ISCA Archive
http://www.isca-speech.org/archive

International Symposium on Chinese Spoken Language Processing (ISCSLP 2004)
Hong Kong
December 15-18, 2004

TEXT-INDEPENDENT SPEAKER VERIFICATION BASED ON RELATION OF MFCC COMPONENTS

Guiwen Ou, Dengfeng Ke
lnsogw@zsu.edu.cn
is99kdf@tom.com

Department of Computer Science
Zhongshan University, Guangzhou, 510275

ABSTPRACT

GMM is prevalent for speaker verification. It performs very well but needs a background model to give a reference value, which would greatly influence the error rate. In order to get better result of generalization, a large database with lots of people is needed to train the background model. In this paper, a new method without background model is proposed, which will be called Correlation and Kernel function method (CK method) later. In CK method, the correlation and un-correlation of MFCC are used to identify individuals, and a kernel function is used to work out the likelihood of two models. It works more than 30 times as fast as GMM method does, but requires fewer data to train and fewer space to store the model. But its performance is nearly identical to that of GMM. So it is suitable for real-time computation.

1. INTRODUCTION

Speaker verification and identification have several applications in security guard and e-Business. Speaker identification means to recognize a person out of a given group of people, while speaker verification means to verify whether the speech is spoken through the given person or not. The later one can also be separated into text-dependent one and text-independent one. In this paper, a method for text-independent speaker verification is proposed.

GMM [1,2] has been being the most classical method for text-independent speaker verification. Reynolds etc. introduced GMM to speaker verification in [1,2], which needs an extra model called background GMM. The background GMM is trained from a large database of different people. The speeches of this database should be carefully selected from different people in order to get better results.

In practice, the selection of speeches for training the background model becomes a hard problem. So Hsu etc. [3] proposed a method to keep from such a pet hate. But their method is too demanding to ask for about 3.5 minutes' speech each to train 20 GMMs for every speaker and then get the thresholds for every one.

Zilca [9] said that GMM mainly uses two statistics — mean and covariance. Through some particular processing, mean would be zero, and then GMM can be simplified to covariance only. So F. Bimbot [4] and R.D.Zilca [5,9] proposed a method for speaker verification called Covariance Modeling, which ignores mean statistics to simplify GMM. This method needs a background model as well. Speaker verification system with a background model has several inconveniences. In order to add a given person to the database, not only the model of the given person should be added but also the background model should be retrained accordingly.

Text-independent speaker verification based on GMM supposed that LPCC or MFCC satisfy Gaussian distribution and that LPCC or MFCC of each frame contain the characteristic of the speaker, So GMM works out probability densities for every frame and then sums up the likelihood to give determination. Our point of view is that LPCC or MFCC of a single frame contains little characteristic of a person. It is the relation between frequency bands that contains the chief characteristic of a person. So in this paper, correlation matrix is used to represent the correlation and kurtosis vector is used to represent the un-correlation of frequency bands.

In fact, Gaussian distribution is only an approximate description of speech feature. There are several assumptions for speech feature. For example, Davenport [6] assumed that the speech feature is of Laplace distribution, which is given as follows:

$$f_x(x) = \frac{1}{\sqrt{2}\delta_x} * e^{\frac{-\sqrt{2}|x|}{\delta_x}} \qquad (1)$$

Paez and Glisson [7] said that gamma distribution is better, which is shown bellow:

$$f_x(x) = \sqrt{\frac{\sqrt{3}}{8\pi\delta_x \mid x \mid}} * e^{\frac{-\sqrt{3}|x|}{2\delta_{\max}}} \qquad (2)$$

In this paper, we need not assume any distribution to the speech feature.

Because of the differences of our vocal organs, every one has a distinguishable accent of speech. In this paper, we consider that this characteristic of different person can be mainly described through the relation between frequency bands. From this point of view, the correlation of MFCC feature is used to represent the correlation of different frequency bands, and the kurtosis vector of MFCC feature is used to represent the un-correlation of different frequency bands. Also, a kernel function is used to work out the likelihood between two models. Background model is not needed in this method.

