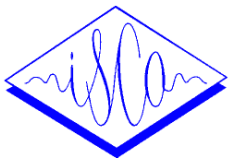

First International Conference on
Spoken Language Processing (ICSLP 90)
Kobe, Japan
November 18-22, 1990

Text-Independent Speaker Recognition Using Vocal Tract and Pitch Information

Tomoko Matsui and Sadaoki Furui

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

ABSTRACT

This paper proposes a new text-independent speaker recognition method based on vector quantization (VQ) using vocal tract and pitch information. The purpose of this research is to create a speaker recognition system robust against the temporal variations of feature parameters. This paper introduces several feature parameters related to both vocal tract and pitch information extracted from spoken vowels, words, and sentences. Interspeaker variability is enhanced, and intraspeaker variability is reduced, by using a new normalization method, Talker Variability Normalization (TVN). A new distance measure, the Distortion-Intersection Measure (DIM), is defined by the size and similarity of the intersection between test vectors and VQ codebook vectors. This proposed method, evaluated using a nine-talker database recorded over three years, achieves 99.0% speaker identification and 98.7% speaker verification accuracy.

1 Introduction

Temporal variations of feature parameters are among the most difficult problems in speaker recognition, and they significantly affect recognition accuracy. Furui [2] has investigated a method combining statistical and dynamic features to achieve highly accurate speaker recognition by using a database recorded over several years. Soong et al. [3], [4] have investigated speaker recognition by a vector quantization (VQ) codebook using a spoken-digit database recorded over several months.

We assume that speakers can be characterized more exactly, and that speaker recognition methods can be made more robust, by using a combination of different types of feature parameters. In conventional speaker recognition systems [1] - [4], a method for combining vocal tract and source features effectively has not yet been established.

This paper describes a VQ-based text-independent speaker recognition method using vocal tract and pitch information to obtain high-performance speaker recognition that is robust against temporal variations of feature parameters.

2 Combination of feature parameters

This paper uses cepstral and delta-cepstral coefficients as vocal tract information, and pitch and delta-pitch frequency as vocal source information. Cepstral coefficients are extracted by conventional LPC analysis. The coefficient order is 16, the analysis interval is 8 ms, and the data window length is 32 ms. Delta-cepstral coefficients are obtained as first-order regression coefficients over 88-ms time-slots. Pitch frequency is also extracted by conventional LPC analysis. Delta-pitch frequency is the first-order regression coefficient of the pitch frequency sequence over a 152-ms time-slot.

We use these feature parameters for VQ-based speaker recognition. There are two codebooks for each speaker: a voiced codebook and an unvoiced codebook. A voiced code vector consists of cepstral

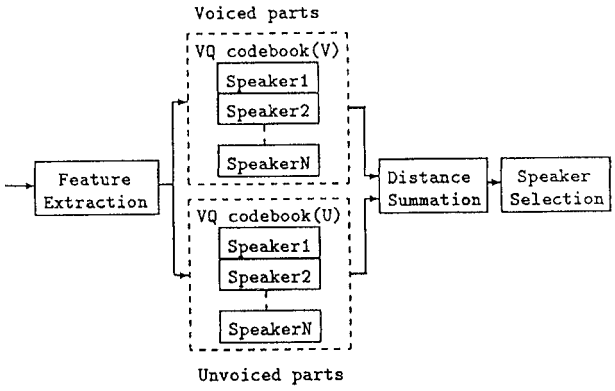

Figure 1. System block diagram.

and delta-cepstral coefficients, and pitch and delta-pitch frequencies, whereas an unvoiced code vector consists only of cepstral and delta-cepstral coefficients. Figure 1 shows a block diagram of the speaker recognition system.

3 Normalization methods

In speaker recognition, the most effective feature parameters have relatively large interspeaker variability and small intraspeaker variability. When treating some different kinds of feature parameters simultaneously, it is necessary to normalize the distribution of each feature parameter according to its effectiveness for speaker recognition. Mahalanobis method normalizes a feature parameter distribution using the intraspeaker standard deviation of that parameter.

This paper introduces Talker Variability Normalization (TVN) to enhance interspeaker variability and to reduce intraspeaker 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 w_i 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 to normal distributions. L_{mni} is given by

$$L_{mni} = \begin{cases} 3(\sigma_{mi} + \sigma_{ni}) - |\mu_{mi} - \mu_{ni}| & \text{if (1)} \\ 6 \cdot min(\sigma_{mi}, \sigma_{ni}) & \text{if (2)} \\ \varepsilon_i & \text{if (3)} \ , \end{cases}$$

(1) $3|\sigma_{mi} + \sigma_{ni}| - \varepsilon_i \geq |\mu_{mi} - \mu_{ni}| > 3|\sigma_{mi} - \sigma_{ni}|$
(2) $3|\sigma_{mi} - \sigma_{ni}| \geq |\mu_{mi} - \mu_{ni}|$
(3) $|\mu_{mi} - \mu_{ni}| > 3|\sigma_{mi} + \sigma_{ni}| - \varepsilon_i \ ,$

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

Figure 2 shows that (1) corresponds to cases in which the distributions of the i-th feature vector element for speakers m and n overlap, (2) fits cases in which the distribution for speaker n is included in the distribution for speaker m, and (3) applies to cases with separate distributions for speakers m and n. The value ε_i is chosen in accordance with the distribution of the i-th feature vector element. The smaller the probability of having the values within the intersection range of L_{mni} is, the more effective the feature parameter is. When the distribution is simply approximated by a triangle, the probability of occurring the i-th feature vector element for speaker n within the intersection with speaker m is approximately proportional to $\frac{L_{mni}^2}{\sigma_{ni}^2}$.

