T2R2 東京科学大学 リサーチリポジトリ
Science Tokyo Research Repository

論文 / 著書情報
Article / Book Information

Title	Comparison of Text-Independent Speaker Recognition Methods Using VQ-Distortion and Discrete/Continuous HMMs
Author	Tomoko Matsui, Sadaoki Furui
Journal/Book name	IEEE ICASSP1992, Vol. , No. , pp. 157-160
発行日 / Issue date	1992,
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.

Powered by T2R2 (Science Tokyo Research Repository)

COMPARISON OF TEXT-INDEPENDENT SPEAKER RECOGNITION METHODS USING VQ-DISTORTION AND DISCRETE/CONTINUOUS HMMS

Tomoko Matsui *Sadaoki Furui*

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

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 identification rates using a continuous ergodic HMM are strongly correlated with the total number of mixtures irrespective of the number of states. It is also found that, for continuous ergodic HMM-based speaker recognition, the Distortion-Intersection Measure (DIM), which was introduced as a VQ-distortion measure to increase the robustness against utterance variations, is effective.

1 INTRODUCTION

For text-independent speaker recognition, VQ-based methods [1]–[2] were proposed many years ago. In recent years, HMM-based methods [3]–[6] have become popular for speech recognition and have also been applied to speaker recognition. However, the effectiveness of HMM-based speaker recognition methods has not been made clear.

Our recent study [7] 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 HMMs, and other studies [4]–[5] have proposed using linear predictive ergodic HMMs. Savic and Gupta [6], 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, even though few studies have directly used the likelihood of an ergodic HMM, and none have yet examined the difference in performance between discrete and continuous HMMs in text-independent speaker recognition. Although Tishby [5] has reported differences between the performance of VQ-distortion and linear predictive ergodic HMMs for digit utterances, the difference between VQ-distortion and regular ergodic HMMs 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.

2 METHODS

In speaker recognition using VQ-distortion [2], VQ codebooks are created for each reference speaker. As shown in Figure 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).

In the ergodic HMM approach, on the other hand, a speaker-dependent ergodic HMM is first made for each reference speaker and the HMM parameters are estimated using the Baum-Welch algorithm, and then the accumulated likelihood of an ergodic HMM for input speech frames is used for recognition decision. The work reported here uses fuzzy-vector-quantization-based discrete models as discrete HMMs, and it uses mixture-Gaussian HMMs with diagonal covariance matrices as continuous HMMs [8] (Figure 1).

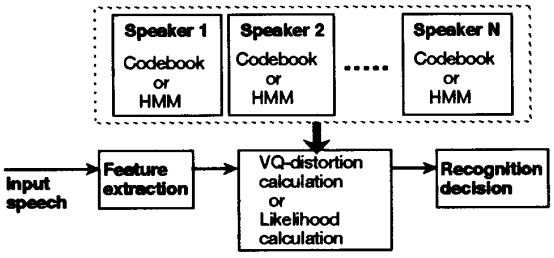


Figure 1. Speaker recognition procedure.

3 EXPERIMENTS

3.1 Experimental conditions

The database consisted of sentence data uttered at three speeds (normal, fast, and slow) by 23 male and 13 female talkers. This database was recorded on three occasions 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 uttered at normal speed on one occasion were used for training, and five sentences uttered at normal, fast, and slow speeds on the other two occasions were used for testing. The duration of each sentence was about 4 s.