2. SPEAKER DATABASE AND PREPROCESSING

In order to confirm this point of view, considering about the impact of different time length, we build 2 databases out of 25 people. The 1st database contains 25 people, about 10 speeches per person, and about 15 seconds per speech. The 2nd database contains 25 people, about 30 speeches per person, and about 5 seconds per speech. Both databases are sampled at 16k Hz 8 bits. In this paper, pre-emphasis coefficient is 0.96, frame size is 16ms, shift size is 8ms, and MFCC feature is calculated through frequency band from 20Hz to 8KHz, filter number 29, and order number 12 with the 1st order deserted.

3. CORRELATION MATRIX

It has some advantages to use correlation matrix to verify a person. It can keep from the negative influence of some environment conditions that should be avoided, such as the convolution of speech. To discuss this, a little knowledge of statistic is briefly introduced as follows:

The covariance of two scalar quantities is defined as

$$\mathrm{cov}(x,y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1} \qquad (3)$$

So, easily we can see that

$$\mathrm{cov}(x + x_0, y + y_0) = \mathrm{cov}(x,y) \qquad (4)$$

$$\mathrm{cov}(\alpha x, \beta y) = \alpha\beta\,\mathrm{cov}(x,y) \qquad (5)$$

where α and β are some non-zero constants.

The variance is defined from the covariance as

$$\delta^2(x) = \mathrm{cov}(x,x) = \frac{\sum (x_i - \bar{x})(x_i - \bar{x})}{n-1} \qquad (6)$$

and the correlation coefficient of two scalar quantities is defined as

$$r(x,y) = \frac{\mathrm{cov}(x,y)}{\delta(x)\delta(y)} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2}\sqrt{\sum (y_i - \bar{y})^2}} \qquad (7)$$

So we can see that

$$r(x,y) = r(x + x_0, y + y_0) \qquad (8)$$

$$r(x,y) = r(ax, by) \qquad (9)$$

This is a very importance characteristic of correlation. It ignores the mean and the scale of MFCC， and considers only the shape of MFCC. The convolution of speech influences the scale of MFCC, which would be ignored by this characteristic of correlation. In the case of n-ary variable X, this characteristic of (8) and (9) holds as well. Because the covariance matrix $\mathrm{cov}(X) = (c_{i,j}), c_{i,j} = \mathrm{cov}(x_i, x_j)$ can be drawn out from (3) and correlation matrix $r(X) = (r_{i,j}), r_{i,j} = r(x_i, x_j)$ can be drawn out from (7).

4. KURTOSIS

We consider the correlation of different frequency bands in the previous part. In the meanwhile, the un-correlation of different frequency bands is also an important relation between frequency bands. On inspiration of independent component analysis [8], we use 4-order statistic, kurtosis function, to quantify how independent the frequency bands are, which can to some extent quantify the un-correlation of frequency bands.

Kurtosis function is defined as

$$kurt(x) = E(x^4) - 3(E(x^2))^2 \qquad (10)$$

kurt(x) is zero when x is of Gaussian distribution, positive when of super-Gassian, negative when of sub-Gaussian. “Non-Gaussian is independent” [11], independent is a kind of un-correlation. So kurtosis value can be used to partly represent the un-correlation of frequency bands.

