論文 / 著書情報
Article / Book Information

Title	Comparison of Text-Independent Speaker Recognition Methods Using VQ-Distortion and Discrete/Continuous HMM's
Author	Tomoko Matsui, Sadaoki Furui
Journal/Book name	IEEE Transactions on Speech and Audio Processing, Vol. 2, No. 3, pp. 456-459
発行日 / Issue date	1994, 7
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.

Comparison of Text-Independent Speaker Recognition Methods Using VQ-Distortion and Discrete/Continuous HMM's

Tomoko Matsui and Sadaoki Furui

Abstract—**This paper compares a VQ (vector quantization)-distortion-based speaker recognition method and discrete /continuous ergodic HMM (hidden Markov model)-based ones, especially from the viewpoint of robustness against utterance variations. We show that a continuous ergodic HMM is as robust as a VQ-distortion method when enough data is available and that a continuous ergodic HMM is far superior to a discrete ergodic HMM. We also show that the information on transitions between different states is ineffective for text-independent speaker recognition. Therefore, the speaker recognition rates using a continuous ergodic HMM are strongly correlated with the total number of mixtures irrespective of the number of states.**

I. Introduction

For text-independent speaker recognition, VQ-based methods [1]–[2] were proposed many years ago. In recent years, HMM-based methods have become popular for speech recognition and have also been applied to speaker recognition [3]–[7]. However, the effectiveness of HMM-based speaker recognition methods in comparison with VQ-based methods has not been made clear.

Our recent study [8] reported a VQ-based method that is robust against utterance variations even when only a short utterance is available. Rosenberg [3] has reported a method using left-to-right HMM's, and other studies [4]–[5] have proposed using linear predictive ergodic HMM's. Rose [6] has examined the effects of the number of mixture components in a single state HMM on speaker recognition performance. Savic and Gupta [7], on the other hand, examined speaker verification by comparing test samples and the reference vectors assigned to each state of an ergodic HMM. Until now, an ergodic HMM has been assumed to be effective for text-independent speaker recognition because it automatically forms broad phonetic classes corresponding to each state. However, few studies have directly used the likelihood of an ergodic HMM, and none have yet examined the difference in performance between discrete and continuous HMM's in text-independent speaker recognition. Although Tishby [5] has reported differences between the performance of VQ-distortion and linear predictive ergodic HMM's for digit utterances, the difference between VQ-distortion and regular ergodic HMM's has not yet been analyzed.

This paper compares a VQ-distortion-based speaker recognition method and discrete/continuous ergodic HMM-based ones, especially from the viewpoint of robustness against utterance variations. As examples of utterance variations, sentences uttered at different speeds and recorded on several sessions were used.

II. Methods

In speaker recognition using VQ-distortion [2], VQ codebooks are created for each reference speaker. As shown in Fig. 1, input speech frames are vector-quantized using the codebooks of reference speakers, and the VQ-distortion values accumulated over all frames are used to identify or verify the speaker (the recognition decision).

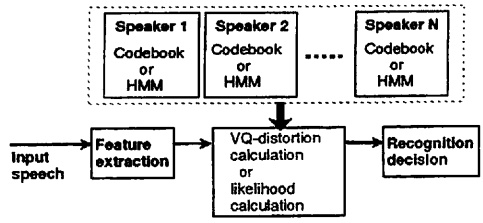

Fig. 1. Speaker recognition procedure.

In the ergodic HMM approach, on the other hand, a speaker-dependent ergodic HMM is first made for each reference speaker by estimating the HMM parameters using the Baum-Welch algorithm. As in the VQ-distortion approach, the accumulated likelihood of an ergodic HMM for input speech frames is used for the recognition decision. The work reported here used fuzzy-vector-quantization-based discrete models [10] as discrete HMM's, and mixture-Gaussian HMM's with diagonal covariance matrices as continuous HMM's (Fig. 1). In the former case, the probability for each codebook was smoothed by the fuzzy-vector-quantization technique to cope with the problem of quantization errors.

