

論文 / 著書情報
Article / Book Information

Title	Free-Text Speaker Recognition Methods Using Phoneme Class Models
Author	Sadaoki Furui, Tomoko Matsui
Journal/Book name	Proc. 1st IEEE Workshop on Interactive Voice Technology for Telecommunications Applications, Vol. , No. , SessionV
発行日 / Issue date	1992, 10
権利情報 / Copyright	(c)1992 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.

FREE-TEXT SPEAKER RECOGNITION METHODS USING PHONEME CLASS MODELS

Sadaoki Furui and Tomoko Matsui

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

ABSTRACT

This paper proposes three speaker recognition methods in which a new and free key-sentence can be used at every recognition. The recognition system prompts each user with a new sentence every time and accepts the input utterance only when it decides that the registered speaker uttered the prompted sentence. These methods use speaker-specific phoneme class HMMs as basic acoustic units, and concatenate these models to create a sentence HMM. The likelihood of input speech against the sentence model is used for the speaker recognition decision. Experimental results confirmed the effectiveness of one of the methods proposed in this paper.

1. INTRODUCTION

Protecting the security of information services using telephone networks will require systems that can recognize the voices of users.

Speaker recognition methods can be classified into text-dependent methods, which use pre-determined key-words or sentences, and text-independent methods, which use arbitrary words or sentences. The text-dependent methods are usually based on template matching techniques in which the time axes of an input speech sample and each reference template or reference model of registered speakers are aligned, and the similarity between them is calculated. The structure of text-dependent recognition systems is therefore rather simple. These systems are also accurate because they exploit phoneme- or syllable-dependent voice individuality.

A text-dependent system can easily be cheated, however, because someone who records the voice of a registered speaker uttering key-words or sentences and then plays it back in front of the microphone can be accepted as the registered speaker. To cope with this problem, there are methods in which a set of words, such as digits, are used as key-words and each user is requested

to utter a given sequence of key-words that is randomly changed every time the system is used [1][2][3]. But even this method is not reliable enough, since it can be cheated by using advanced electronic recording equipment that can reproduce key-words in a requested order.

This paper therefore proposes methods in which completely free key-sentences can be used at every recognition. The system accepts the input utterance only when it decides that the registered speaker correctly uttered the prompted sentence. Because the vocabulary is unlimited, prospective imposters cannot know in advance the sentence they will be prompted to say. These methods are facilitated by using speaker-specific phoneme class models as basic acoustic units. This method can not only accurately recognize speakers but can also reject utterances whose text differs from the prompted text, even if it is uttered by the registered speaker. A recorded and played back voice can thus be correctly rejected.

One of the major issues in these methods is how to properly create speaker-specific phoneme class models when using training utterances of a limited size. The following sections introduce variations of these methods.

2. BASIC STRUCTURES OF THE METHODS

2.1 Method I [4]

Figure 1 shows a block diagram of Method I. First, training utterances for each speaker are used to create a one-state Gaussian mixture HMM as the text-independent speaker model [5]. Then the same training utterances are used to create speaker-specific phoneme class models. The number and content of phoneme

classes are assumed to be determined beforehand. Because the text of training utterances is known, these utterances can be modeled as the concatenation of phoneme class models that can be automatically trained by an iterative algorithm[6][7]. The text-independent speaker model is used as the initial condition of all the phoneme class models, and only the weighting factors of Gaussian mixtures for each model are trained. Each phoneme class model is therefore also represented by a one-state Gaussian mixture HMM.

In the recognition stage, the system prompts a text, which could be different from those used for training, to be spoken. The speaker is requested to utter the prompted text, and at the same time the system concatenates the phoneme class models of each registered speaker according to the text to create the sentence HMM model. The likelihood of the input speech against the sentence model is calculated for each registered speaker, and the similarity of both voice and speech material is evaluated from the likelihood value. If the similarity from these two points of view is high enough, the speaker is accepted as the registered speaker.