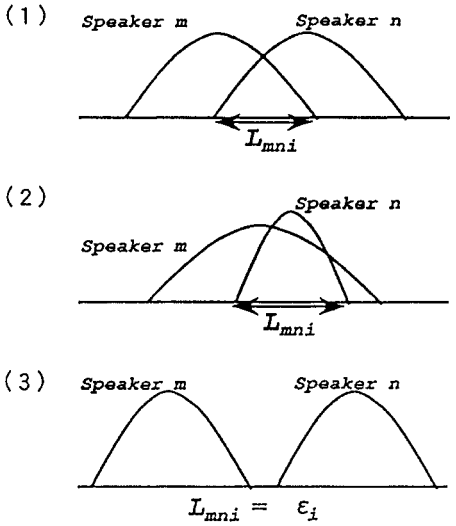

Figure 2. Illustration of L_{mni} for the three conditions, (1), (2), and (3).

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

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

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

4 Distance measures

In the conventional methods, the distance between a set of test vectors and a set of VQ codebook vectors is defined as the average quantization distortion over all test vectors. Some test vectors that are far from VQ codebook vectors, however, may be outliers corresponding to phonemes that are not included in the training data, or corresponding to feature parameters that vary from session to session. It is therefore

possible that these vectors impair text-independent speaker recognition that uses feature parameters with intrinsically wide variability.

We propose a new distance measure called the Distortion-Intersection Measure (DIM). 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 $\mathcal{D}(\{\vec{y}_j\},\{\vec{c}_{\mathbf{n}k}\})$ between a set of test vectors $\{\vec{y}_j\}$ and a set of VQ codebook vectors for speaker n $\{\vec{c}_{\mathbf{n}k}\}$ as

$$\mathcal{D}(\{\vec{y}_j\},\{\vec{c}_{\mathbf{n}k}\}) = \frac{\sum_{j=1}^{J} d_{\mathbf{n}j} + R_{\mathbf{n}}\,(U - u_{\mathbf{n}})}{U}$$

$$d_{\mathbf{n}j} = \begin{cases} \min\limits_{k} \| \vec{y}_j - \vec{c}_{\mathbf{n}k} \|^2 \\ \qquad \text{if } \min\limits_{k} \| \vec{y}_j - \vec{c}_{\mathbf{n}k} \|^2 \leq r_{\mathbf{n}k} & (4) \\ 0 \qquad \text{otherwise} & (5) \end{cases}$$

$$R_{\mathbf{n}} = \max_{k} r_{\mathbf{n}k}\ , \qquad U = \max_{n} u_{\mathbf{n}}\ ,$$

where $\| \cdot \|$ is the Euclidean distance and J is the total number of test vectors. The radius $r_{\mathbf{n}k}$ of a hypersphere approximating a cluster whose centroid vector is the VQ codebook vector $\vec{c}_{\mathbf{n}k}$ indicates the scope of that vector. This radius is set to the maximum Euclidean distance between the VQ codebook vector $\vec{c}_{\mathbf{n}k}$ and a training vector in the cluster. The term $u_{\mathbf{n}}$ is the number of test vectors corresponding to case (4).

The left-hand term $\sum_{j=1}^{J} d_{\mathbf{n}j}$ 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 defined by the scope of a VQ codebook vector. For this, we assume that a test vector $\vec{y}_j$ is the nearest test vector to VQ codebook vector $\vec{c}_{\mathbf{n}k}$. If, as in Figure 3a, a test vector $\vec{y}_j$ is included in the scope of a VQ codebook vector $\vec{c}_{\mathbf{n}k}$, the quantization distortion $d_{\mathbf{n}j}$ is calculated according to (4). Otherwise (as in Figure 3b) the quantization distortion $d_{\mathbf{n}j}$ is set to 0 according to (5).