0-7803-0532-9/92 $3.00 © 1992 IEEE

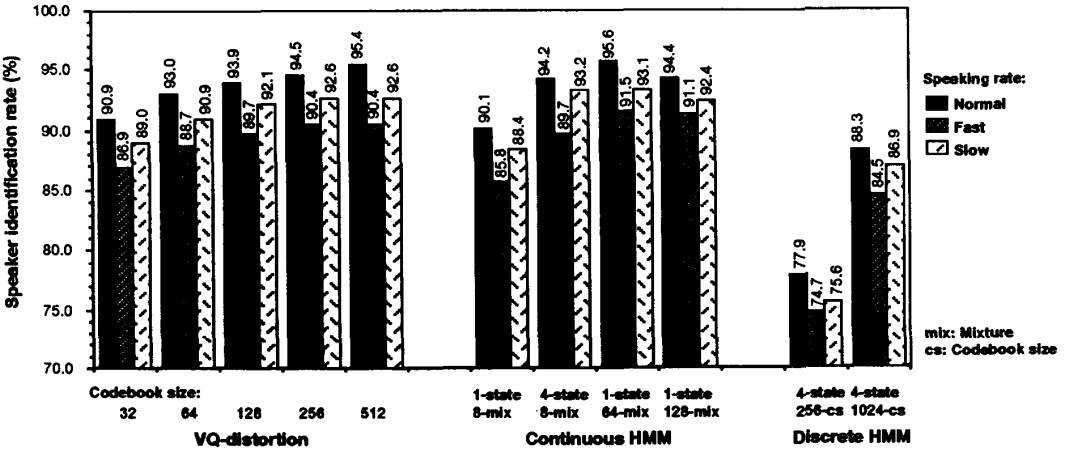


Figure 2. Speaker identification rate (%).

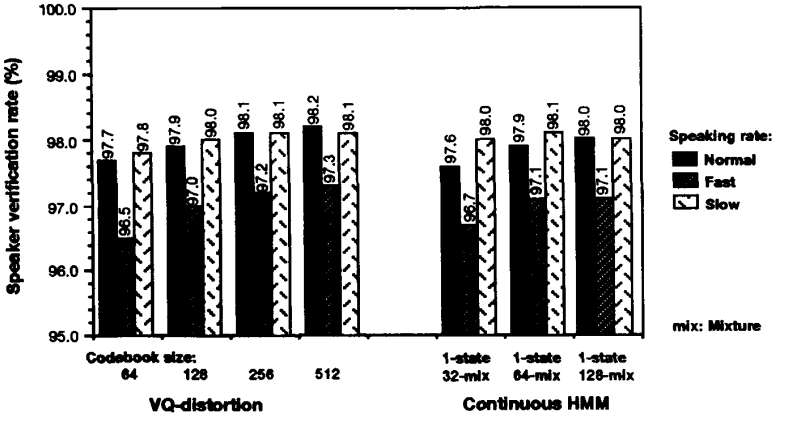


Figure 3. Speaker verification rate (%).

LBG algorithm was used for creating VQ codebooks. HMM parameters were initialized as follows. For discrete HMMs, output probabilities were initialized using histograms of codewords. For continuous HMMs, 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 each segment were calculated. Two transition probabilities derived from the same state were initialized identically. Two arcs derived from the same state had the same output probabilities.

3.2 Results

Figures 2 and 3 show the results of speaker recognition experiments. They indicate that a continuous ergodic HMM is as robust as a VQ-distortion method against utterance variations, and it is far superior to a discrete ergodic HMM. In a VQ-distortion method, the codebook size of 256 is enough for speaker recognition under the experimental conditions.

4 DISCUSSION

4.1 Difference between discrete and continuous ergodic HMMs

Let us consider the difference in performance between discrete and continuous ergodic HMMs. 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 Figure 4. In text-independent speaker recognition using a short utterance with intrinsically wide variability, the test vector distribution deviates from the training vector distribution. In such a case, if there is a significant number of test vectors for which the output probability of the nearest VQ codebook vector associated with a different speaker is high, 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.