A vector quantization codebook instead of the

Gaussian mixture HMM can also be used for representing speaker characteristics.

2.2 Method II

With Method I, phoneme class models are made by using only training utterances by each speaker. The number of phoneme classes is therefore limited by the size of training utterances. In Method II (Fig. 2), phoneme models for each speaker are made by adapting universal models to each speaker's voice. These universal models are originally created for speaker-independent speech recognition.

As with Method I, training utterances by the speaker are first used to create a one-state Gaussian mixture HMM as the text-independent speaker model. Then speaker-independent phoneme models, which are typically represented by Gaussian mixture continuous HMMs, are adapted to each speaker by using the text-independent speaker model. The adaptation can also be done directly by using the speaker's training utterances. After the speaker-adapted phoneme models are made, the same speaker recognition method as that used in Method I can be applied. Thus Method II uses more precise phoneme models than Method I, and is, therefore, ex-

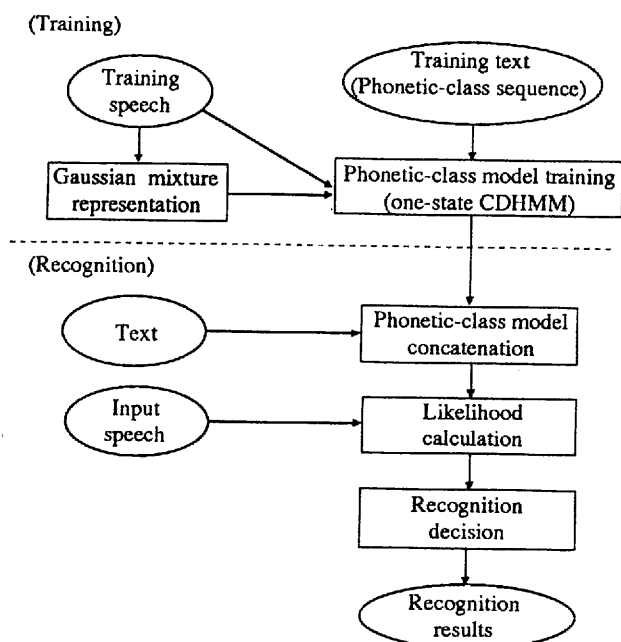


Fig. 1 - Block diagram of Method I.

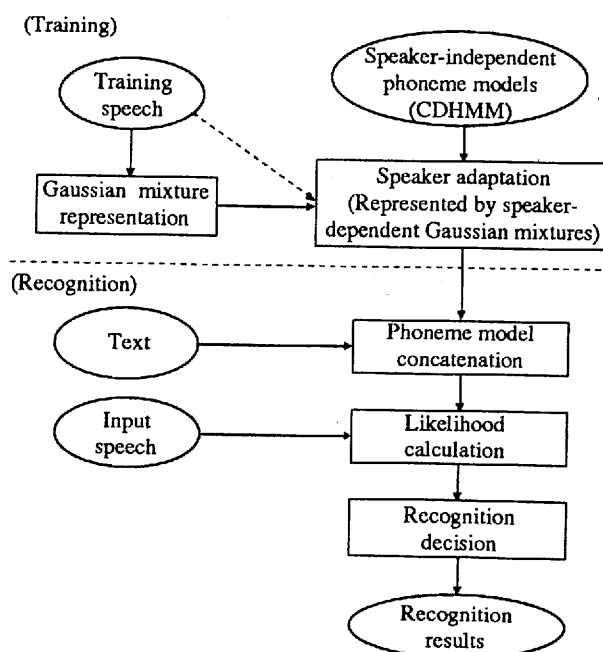


Fig. 2 - Block diagram of Method II.

pected to produce better results, especially for rejecting utterances different from prompted texts.

The Method II can also model the speaker by using the vector quantization codebook instead of the Gaussian mixture HMM.