III. Experiments

A. Experimental Conditions

The database consisted of sentence data uttered at three speaking rates (normal, fast, and slow) by 23 male and 13 female talkers. As one example of utterance variations, sentences uttered at different speeds were used. The sentences were selected from phonetically balanced sentences [9] and were read. This database was recorded on three sessions over six months and was recorded in the same recording room using the same microphone for all speakers for all sessions. The sampling rate was 12 kHz. 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. (The speech power was not retained.) Ten sentences uttered at normal rate in one session were used for training. The utterances recorded in two other sessions were used for testing. The combination of training and testing sessions was rotated. 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 in each session consisted of five sentences uttered at normal, fast, and slow rates. The sentences for testing were different from those for training, and were the same for all speakers and all recording sessions. Each sentence was evaluated individually for testing. The average durations of each class of sentences uttered at normal, fast, and slow rates were 4.2 s, 3.2 s, and 5.8 s.

The performances of VQ-distortion–and ergodic HMM-based methods were evaluated by the speaker identification and verification rates. In speaker identification using the VQ-distortion approach the speaker who had the minimum distortion was selected from the registered speakers, while using the HMM approach the speaker having the maximum likelihood was selected. In speaker verification, the threshold was used to accept or reject a speaker, and the speaker was accepted only when the VQ-distortion was smaller than the threshold in the VQ approach and when the likelihood was bigger than the threshold in the HMM approach. The threshold was set a

Manuscript received September 15, 1992; revised December 3, 1993. The associate editor coordinating the review of this paper and approving it for publication was Prof. Huseyin Abut.

The authors are with NTT Human Interface Laboratories, Musashino-Shi, Tokyo 180, Japan.

IEEE Log Number 9400758.

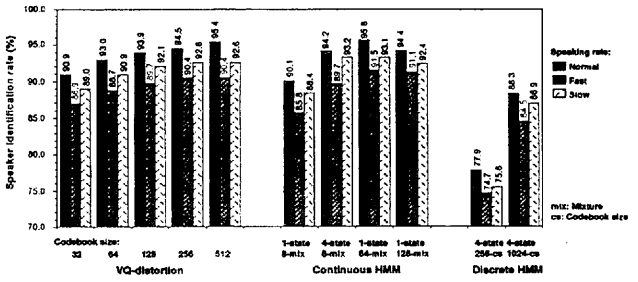

Fig. 2. Speaker identification rate.

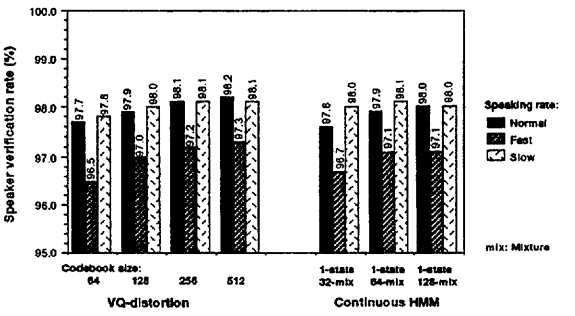

Fig. 3. Speaker verification rate.

posteriori to equalize the probability of false acceptance and false rejection, and was set for individual speakers. The verification was performed by using one speaker as the customer and the other 35 speakers as impostors, rotating through all speakers and then averaging the results.

In the experiments using VQ-distortion, the codebook size was varied from 32 to 512. In the experiments using discrete ergodic HMM's, the codebook size was varied from 256 to 1024 and the number of states was set at 4. In the experiments using continuous ergodic HMM's, the number of mixture Gaussian distributions was varied from 8 to 128 and the number of states was varied from 1 to 8.

The generalized Lloyd algorithm (LBG algorithm) was used for creating VQ codebooks. For the fuzzy-vector-quantization-based discrete HMM's, the fuzziness value to control the smoothness was set to 1.5 and the number of nearest neighbors used for smoothing was set to 5. HMM parameters were initialized as follows. For the discrete HMM, the length of each training sample was divided by the number of states and assigned to each state. The output probabilities were initialized using histograms of the codewords for each state. For continuous HMM's, the length of each training sample was divided by the total number of mixtures (the number of states times the number of mixtures assigned to each state), and the mean and covariance values of samples assigned to each mixture were calculated. Each state was, therefore, initialized without any special information about broad phonetic classes. Two transition probabilities derived from the same state were initialized identically. Two arcs derived from the same state had the same output probabilities.

B. Speaker Identification

Fig. 2 shows the results of speaker identification experiments. For continuous HMM's, the total number of mixtures was varied from 8 to 128 and the number of states was one or four. Since the number of states was fixed to four for discrete HMM's, the results for continuous HMM's include the case of four states in order to clearly compare the discrete and continuous HMM-based methods. In these experiments, a continuous HMM with one state and 64 mixtures performed best, especially for test data that was uttered at a fast rate. The performance of the continuous ergodic HMM was about the same as the VQ-distortion method and was much higher than the discrete ergodic HMM. From the viewpoint of the number of model parameters, the continuous ergodic HMM outperformed the VQ-distortion method, since a 128 VQ codebook has about as many parameters as a 64-mixture Gaussian mixture with diagonal covariance.