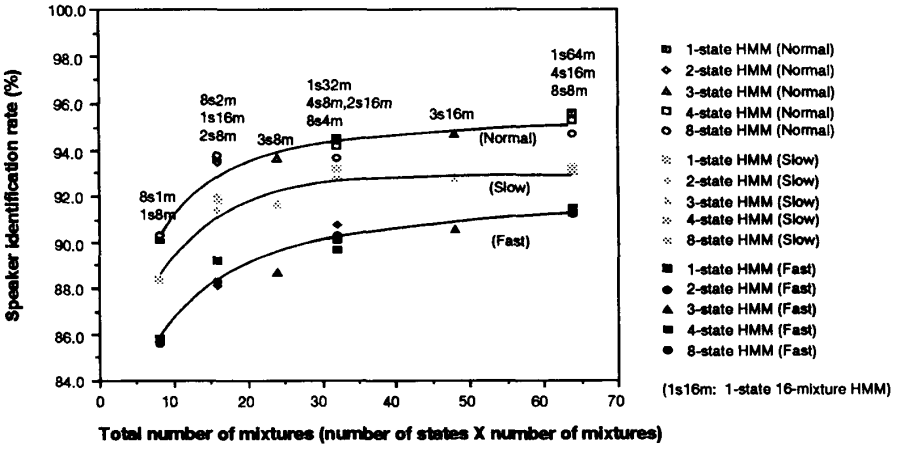


Figure 5. Speaker identification rates as functions of the numbers of states and mixtures.

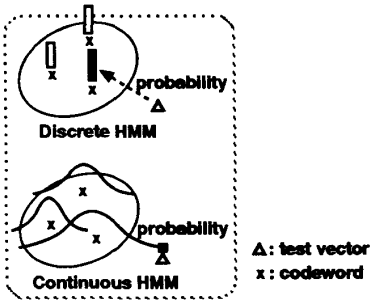


Figure 4. Illustration of discrete HMM *vs.* continuous HMM.

As shown in Figure 2, in a discrete HMM, when the codebook size is bigger than 1024, the identification rates may be higher, but the amount of training data and calculation becomes enormous.

4.2 Performance of continuous HMMs with different numbers of states and mixtures

Speaker identification experiments were also carried out using continuous ergodic HMMs with different numbers of states and mixtures. For all utterance speeds, the identification rate increased as the number of states and mixtures increased (Figure 5). The identification rates are highly correlated with the total number of mixtures (the number of states times the number of mixtures assigned to each state). The identification rates using 32 mixtures are almost saturated except for the fast speed case. These results indicate that information on transitions between different states is not effective for text-independent speaker recognition. All the transition probabilities between different states in these experiments were between 0.1 and 0.2.

4.3 Robustness against different amounts of training data

The performance of VQ-distortion and continuous HMMs for different amounts of training data was also investigated. Figure 6 shows the results of speaker identification experiments using two different training sets: one training set consisted of the 10 sentences used in the experiments reported in the previous sections, and the other training set consisted of 5 sentences selected out of the 10 sentences. The VQ codebook size was 256 or 512 in the VQ-distortion method. The continuous HMMs had one state and 16, 32, or 64 mixtures. Figure 6 indicates that identifying speakers using continuous HMMs needs more training data, and also indicates that when the amount of training data is small, the results for 32 mixtures are much better than those for 64 mixtures. This is probably because the estimation of the continuous HMM parameters is difficult when the amount of available data is small, so the identification rates for a continuous HMM-based method become lower than those for a non-parametric method such as a VQ-distortion-based method.

4.4 Effect of applying the DIM to continuous HMMs

In a VQ-distortion method, the Distortion-Intersection Measure (DIM) [7] is characterized by selective matching using only a stable subset of test vectors in the distortion calculation. The stable subset is defined as the intersection space between a set of test vectors and a set of VQ codebook vectors. The intersection space is determined by using the scope of VQ codebook vectors. If, as shown in Figure 7 (a), a test vector is not included in the scope of the nearest VQ codebook vector, the quantization distortion is set to the boundary value of the scope. With continuous HMMs, the idea of DIM is implemented by flattening the tail of each Gaussian distribution. If a test vector corresponds to the tail of the Gaussian distribution, the output probability is set to the flattening value (Figure 7 (b)). In this paper, the flattening was experimentally started at the value of 3 σ for each Gaussian distribution.