2.3 Method III

In Methods I and II, phoneme or phoneme class models for each speaker are used commonly for calculating the likelihood of both voice and speech material. Method III (Fig. 3), in contrast, uses different models to make the decisions of voice and speech material, and it makes the final decision by combining these results.

First, training utterances by each speaker are used to create a text-independent speaker model, represented by a one-state Gaussian mixture HMM, for each speaker. At the same time, a sentence model is made by concatenating speaker-independent phoneme models (e. g., continuous HMMs using Gaussian mixture distribution) according to the input speech text. In the recognition stage, the likelihood values of input speech against each speaker model and the

sentence model are calculated separately. These two likelihood values are combined logically or numerically for the final decision. These values can be combined either (a) by calculating weighted average of these likelihood values, or (b), only when the likelihood against the sentence model is higher than a threshold, by performing speaker recognition according to the likelihood against the speaker model.

3. RESULTS OF EXPERIMENTS USING METHOD I

3.1 Experimental Conditions [4]

Recognition experiments were performed to evaluate the effectiveness of Method I. Various sentences uttered by 15 speakers (10 male and 5 female) at three sessions over a period of roughly five months were used. We call these sessions, in order, Sessions A, B, and C. The utterances at the first session (A) were used for training and those at the other two sessions (B and C) were used for testing. The time intervals between training and testing utterances were four months and five months. The average sentence length was about 4 s.

LPC cepstral coefficients (16th order) were used as feature parameters. A text-independent speaker model represented by a diagonal covariance Gaussian mixture HMM with one state and 64 mixtures was made for each speaker by using training utterances consisting of 10 sentences. Then the same training utterances were used to create phoneme class models for each speaker. The number of phoneme classes was either 1, 6, 15, or 25.

Five kinds of sentences uttered at each of two sessions were used for testing. The total number of test utterances was 150 (5 sentences $\times$ 15 speakers $\times$ 2 sessions). One of the five sentences was used in turn as the prompted sentence. The likelihood value for each test utterance against the sentence model, which was built by concatenating the speaker-specific phoneme class models in accordance with the prompted text, was calculated. Utterances that were spoken by the same speaker but whose speech material differed from the prompted text was to have been rejected. Duration control for

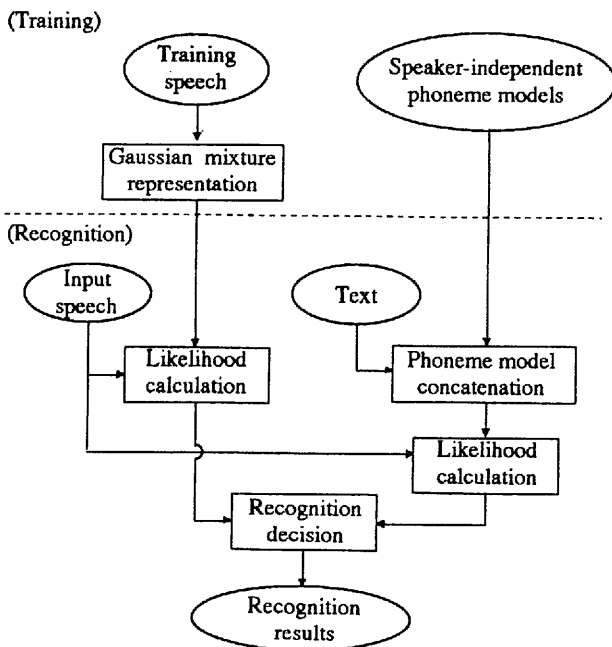


Fig. 3 - Block diagram of Method III.

each phoneme class model was not performed in the likelihood calculation.

The threshold likelihood values for rejecting the utterances whose speech material differed from the prompted material was set a posteriori at the maximum value that assured that the correct utterances were not rejected. Because the likelihood value can be widely varied according to the text of an utterance to be recognized and to the session-to-session variation of speech, it is difficult to set a universal threshold that can be applied to the utterances of different speech material spoken at different sessions. It is therefore necessary to introduce a technique for normalizing the likelihood values. We have tried to normalize them by the value obtained without text constraints -that is, by the value obtained by using the text-independent speaker model. In other words, the likelihood value was normalized by the value obtained by directly using the Gaussian mixture of the speaker.