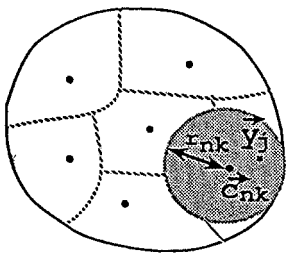

Figure 3a. Condition (4)

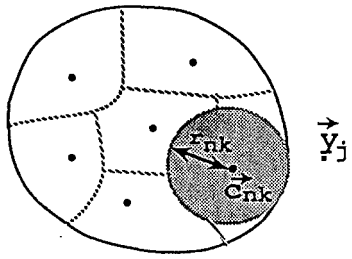

Figure 3b. Condition (5)

The right-hand term $R_{\mathbf{n}}\,(U - u_{\mathbf{n}})$ 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 of any speaker. The larger the intersection space is, the smaller the distance $\mathcal{D}(\{\vec{y}_j\},\{\vec{c}_{\mathbf{n}k}\})$ is. The coefficient $R_{\mathbf{n}}$ determines which is more important, the quantization distortion or the size of the intersection space.

5 Experiments

The speech database used for the experiment contains data from nine male talkers recorded on four occasions over three years. It contains five vowels, five words, and three sentences. The durations of the three sentences are about 5, 12, and 30 s. The vowels, words, and two of the sentences (5 and 12 s) were used for training. The longest sentence (30 s) was used for testing. Since the texts of the test and training data in this experiment are not the same, this speaker-recognition experiment is text-independent. VQ codebooks were made using the LBG algorithm. A codebook was made on every occasion for each speaker. The test data from one occasion was tested using each codebook for the other three occasions as the reference. The proposed method was evaluated by averaging error rates for speaker identification (I) and verification (V).

5.1 Feature parameter combination effects

Table 1 lists the results for speaker identification using several combinations of feature parameters. We used Mahalanobis method for normalization. The average quantization distortion over all test vectors was used as a distance measure. The error rates are smaller when more parameters are used, and vocal source features combine effectively with vocal tract features.

Table 1. Evaluation of feature parameter combinations

Feature parameters	Error rate (%)
Cepstrum	9.43
ΔCepstrum	11.81
Pitch	74.42
ΔPitch	85.88
Cepstrum, ΔCepstrum	7.93
Cepstrum, ΔCepstrum Pitch, ΔPitch	2.89

5.2 Effects of TVN

Table 2 gives error rates for speaker recognition by Mahalanobis method and by TVN. The average quantization distortion over all test vectors was used as a distance measure. For both speaker identification and verification, TVN is clearly more effective than Mahalanobis method.

Table 2. TVN vs. Mahalanobis error rate (%)
I: identification, V: verification

Feature parameters		TVN	Mahalanobis
Cepstrum	I	5.50	9.43
	V	5.12	5.69
Cepstrum, ΔCepstrum	I	6.02	7.93
	V	4.83	6.77
Cepstrum, ΔCepstrum	I	1.91	2.89
Pitch, ΔPitch	V	5.58	9.93

5.3 Effects of DIM

Table 3 allows comparison of speaker recognition by DIM with recognition using the average quantization distortion over all test vectors (ALL) when TVN was used for normalization. DIM is more effective than the conventional distortion measure of VQ.

Table 3. DIM vs. ALL error rate (%)

Feature parameters		DIM	ALL
Cepstrum	I	5.15	5.50
	V	2.20	5.12
Cepstrum, ΔCepstrum	I	4.05	6.02
	V	2.46	4.83
Cepstrum, ΔCepstrum	I	1.00	1.91
Pitch, ΔPitch	V	1.27	5.58

6 Summary and Conclusions

This paper demonstrates that the combination of vocal tract and pitch information is effective in speaker recognition. The vocal tract information (cepstral and delta-cepstral coefficients) and the vocal source information (pitch and delta-pitch frequency) were combined into a vector, and voiced and unvoiced codebooks were used. The error rate decreases as the number of feature parameters increases. Identification error rates using cepstral and delta-cepstral coefficients and pitch and delta-pitch frequency are roughly 1/3 of those using only the cepstral coefficients. This paper also introduces TVN as a feature-parameter normalization method. Experimental results show that TVN is more effective for speaker recognition than is Mahalanobis method. The error rates using TVN are roughly 2/3 of those using Mahalanobis method. This paper proposes DIM for measuring the distance between the VQ codebook and test vectors. For text-independent speech data with wide variability, we found that DIM is more efficient than the conventional distortion measure of VQ. The error rates using DIM are roughly 1/2 of those using the conventional distortion measure of VQ. The speaker identification rate was as high as 99.0%, and the speaker verification rate was 98.7%.

We are presently analyzing DIM and applying it in an HMM-based speaker recognition system. To evaluate the proposed method more precisely, we are also increasing the number of talkers in our database.

7 Acknowledgment

The authors wish to acknowledge the members of NTT Basic Research Laboratories and Human Interface Laboratories for their valuable and stimulating discussions.

REFERENCES

[1] S.Furui "Research on individuality information in speech waves," *Ph.D Thesis, Tokyo University (1978)*

[2] S.Furui, "Comparison of speaker recognition methods using statistical features and dynamic features," *IEEE Trans. ASSP,pp.342-350 (1981)*

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

[4] F.K.Soong et al., "On the use of instantaneous and transitional spectral information in speaker recognition," *Proc. ICASSP,pp.877-880 (1986)*