論文 / 著書情報
Article / Book Information

Title	A Text-Independent Speaker Recognition Method Robust Against Utterance Variations
Author	Tomoko Matsui, Sadaoki Furui
Journal/Book name	IEEE ICASSP1991, Vol. , No. , pp. 377-380
発行日 / Issue date	1991, 5
権利情報 / Copyright	(c)1991 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 PROCEEDINGS OF THE ICASSP 91 — 1991 INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING, Toronto, Canada May 14-17, 1991

S6.3

A TEXT-INDEPENDENT SPEAKER RECOGNITION METHOD ROBUST AGAINST UTTERANCE VARIATIONS

Tomoko Matsui and Sadaoki Furui

NTT Human Interface Laboratories
Musashino-shi, Tokyo 180, Japan

ABSTRACT

This paper describes a VQ (vector quantization)-based text-independent speaker recognition method which is robust against utterance variations. Three key techniques are introduced to cope with temporal and text-dependent spectral variations. First, either an ergodic hidden Markov model (HMM) or a V/UV (voiced/unvoiced) decision is used to classify input speech into broad phonetic classes. Second, a new distance measure, Distortion-Intersection Measure (DIM), is introduced for calculating VQ distortion of input speech compared to speaker-dependent codebooks. DIM is characterized by selective matching using only a stable subset of test speech in the distortion calculation. Third, a new normalization method, Talker Variability Normalization (TVN), is introduced. TVN normalizes parameter variation taking both inter- and intra-speaker variability into consideration. TVN emphasizes feature parameters that have relatively large inter-speaker variability and small intra-speaker variability. The system is tested using utterances of nine speakers recorded over three years. The combination of the three techniques achieves high speaker identification accuracies of 98.5% using only vocal tract information, and 99.0% using both vocal tract and pitch information.

1 INTRODUCTION

The VQ-based method using speaker-specific codebooks is one of the well-known text-independent speaker recognition methods. This method is robust against utterance variations, such as session-to-session variation and text-dependent variation, if sufficient training and test data are available (Soong [2]). When the amount of available data is small, however, the performance is greatly decreased.

One of the authors, Furui [1], has shown that there is a strong interaction between speaker and phoneme factors of speech spectral vectors, and that broad-phonetic-class-dependent features are effective in speaker recognition. This fact suggests speaker recognition using broad-phonetic-class-dependent VQ codebooks. This approach is effective in avoiding mismatching between an input frame and codebooks corresponding to different kinds of phonemes, which is an especially serious problem with a small data set. Recently, Savic [3] and Eatock [4] have reported that speaker recognition methods which classify speech frames into broad phonetic classes achieve high performance.

This paper proposes a new robust VQ-based text-independent speaker recognition method which incorporates the broad phonetic classification approach and new distance measures. This method gives stable performance even with limited data [5].

2 SPEAKER RECOGNITION SYSTEM

The speaker recognition system uses broad-phonetic-class-dependent VQ codebooks created for each reference speaker. Figure 1 illustrates the training procedure.

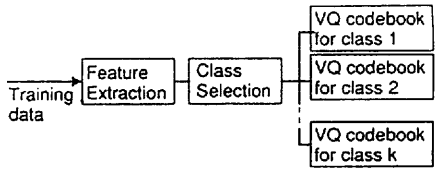

Figure 1. Training procedure for each reference speaker.

The speaker identification procedure is shown in Figure 2. Input speech frames classified into broad phonetic classes are vector-quantized using the codebooks of reference speakers, and distance (VQ-distortion) values are accumulated over all frames of each class. Then the distances for all classes are summed and used for the identification decision.

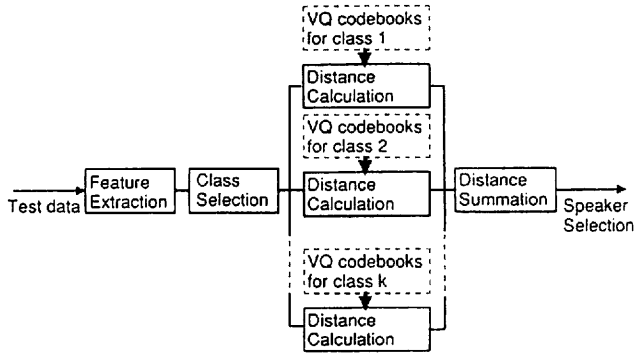

Figure 2. Speaker identification procedure.