3.2 Experimental Results [4]

(1) Speaker Verification and Identification

The threshold for speaker verification decision was set a posteriori for each registered speaker to equalize the two kinds of errors: rejection of the correct speaker, and acceptance of imposters. The results listed in Table 1 show that the mean error rate could be reduced by about 20% by using 15 or 25 phoneme classes and that very low verification error rates could be obtained by this method. Experimental results for speaker identification indicated that the mean identification error rate was 4.7% and that this error rate was not affected by the number of phoneme

Table 1. Verification error rates.

	Number of phonetic 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%

classes.

(2) Rejection of Incorrect Speech - Without Normalization -

The percentage of rejecting the utterances by the true speaker but whose texts differed from the prompted text listed in Table 2 indicate that the rejection performance improves as the number of phoneme classes increases. But even when using 25 classes, this method rejected only about half of the utterances that should have been rejected.

Table 2. Different sentence rejection rates (without likelihood normalization).

	Number of phonetic classes		
	6	15	25
Session B	41.3%	63.7%	69.0%
Session C	37.7%	54.0%	60.7%
Session B + Session C	24.3%	39.7%	48.3%

(3) Effects of Likelihood Normalization

The average rejection rates listed in Table 3 show that the rejection performance is improved by likelihood normalization. Under the condition that the threshold is set a posteriori so that the utterances of correct texts are not rejected, roughly 80% of the utterances of incorrect texts can be rejected. The rejection error is reduced to roughly half of that obtained without normalization.

Table 3. Rate of rejecting incorrect sentences.

	Normalization		Improvement
	No	Yes	
Session B	69.0%	82.0%	13.0%
Session C	60.7%	81.3%	20.6%
Session B + Session C	48.3%	80.7%	32.4%

4. CONCLUSION

This paper has proposed methods that reliably recognize speakers by using arbitrary texts that can be designated by the system for each instance of recognition. Recognition experiments using one of these methods confirmed that this method achieves very high speaker recognition performances even when the texts of utterances are changed at every instance of recognition. These experiments also confirmed that the utterances differing from prompted texts can reliably be rejected even if they are uttered by registered speakers.

The methods proposed in this paper can therefore solve one of the problems of conventional text-dependent speaker recognition methods: vulnerability to cheating the recognition system by using recorded and played-back voices.

The following topics remain to be investigated:

- (a) Evaluation of Methods II and III,
- (b) Distance (similarity) measures to more effectively use the phonemes that have relatively large and stable individual information,
- (c) Segmental approach to use dynamic individual information instead of frame-by-frame information,
- (d) Phoneme modeling that can more precisely represent speaker-specific information and context-dependent spectral variation,
- (e) Improvement of recognition performances. For example, by incorporating duration control for each phoneme HMM.

ACKNOWLEDGMENTS

We thank the members of Furui Research Laboratory for their valuable discussions.

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] T. Matsui and S. Furui, "Speaker recognition using concatenated phoneme HMMs," *Proc. Int. Conf. Spoken Language Processing*, Banff (1992).
- [5] T. Matsui and S. Furui, "Comparison of text-independent speaker recognition methods using VQ-distortion and discrete/continuous HMMs," *Proc. IEEE Int. Conf. Acoust. Speech, Signal Processing*, San Francisco, pp. II-157-160 (1992).
- [6] K. F. Lee and H. W. Hon, "Large-vocabulary speaker-independent continuous speech recognition using HMM," *Proc. IEEE Int. Conf. Acoust. Speech, Signal Processing*, New York, pp. 123-126 (1988).
- [7] K. Maruyama, T. Hanazawa, and K. Shikano, "English word recognition using HMM phone concatenated training," *IEICE Technical Report*, SP88-119 (1989).