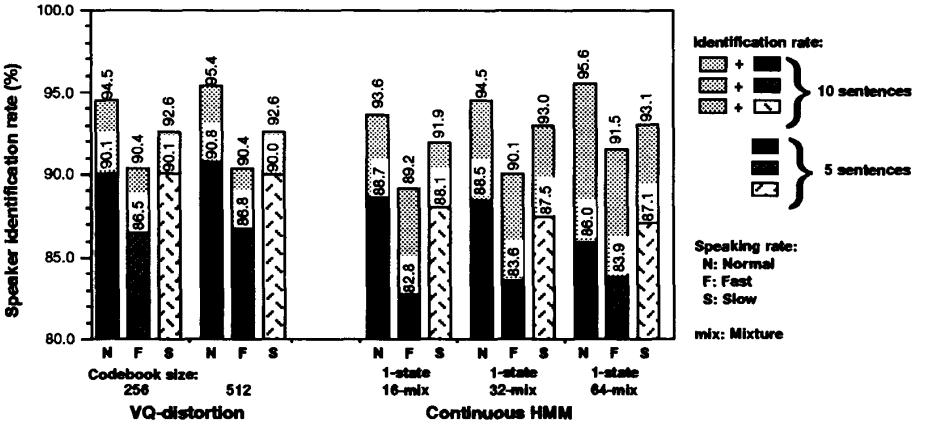


Figure 6. Speaker identification rates with different amounts of training data.

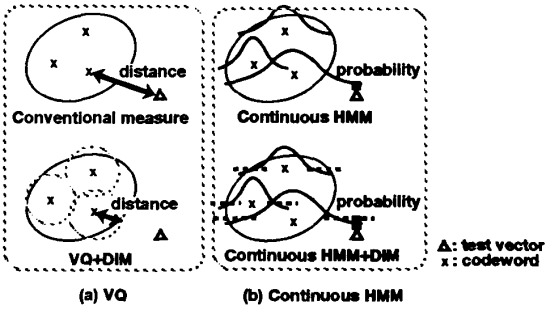


Figure 7. Illustration of applying the DIM to (a) VQ and (b) continuous HMM.

Table 1 lists the results of speaker identification experiments using DIM in the framework of a continuous HMM. HMMs had one state and 32 mixtures in the experiments. These results indicate that DIM can be applied to continuous HMMs effectively.

Table 1. Speaker identification rates applying the DIM.

speaking rate	continuous HMM 1-*state* 32-*mixture*	
	-	+DIM
normal	88.5	89.1
fast	83.6	84.8
slow	87.5	88.6

5 CONCLUSION

This paper has compared text-independent speaker recognition methods that use VQ-distortion and discrete/continuous ergodic HMMs. A continuous ergodic HMM is as robust as a VQ-distortion method against utterance variations and it is much better than a discrete ergodic HMM. The speaker identification rates using a continuous ergodic HMM are strongly correlated with the total number of mixtures irrespective of the number of states. The information on transitions between different states is not effective for text-independent speaker recognition. Moreover, when the amount of available data is small, a VQ-distortion method is more robust than a continuous HMM. The DIM method improves the continuous ergodic HMM-based speaker recognition.

ACKNOWLEDGMENT

The authors wish to acknowledge 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," *Proc. ICASSP, pp.555-558 (1983)*

[2] F.K. Soong et al., "A vector quantization approach to speaker recognition," *Proc. ICASSP, pp.387-390 (1985)*

[3] A.E. Rosenberg et al., "Connected word talker verification using whole word Hidden Markov Models," *Proc. ICASSP,pp.381-384 (1991)*

[4] A.B. Poritz, "Linear predictive Hidden Markov Models and the speech signal," *Proc. ICASSP, pp.1291-1294 (1982)*

[5] N.Z. Tishby, "On the application of Mixture AR Hidden Markov Models to text independent speaker recognition," *IEEE, Trans. ASSP, pp.563-570 (1991)*

[6] M. Savic and S.K. Gupta, "Variable parameter speaker verification system based on hidden Markov modeling," *Proc. ICASSP,pp.281-284 (1990)*

[7] T. Matsui and S. Furui, "A text-independent speaker recognition method robust against utterance variations," *Proc. ICASSP,pp.377-380 (1991)*

[8] L.A. Liporace., "Maximum likelihood estimation for multivariate observations of Markov process," *IEEE Trans. Inf. Theory, IT-28, 5, pp.729-734 (1982)*