2.1 Speech Classification

Two methods of broad phonetic classification were examined. One method classifies speech data into voiced and unvoiced parts, and makes a voiced and an unvoiced codebook for each speaker [5].

The second method uses a discrete ergodic HMM for classification. This method automatically creates phonetic classes by using training algorithms, and classifies speech frames into these classes. Figure 3 shows an example of an ergodic HMM.

CH2977-7/91/0000-0377 $1.00 © 1991 IEEE

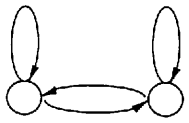

Figure 3. 2-state ergodic HMM.

Classification into k classes is performed using k-state ergodic HMMs, designed for each speaker using speaker-specific codebooks. The codebook size is set at 32. Initial probabilities of each state and each codebook element are set in proportion to the number of training samples assigned by VQ to each codebook element of each state. The HMM parameters are estimated using the Baum-Welch algorithm. The broad phonetic class for each input speech frame is determined by the alignment with the ergodic HMM using the Viterbi algorithm.

These two speech classification methods are compared from the viewpoint of speaker identification performance.

2.2 Distortion-Intersection Measure

In the conventional methods of calculating the overall distortion, that is, the distance between a set of test vectors and a set of VQ codebook vectors, quantization distortion is averaged over all test vectors. Some test vectors that are far from VQ codebook vectors, however, may correspond to phonemes not in the training data, or to feature parameters that vary from session to session. It is therefore possible for these vectors to impair speaker recognition.

We propose a new distance measure, which we call the Distortion-Intersection Measure (DIM) [5]. This distance between a set of test vectors and a set of VQ codebook vectors is defined in terms of the size of the intersection space between the two sets and the average quantization distortion for the intersection space.

DIM defines the distance $\mathcal{D}(\{\vec{y}_j\},\{\vec{c_{n}}_k\})$ between a set of test vectors $\{\vec{y}_j\}$ and a set of VQ codebook vectors for speaker n $\{\vec{c_{n}}_k\}$ as

$$\mathcal{D}(\{\vec{y}_j\},\{\vec{c_{n}}_k\}) = \frac{\sum_{j=1}^{J} d_{nj} + R_n (U - u_n)}{U}$$

$$d_{nj} = \begin{cases} min \parallel \vec{y}_j - \vec{c_{n}}_k \parallel^2 \\ \qquad \text{if } min \parallel \vec{y}_j - \vec{c_{n}}_k \parallel^2 \leq r_{nk} & (1) \\ 0 \qquad \text{otherwise} & (2) \end{cases}$$

$$R_n = max\, r_{nk} , \qquad U = max\, u_n ,$$

where $\| \cdot \|$ is the Euclidean distance and J is the total number of test vectors. The radius r_{nk} of a hypersphere approximating a cluster whose centroid vector is the VQ codebook vector $\vec{c_{n}}_k$ indicates the scope of that vector. This radius is set to the maximum Euclidean distance between the VQ codebook vector $\vec{c_{n}}_k$ and a training vector in the cluster. The term u_n is the number of test vectors corresponding to case (1).

The left-hand term $\sum_{j=1}^{J} d_{nj}$ in the numerator of the DIM equation represents the quantization distortion for the intersection space between a set of test vectors and a set of VQ codebook vectors for speaker n. This intersection space is determined as follows using the scope of the codebook vectors. Let $\vec{c_{n}}_k$ be the nearest VQ codebook vector to a test vector $\vec{y}_j$. If, as in Figure 4 (1), the test vector $\vec{y}_j$ is included in the scope of the VQ codebook vector $\vec{c_{n}}_k$, the quantization distortion d_{nj} is calculated according to (1). Otherwise (as in Figure 4 (2)) the quantization distortion d_{nj} is set to 0 according to (2).

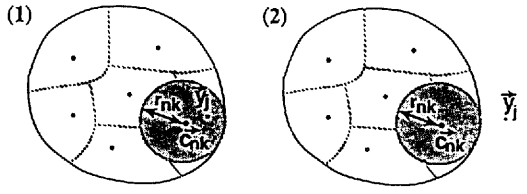

Figure 4. Illustration of two conditions, (1) and (2), for calculating the quantization distortion.

