The speaker recognition procedure is shown in Figure 2. The phoneme-class-concatenation model corresponding to the key text is made, and the accumulated likelihood of the model for input speech frames is used to confirm the text and to accept or reject the speaker. The concatenation model produces a relatively high likelihood for the correct key texts uttered by the true

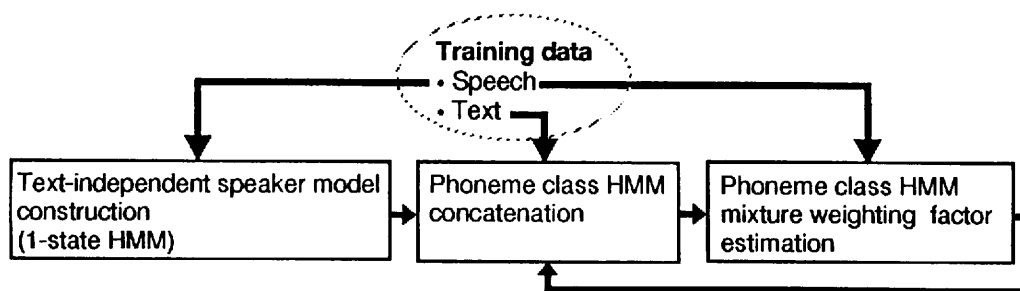


Figure 1. Training procedure for each reference speaker.

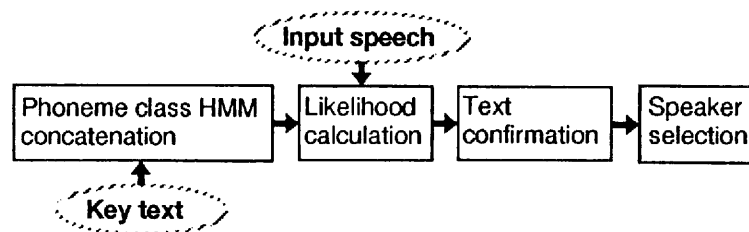


Figure 2. Speaker recognition procedure.

speaker, and a relatively low likelihood for other speakers or for the wrong key text uttered even by the same speaker. The thresholds of the likelihood value are set for the text confirmation and the speaker verification.

3 EXPERIMENTS

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 $\times$ 5 sentences $\times$ 2 sessions) were used for evaluation. The duration of each sentence was about 4 s. For HMMs, two transition probabilities derived from the same state were initialized identically. Two arcs derived from the same state had the same output probabilities.

The performance of our method was evaluated by the following three measures. One is the speaker verification rate and the next one is the speaker identification rate. The threshold for speaker verification 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 rejection 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 key texts uttered by the true speaker.

4 EXPERIMENTAL RESULTS

4.1 Speaker Verification and Identification

Table 1 lists speaker verification error rates. The error rate for speaker verification using 25 phoneme classes was roughly 80% that of methods not using phoneme classes.

Table 1. Verification error rates (%).

Phoneme Classes	1	15	25
Session B	4.4	3.4	3.4
Session C	4.1	3.7	3.7
Average	4.3	3.6	3.6

For speaker identification, the error rates were constant, independent of the number of phoneme classes. The average error rate for test data on sessions B and C was 4.7%.

Table 2. Likelihood normalization effects for speaker verification (%).

Normalization	Not used	Used					
The n highest	-	1	2	3	5	10	15
Session B	3.4	0.6	0.1	0.3	0.8	1.3	1.1
Session C	3.7	0.2	0.1	0.0	0.1	0.9	1.0
Average	3.6	0.4	0.1	0.2	0.5	1.1	1.1

4.2 Likelihood Normalization for Speaker Verification

The likelihood has a wide range for different input speech data. It is difficult to set stable thresholds of speaker verification for all speech recorded on several sessions that have different texts. Higgins [3] has reported the normalization method of similarity values (in this paper, likelihood of HMMs) that uses the similarity values between input speech and models of other reference speakers.