For the VQ-distortion-based method, the identification rate increased as the codebook size increased. For the continuous HMM-based method, a 1-state 128-mixture HMM was worse than a 1-state 64-mixture HMM. This was because the difficulty of estimating the variance of each Gaussian distribution increased when the number of mixtures was too large. For the discrete HMM-based method, when the codebook size is greater than 1024, the identification rates may be higher, but the amount of training data and the calculation becomes enormous.

As for speaking rates, the recognition rates were highest for test data uttered at normal rate, and lowest for data uttered at a fast rate.

C. Speaker Verification

Fig. 3 shows the results of speaker verification experiments. The verification rate was defined as the probability of true acceptance and true rejection at the equal error operating point. In these experiments, VQ-distortion using a codebook size of 512 performed best, although there was no statistically significant difference between any of the verification results.

The verification rates were higher for test data uttered at normal and slow rates than for test data uttered at a fast rate. This shows that the test vector distribution uttered at a fast rate deviates from those uttered at normal and slow rates.

D. Robustness Against Different Amounts of Training Data

The performance of VQ-distortion and continuous HMM's for different amounts of training data was also investigated. Fig. 4 shows the results of speaker identification experiments using two different training sets: one training set consisted of the ten sentences used in the experiments reported in the previous sections, and the other training set consisted of five sentences selected out of the ten sentences. The VQ codebook size was 256 or 512 in the VQ-distortion method. The continuous HMM's had one state and 16, 32, or 64 mixtures. Fig. 4 indicates that, when only five sentences were used for training, the performance of the VQ-distortion method was much better than that of the continuous HMM method, although the performances of both methods were almost the same when ten sentences were used for training. Identifying speakers using continuous HMM's needs more training data. The figure also indicates that when the amount of training data was small, the results for 32 mixtures were relatively better than those for 64 mixtures. This is probably because it is difficult to estimate the continuous HMM parameters when the amount of available data is small.

For the VQ-distortion-based method, the identification rate increased as the codebook size increased up to 512, when ten sentences were used for training. When five sentences were used for training, the performance was almost saturated at a codebook size of 256. The average number of the training vectors per cluster at the saturation level was 11. For the continuous HMM-based method, a 1-state 64-mixture HMM performed best using ten sentences for training. When five sentences were used for training, the performance was almost saturated for the 1-state 32-mixture condition. The average number of training vectors per mixture component at the saturation level was 89. These experimental results indicate that the amount of training

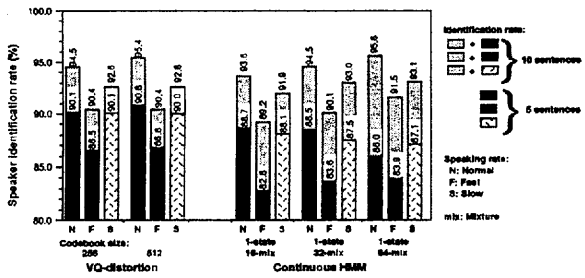

Fig. 4. Speaker identification rates with different amounts of training data.

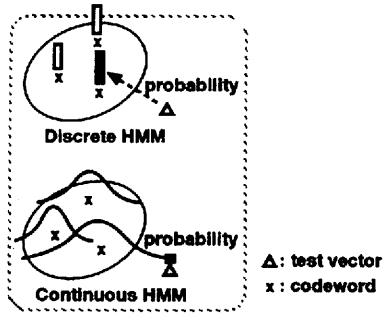

Fig. 5. Illustration of discrete HMM versus continuous HMM.

data per mixture for creating continuous HMM's needs to be eight times more than that per cluster for creating VQ codebooks. It may be useful to clip the variance values to prevent bad estimates when the amount of training data is small.