The right-hand term $R_n(U - u_n)$ corresponds to the penalty for the size of the space in a set of test vectors outside the intersection space. This term is proportional to the difference between the size of the intersection space for speaker n and the maximum size for all speakers. The larger the intersection space, the smaller the distance $\mathcal{D}(\{\vec{y}_j\},\{\vec{c_{n}}_k\})$. The coefficient R_n balances the importance of the quantization distortion and the size of the intersection space.

2.3 Talker Variability Normalization

In speaker recognition, feature parameters which have relatively large inter-speaker variability and small intra-speaker variability are most effective. When treating different feature parameters simultaneously, it is necessary to normalize the distribution of each feature parameter according to its effectiveness for speaker recognition. The Mahalanobis method normalizes the feature parameter distribution using the intra-speaker standard deviation of that parameter.

We propose the Talker Variability Normalization (TVN) method [5] which enhances inter-speaker variability and reduces intra-speaker variability. In TVN, the normalized i-th feature vector element y_i is given by

$$y_i = w_i \cdot x_i ,$$

where x_i is the i-th feature vector element and w_i is its normalization weight. The normalization weight is defined as

$$w_i = \sum_{n=1}^{N} \sum_{m=1, m \neq n}^{N} \frac{\sigma_{ni}}{L_{mni}^2} ,$$

where N is the number of speakers, σ_{ni} is the standard deviation of the i-th feature vector element for speaker n, and L_{mni} is a measure for the size of the intersection between the distributions of the i-th feature vector elements for speakers m and n, when the distributions are approximated by normal distributions.

L_{mni} is given by

$$L_{mni} = \begin{cases} 3(\sigma_{mi} + \sigma_{ni}) - |\mu_{mi} - \mu_{ni}| & \text{if (3)} \\ 6 \cdot min(\sigma_{mi}, \sigma_{ni}) & \text{if (4)} \\ \varepsilon_i & \text{if (5)} , \end{cases}$$

(3) $3|\sigma_{mi} + \sigma_{ni}| - \varepsilon_i \geq |\mu_{mi} - \mu_{ni}| > 3|\sigma_{mi} - \sigma_{ni}|$
(4) $3|\sigma_{mi} - \sigma_{ni}| \geq |\mu_{mi} - \mu_{ni}|$
(5) $|\mu_{mi} - \mu_{ni}| > 3|\sigma_{mi} + \sigma_{ni}| - \varepsilon_i$,

where μ_{ni} is the mean of the i-th feature vector element for speaker n, and ε_i is a positive constant.

Figure 5 shows that (3) corresponds to cases in which the distributions of the i-th feature vector element for speakers

m and n overlap, (4) fits cases in which the distribution for speaker n is included in the distribution for speaker m, and (5) applies when the distributions for speakers m and n are separate. The value ε_i is chosen in accordance with the distribution of the i-th feature vector element.

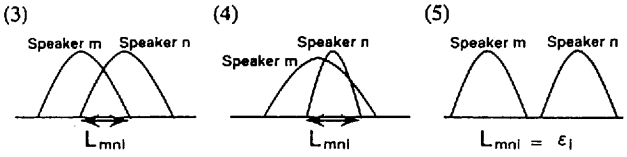

Figure 5. Illustration of L_{mni} under three different conditions, (3), (4), and (5).

The smaller the probability of the i-th feature vector element for speaker n occurring within the intersection with speaker m, the more effective the feature parameter. When the distribution is simply approximated by a triangle, the probability is approximately proportional to $\frac{L^2_{mni}}{\sigma^2_{ni}}$.

The weight w_i is created from the product of the reciprocal of this factor and the factor $\frac{1}{\sigma_{ni}}$ which is usually used in the Mahalanobis method,

$$w_i = \sum_{n=1}^{N} \sum_{m=1, m\neq n}^{N} \frac{1}{\frac{L^2_{mni}}{\sigma^2_{ni}}} \frac{1}{\sigma_{ni}}.$$

In this way, the weight w_i normalizes the intra-speaker variability and enhances the inter-speaker variability. This method is expected to be more effective than the conventional Mahalanobis normalization method.