5. KERNEL FUNCTION

Generally, a kernel function used in Hilbert space is

$$K(x,y) = \phi(x)^T \phi(y) = \sum_{i=1}^{d_\phi} \phi_i(x)\phi_i(y) \qquad (11)$$

where ϕ is a non-linear function, d_ϕ is the dimension of $\phi(x)$. In this paper, a kernel used in SVM is chosen, which is defined as

$$K(x,y) = \exp(-\frac{\| x - y \|^2}{2}) \qquad (12)$$

Here ||x-y|| means the 1-norm of x-y. 1-norm is defined as

$$\| X \| = \max \sum | x_i | \qquad (13)$$

In (12)，x and y can be either vector or matrix. The time complexity is O(n) and $O(n^2)$ respectively for vector and matrix. In some papers [3,4], covariance matrix is treated as the characteristic of different person. The distance of two persons is defined as the distance of two covariance matrices which is

$$D(X,Y)=\log\frac{trace(\mathrm{cov}^{-1}(X)\mathrm{cov}(Y))trace(\mathrm{cov}(X)\mathrm{cov}^{-1}(Y))}{n^2} \quad (14)$$

Formula (14) need to compute the inverts of 2 matrices, whose computation complexity is $O(n^3)$， so the process is time consuming. In GMM model, the invert of covariance matrix, the determinant of covariance matrix should be computed. They are $O(n^3)$as well. In some cases, the covariance matrix in GMM may be nearly singular so that the computation process had to be stopped. The kernel function method would never meet such awkward.

Another advantage for (12) is that the value of K(x,y) falls in (0,1], which is simple for determination. Simply 1 means same person while 0 means different one.

6. VERICATION ALGORITHM

Let M0, K0 be the trained model, M0 the correlation matrix and K0 the kurtosis vector. Let M1, K1 be the correlation matrix and kurtosis vector of an unknown speaker. Use this algorithm to verify:

Step 1: Likelihood:= K(M0,M1)

Step 2: if $\text{Likelihood} > \theta_{\min}$ and $Likelihood < \theta_{\max}$ then

$NewLikelihood := K(K0, K1)$

$NewLikelihood := (1-p)Likelihood + p * NewLilihood$

End if

Setup 3: if $\text{Likelihood} > \theta$ then

Accept

else

Reject

End if

In the above algorithm, the correlation matrix plays the predominant part, and the kurtosis vector assists the determination. It simulates the recognition of human intuition - if the likelihood is larger than a threshold $\theta_{\max}$, it should be accepted, if the likelihood is smaller than a threshold $\theta_{\min}$, it should be rejected, if the likelihood between $\theta_{\min}$ and $\theta_{\max}$, it may be yes may be no, so a kurtosis vector is used to assist this determination. The forget factor $p \in (0,1)$ and the threshold θ should be set between $\theta_{\min}$ and $\theta_{\max}$. In this paper, we set $\theta_{\max}=0.6$, $\theta_{\min}=0.5$, $\theta=0.54$ and $p=0.8$, and get wonderful results.

7. TIME AND SPACE COMPLEXITY

Time complexity can be divided into 3 parts — feature extraction time, training time and verification time.

(1) Feature extraction: GMM needs to compute MFCC only, which is $O(n\log(n)\mathrm{T})$, where n is frame size and T is sequence length. Correlation and Kurtosis Method (CK) needs to compute correlation matrix and kurtosis vector, which needs an extra time $O(m^2T)$ to compute correlation and $O(mT)$ to compute kurtosis vector, where m is the order of MFCC. The computation time seems much longer than GMM. Noticed that m is often chosen as 12, n is often chosen as 256, m^2 is smaller than n*log(n), so the computation time for CK is also $O(n\log(n)\mathrm{T})$. It is only a little longer than GMM.

(2) Training：GMM uses VQ and EM algorithm to separate data, and then computes the mean vector and the covariance matrix for every component. Every step is time consuming, which needs $\Omega(n^3T)$. But CK method needs no training. The feature—correlation matrix and kurtosis vector, is also the model. So the training time is $O(1)$.

(3) Verification：GMM needs to calculate probability density for MFCC of each frame, which is $O(nT)$. CK method needs to work out likelihood of train matrix or vector, which respectively costs $O(m^2)$ and $O(m)$. So CK method needs much less time than GMM does.