IV. Discussion

A. Difference Between Discrete And Continuous Ergodic HMM's

Let us consider the difference in performance between discrete and continuous ergodic HMM's. In a discrete ergodic HMM, the output probability of each test vector is set to the output probability of the nearest VQ codebook vector as shown in Fig. 5. Our speaker recognition experiments were text-independent and used utterances recorded on several sessions. Since the training and test vectors have session-to-session, text-dependent, and speaking-rate-dependent variations, the test vector distribution deviates from the training vector distribution. Even in such a case, every input vector is assigned the output probability of the nearest codebook vector in the discrete HMM method. If a significant number of test vectors have high output probability of the nearest VQ codebook vector associated with a different speaker, the recognition is poor. With a continuous ergodic HMM, the output probability of such a test vector is low because it corresponds to the tail of the Gaussian distribution. Here, a continuous ergodic HMM is therefore superior to a discrete ergodic HMM.

B. Performance of Continuous HMM's With Different Numbers of States And Mixtures

Speaker identification experiments were also carried out using continuous ergodic HMM's with different numbers of states and mixtures. For all speaking rates, the identification rate increased as the number of states and mixtures increased (Fig. 6). The identification rates were highly correlated with the total number of mixtures (the number of states times the number of mixtures assigned to each state). These results indicate that information on transitions between different states is not effective for text-independent speaker

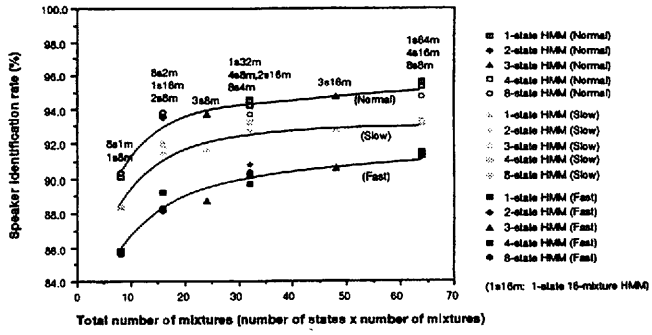

Fig. 6. Speaker identification rates as functions of the numbers of states and mixtures.

recognition. All the transition probabilities between different states in these experiments were between 0.1 and 0.2. The identification rates were almost saturated when 32 or more mixtures were used except for the fast rate case.

V. Conclusion

This paper compared text-independent speaker recognition methods that use VQ-distortion and discrete/continuous ergodic HMM's. Continuous ergodic HMM's identified speakers much more accurately than discrete ergodic HMM's did. The continuous HMM's are as resistant to session-to-session variations in speech and to those due to different utterance rates as the VQ-distortion-based method. However, when little data is available, the VQ-distortion-based method is more robust than a continuous HMM-based method.

With continuous ergodic HMM's, the speaker identification rates are strongly correlated with the total number of mixtures, irrespective of the number of states. This means that the information about transitions between different states does not contribute to text-independent speaker recognition.

Future research items include methods that effectively use phoneme class information and also use Δcepstrum features in combination with cepstrum features.

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] K. P. Li and E. H. Wrench Jr., "An approach to text-independent speaker recognition with short utterances," in *Proc. ICASSP*, 1983, pp. 555–558.
[2] F. K. Soong *et al.*, "A vector quantization approach to speaker recognition," in *Proc. ICASSP*, 1985, pp. 387–390.
[3] A. E. Rosenberg *et al.*, "Connected word talker verification using whole word hidden Markov models," in *Proc. ICASSP*, 1991, pp. 381–384.
[4] A. B. Poritz, "Linear predictive hidden Markov models and the speech signal," in *Proc. ICASSP*, 1982, pp. 1291–1294.
[5] N. Z. Tishby, "On the application of mixture AR hidden Markov models to text independent speaker recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, pp. 563–570, 1991.
[6] R. C. Rose and D. A. Reynolds, "Text independent speaker identification using automatic acoustic segmentation," in *Proc. ICASSP*, 1990, pp. 293–296.
[7] M. Savic and S. K. Gupta, "Variable parameter speaker verification system based on hidden Markov modeling," in *Proc. ICASSP*, 1990, pp. 281–284.
[8] T. Matsui and S. Furui, "A text-independent speaker recognition method robust against utterance variations," in *Proc. ICASSP*, 1991, pp. 377–380.

[9] H. Kuwabara *et al.*, "Construction of ATR Japanese speech database as a research tool," ATR Tech. Rep. TR-I-0086, 1989.
[10] H.-P. Tseng, M. J. Sabin, and E. A. Lee, "Fuzzy vector quantazation applied to hidden Markov modeling," in *Proc. ICASSP*, 1987, pp. 641–644.