2.4 Pitch Feature Combination

We expect that speakers can be characterized more exactly, and that speaker recognition methods can be made more robust, by combining different types of feature parameters. This paper investigates the usefulness of combining pitch information, pitch and delta-pitch frequencies, with vocal tract information, cepstral and delta-cepstral coefficients. Input speech frames are classified into voiced and unvoiced frames, and vector-quantized using the codebook of the assigned class. A voiced code vector consists of cepstral and delta-cepstral coefficients and pitch and delta-pitch frequencies, whereas an unvoiced code vector consists only of cepstral and delta-cepstral coefficients.

3 RECOGNITION EXPERIMENTS

The database consists of utterances by nine male talkers recorded on four occasions over three years. It includes five vowels, five words, and three sentences. Durations of the sentences are about 5, 12, and 30 sec. Cepstral coefficients are calculated by conventional LPC analysis. The analysis order is 16, the frame period is 8 ms, and the frame length is 32 ms. Delta-cepstral coefficients are calculated as the first-order regression coefficients over an 88-ms period. Pitch frequency is calculated using LPC residual waves. Delta-pitch frequency is calculated as the first-order regression coefficient of the pitch frequency sequence over a 152-ms period. The TVN method is used for normalization.

The vowels, words, and two of the sentences (5 and 12 sec) are used for training. The longest sentence (30 sec), different from training sentences, is used as a whole or partially, for testing to evaluate the robustness of text-independent speaker recognition. Codebooks are made on every occasion for each speaker, and test data from one occasion are tested using the codebooks for the other three occasions.

3.1 Speech Classification Effects

The results of speaker identification experiments under the six conditions indicated in Table 1 are shown in Figure 6. Pitch information is not used in these experiments. The length of test sentences, varied in the experiments, is normalized by the total length (30 sec) and indicated on the horizontal axis.

Table 1. Experimental conditions.

Condition	Speech Classification	Distance Measure
HMM+VQ(DIM)	HMM	VQ(DIM)
V/UV+VQ(DIM)	V/UV	
VQ(DIM)	no	
HMM+VQ	HMM	VQ
V/UV+VQ	V/UV	
VQ	no	

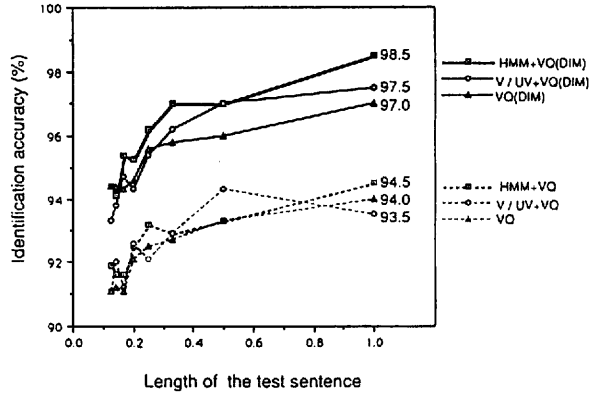

Figure 6. Speaker identification accuracy as a function of test sentence length.

The figure demonstrates that speech classification by an (2-state) ergodic HMM is effective in improving the identification accuracy. It is clearly shown that DIM is more effective than the conventional VQ-distortion measure.

3.2 Pitch Feature Combination Effects

The results for speaker identification using pitch and delta-pitch frequencies in addition to cepstral and delta-cepstral coefficients are shown in Figure 7.

Identification error rates can be reduced to roughly 1/3 by also using pitch and delta-pitch frequencies. These results indicate the usefulness of pitch information in text-independent speaker recognition.

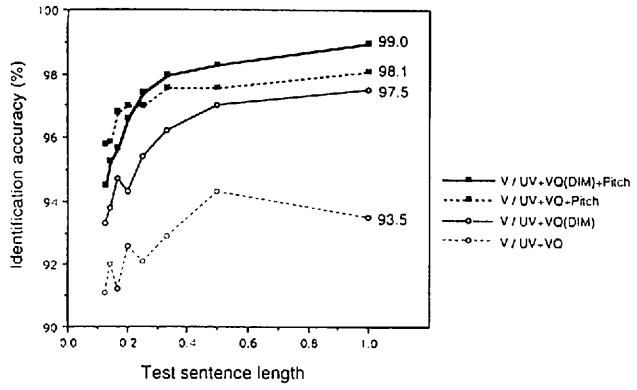

Figure 7. Results of speaker identification using pitch information.