Here the likelihood value is normalized by subtracting the average value of the n highest likelihoods. Table 2 lists speaker verification error rates. The small error rates show that the verification error rate is almost one-thirtieth that not using the normalization method. These results confirm the effectiveness of the normalization method proposed in this paper.

4.3 Rejection of Incorrect Speech

Table 3 lists the rejection rates for incorrect speech using various numbers of phoneme classes. In this section, no methods for likelihood normalization were used. Session B + C means that the threshold was set not to reject incorrect speech of key texts on either session B or C. For 25 phoneme classes, this method could reject roughly half the incorrect speech samples.

Table 3. Rejection rates (%).

Phoneme Classes	6	15	25
Session B	41.3	63.7	69.0
Session C	37.7	54.0	60.7
Session B + C	24.3	39.7	48.3

4.4 Likelihood Normalization for Rejection of Incorrect Speech

In order to improve the performance of rejecting speech that differs from the key text, the likelihood value was normalized by subtracting the likelihood value using the text-independent 1-state 64-mixture Gaussian hidden Markov model for the speaker. Table 4 lists the results for the rejection rates using or not using the normalization method. With the normalization method, the rejection rates become more than 80% and the error rates are roughly half those without the normalization method.

Table 4. Rejection rates (%).

Normalization	Not used	Used
Session B	69.0	82.0
Session C	60.7	81.3
Session B + C	48.3	80.7

5 ADDITIONAL EXPERIMENTS ON PHONEME CLASS MODELS

Our method uses speaker-dependent phoneme class models, and the likelihood of the concatenated models is used to reject speech that differs from the key text and to verify and identify the speaker. It would be possible to use the conventional speaker-independent phoneme models, which are usually used to recognize words and sentences, to reject incorrect speech. Here we compare the performance for incorrect speech rejection and sentence recognition of our new method with the conventional speaker-independent method, and evaluate the performance of our method.

5.1 Performance of Incorrect Speech Rejection

The speaker-independent phoneme class models were trained using a large amount of sentence data uttered by eight male and female speakers, who were

different from the set of 15 speakers used in the experiments just mentioned. The number of phoneme classes was 65. Table 5 lists the results for the rejection rates by our method (SD) that uses 25 speaker-dependent phoneme class models and the method (SI) that uses speaker-independent phoneme class models. No methods for likelihood normalization were used. Both rejection rates are almost same. The proposed method, however, has an advantage that it can confirm the text and verify the speaker in one pass.

Table 5. Rejection rates (%).

Phoneme Class	SD	SI
Session C	60.7	61.7

5.2 Performance of Sentence Recognition

As an additional experiment to check the performance of phoneme class models, a sentence recognition test was performed. In this experiment, one of the five sentences that produced the highest likelihood for each input speech was selected. Other experimental conditions were the same as above.

Table 6 lists the sentence recognition rates. These results indicate that the proposed method is effective for making speaker-dependent phoneme class models.

Table 6. Sentence recognition rates (%).

Phoneme Class	SD	SI
Session C	98.0	93.0

6 CONCLUSIONS

This paper proposed a text-dependent speaker recognition method in which the key texts can be changed every time the recognizer is used and the voice is accepted only when the true speaker utters the prompted text. For setting stable thresholds of speech rejection and speaker verification, normalization methods for the likelihood values were investigated.

The proposed method was evaluated for various numbers of phoneme classes (1, 6, 15 and 25) using a 15-talker database recorded over several months. The rate for rejecting speech that differs from the prompted

key text increases as the number of phoneme classes increases. For 25 phoneme classes, our method could reject 80.7% of incorrect speech samples. The speaker verification rate was 99.9%. The error rates for speaker identification were constant, independent of the number of phoneme classes.

The following topics remain to be investigated:

- similarity measures to more effectively use the phonemes that have relatively stable individual information,
- segmental approach to use dynamic individual information.

ACKNOWLEDGMENT

The authors wish to acknowledge the members of the Furui Research Laboratory of NTT Human Interface Laboratories for their valuable and stimulating discussions, and are grateful to Dr. A.E. Rosenberg at AT&T Bell Laboratories for suggesting we read paper [3].

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