The computation time advantage will be shown in Table 2 in the next part.

Space complexity here mainly refers to the storage space of speaker models. A GMM with k components needs to store k mean vectors, k covariance matrices and k weight values, which is $\Theta(k(m^2+m))$. CK method needs only $\Theta(m^2+m)$.

8. EXPERIMENT AND RESULTS

The 2 testing databases are as described in part 2. The 1st speech of each person is used to train the model. The other speeches of the same person are used to test the error rejection rate. And the speeches of the other people are used to test the error acceptance rate.

The register and imposter speech counts are as shown in Table 1.

Table 1 Database for Experiment

Speech Length	Register Speech Count	Imposter Speech Count
15 s	232	5568
5 s	808	19392

In order to show the wonderful results of this CK method, the results of a GMM method of 4 components is used to compared with. The following results are gained from a personal computer（Pentium IV 2G）.

Table 2 Time Complexities

Speech Length	Training Time		Recognition Time	
	GMM	CK	GMM	CK
15 s	8m14s	6s	55m57s	48s
5 s	5m17s	5s	54m36s	2m25s

Table 3 Error Rate

Speech Length	False Rejection Rate		False Accepttion Rate	
	GMM	CK	GMM	CK
15 s	6.03%	2.16%	5.78%	5.50%
5 s	21.7%	34.2%	1.75%	1.73%

These tables show that, the training and recognition time of CK is much less than that of GMM, but the results are nearly equivalent to that of GMM. When the training data is limited (only a speech of about 15 seconds), CK method performs even a little better than GMM does.

9. CONCLUSIONS

In this paper, we consider that the characteristic of a person is mainly described through the relation between frequency bands. According to this point of view, correlation and kurtosis of MFCC are used to denote this relation. A kernel function is used to work out the likelihood between two models. This CK method needs very much less time than GMM does, while the result is nearly equivalent to GMM. Also, it needs very few data to train and little space to store the model. So it is suitable for real-time computation. This method has been implemented to a real-time verification system with the co-operation of Crystal Ball Company, and the further results will be shown in the next paper.

10. REFERENCES

[1] D. A. Reynolds, “Speaker Identification and Verification using Gaussian Mixture Speaker Models,” *Speech Communication*, Vol. 17, pp. 91-108, 1995.

[2] Douglas A. Reynolds, Thomas F. Quatieri and Robert B. Dunn. “Speaker verification using adapted Gaussian mixture models.” *Digital Signal Processing*, Academic Press, 2000.

[3] *Chun-Nan Hsu, Hau-Chung Yu, Bo-Hou Yang,* SPEAKER VERIFICATION WITHOUT BACKGROUND SPEAKER MODELS, http://chunnan.iis.sinica.edu.tw/OSCILLO/oscillo10252002.pdf

[4] F. Bimbot, I. M. Magrin -Chagnolleau and L. Mathan, “Second Order Statistical Measures for Text Independent Speaker Identification,” *Speech Communication* , Vol. 17, pp. 177-192.

[5] R. D. Zilca, “Text Independent Speaker Verification Using Covariance Modeling,” *IEEE Signal Processing Letters,* April 2001.

[6]. W. B ,Davenport,. “An Experimental Study of Speech Wave Probability Distributions”, *Journal of Acoust. Soc. America*, Vol 24, pp: 390-399, 1952.

[7]. M. D. Paez,., and T. H Glisson,. “Minimum Mean Squared Quantization in Speech”, *IEEE Transactions on Communications*, Vol. Com-20, pp: 225-230, 1972.

[8] A.Hyvarinen, E.Oja, “Independent component analysis: algorithms and applications”, *Neural Networks 13(2000)*, pp.411-430.

[9] R. D. Zilca, “Text-Independent Speaker Verification Using Utterance Level Scoring and Covariance Modeling”, *IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING*, VOL. 10, NO. 6, pp.363-370,SEPTEMBER 2002