4 DISCUSSION

Additional speaker identification experiments using 2, 3, and 4-state ergodic HMMs were carried out. Figure 8 shows error rates for two different training sets, A and B. Training set A corresponds to that used in the recognition experiments thus far. Training set B consists of training set A and the first half of the 30-sec sentence. The length of the utterances in set B is about 1.75 times that in set A. The second half of the 30-sec sentence is used for testing. The results show that, as the number of HMM states increases, more training data is needed. If there is enough training data, the recognition performance seems to be increased by using an HMM with more than 2 states.

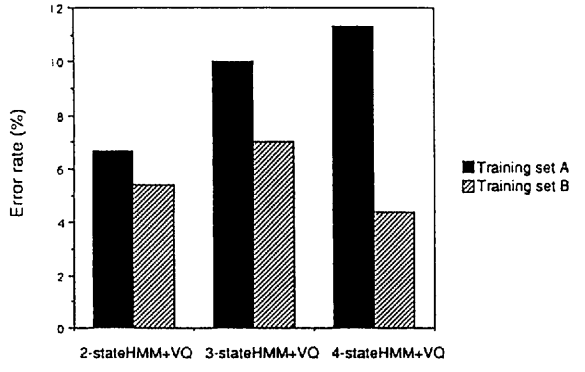

Figure 8. Speaker identification error rate as a function of the number of HMM states.

An additional experiment was done directly using speaker-specific ergodic HMMs for speaker recognition, instead of using them for broad phonetic classification. In this experiment, likelihood values obtained for each frame using the speaker-specific ergodic HMM are accumulated over all test frames and used for the recognition decision. Ergodic HMMs with 2, 3, and 4 states were tested. The HMM, with a common codebook used for all speakers, is trained for each reference speaker. The universal codebook size is set at 256. The results shown in Figure 9 indicate that the ergodic HMM itself cannot achieve high recognition performance but that high performance is achieved when it is combined with VQ-based recognition.

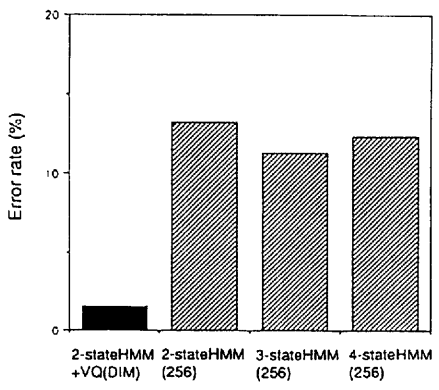

Figure 9. Speaker identification error rate using only HMM or the combination of HMM and VQ.

5 SUMMARY

This paper has demonstrated the effectiveness of broad-phonetic-class-dependent VQ codebooks in text-independent speaker recognition. Identification error rates using a 2-state ergodic HMM as the broad phonetic classifier are roughly half those without speech classification. DIM has been proposed for measuring the distance between a VQ codebook and a set of test vectors. For text-independent speech data with wide variability, DIM is more efficient than the conventional VQ-distortion measure. The error rates using DIM are roughly half those using the conventional measure. TVN has been introduced for feature-parameter normalization. This paper has also demonstrated that pitch features can be effectively combined with vocal tract features. Identification error rates using cepstral and delta-cepstral coefficients and pitch and delta-pitch frequencies are roughly a third of those using only cepstral and delta-cepstral coefficients. When these new methods are combined, speaker identification rate using only vocal tract information is as high as 98.5%, and that using vocal tract and pitch information is 99.0%.

Further study includes the evaluation of these methods using a large database.

6 ACKNOWLEDGMENT

The authors wish to acknowledge the members of the Speech and Acoustics Laboratory of NTT Human Interface Laboratories for their valuable and stimulating discussions.

REFERENCES

[1] S.Furui, "Research on individuality features in speech waves and automatic speaker recognition techniques," *Speech Communication 5,pp.183-197 (1986)*

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

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

[4] J.Eatock and J.S.Mason, "Automatically focusing on good discriminating speech segments in speaker recognition," *Proc. ICSLP,pp.133-136 (1990)*

[5] T.Matsui and S.Furui, "Text-independent speaker recognition using vocal tract and pitch information," *Proc. ICSLP,pp.137-140 (1990)*