PAN ARCHIVES OF ACOUSTICS
Vol. **37**, No. 4, pp. 555–559 (2012)

DOI: 10.2478/v10168-012-0065-9

Joint Factor Analysis of Channel Mismatch in Whispering Speaker Verification

Gang LV, Heming ZHAO

School of Electronic Information, Soochow University
Suzhou, 215003, P.R.China; e-mail: lvgang@suda.edu.cn

(*received May 28, 2011; accepted November 5, 2012*)

A speaker recognition system based on joint factor analysis (JFA) is proposed to improve whispering speakers' recognition rate under channel mismatch. The system estimated separately the eigenvoice and the eigenchannel before calculating the corresponding speaker and the channel factors. Finally, a channel-free speaker model was built to describe accurately a speaker using model compensation. The test results from the whispered speech databases obtained under eight different channels showed that the correct recognition rate of a recognition system based on JFA was higher than that of the Gaussian Mixture Model–Universal Background Model. In particular, the recognition rate in cellphone channel tests increased significantly.

Keywords: joint factor analysis, whisper, speaker verification.

1. Introduction

In cellphone conversations in public places, if the messages of one party involve a password, band card number, ID card number, or other sensitive private information, the conversation generally proceeds on a special articulation mode of whispered speech. The other party of the conversation generally understands and agrees. However, criminals hide their identities through whispered speech and conduct financial frauds via cellphone by taking advantages of these secret modes of information exchange. Based on the precondition that clients' privacy protection through whispered conversations is conducted, determining the identity of a whispering speaker has effectively become an urgent issue for financial security sectors.

Compared with the recognition of speakers using normal speech, identifying whispering speakers is more difficult. First, the vocal cord does not vibrate in whispered speech, resulting in the absence of F0 in the speech signals. The first and the second formants shift toward high frequency, and the corresponding bandwidth increase (Gang, Heming, 2009). Therefore, an efficiency parameter that characterizes whispered speech is lacking. Second, during the development of a whispering speaker recognition system, samples of the training data generally are generated from recorded signals in indoor environment with high signal-to-noise ratio (SNR); in contrast, the test samples generally are generated from signals recorded in cellphones with low SNR. The mismatch between training and testing channels results in low recognition rates and inadequate robustness in the whole recognition system.

Regarding the issue of characteristic extraction of whispered speech, many research results have been reported. A system was established using 13 Mel frequency cepstral coefficients (MFCC) as speaker features and Gaussian Mixture Model (GMM) with 128 Gaussian components as speaker models (Jin, *et al.*, 2007). Altogether, from 8% to 33% relative improvement in identification accuracy was achieved compared with the performance under corresponding matched conditions. Fan introduced a whisper speaker ID system based on modified linear frequency cepstral coefficient and feature mapping, which achieved an absolute +20% enhancement compared with an MFCC baseline system (Fan, Hansen, 2009). In another paper, Fan discussed the various differences between whispered and neutral speeches among different speakers and their effect on neutral trained MFCC–GMM systems, and he proposed a confidence space to measure the quality of whispered speech for the task of speaker ID (Fan, Hansen, 2010). However, no articles concerning channel mismatch of whispered speech are available. In this study, the joint factor analysis (JFA)

model proposed by Kenny was adopted (Kenny *et al.*, 2007). First, the model was simplified, and the eigenvoice and the eigenchannel were estimated separately. Then, the speaker and the channel factors were determined using training to obtain a channel-free speaker model. Finally, the trained model was tested. The test result suggested that, for whispered speech signals, the recognition rate of the system based on JFA was superior to that of GMM – Universal Background Model (UBM) under channel mismatch.

This paper proceeds as follows: the JFA and the estimation of the eigenvoice and eigenchannel are introduced in Sec. 2. The training and testing methods for the whispering speaker model are introduced in Sec. 3. In Sec. 4, the whispered speech database is described, and the test environment is presented with test results given. Conclusions are presented in Sec. 5.

2. JFA

JFA is a speaker and session variability model in GMM. There are a number of C Gaussian components in the fixed Gaussian structure, and the mean vector of each Gaussian element is F-dimensional. Thus, a $C \times F$ mean supervector with C Gaussian mean connections can be obtained. Similarly, a $CF \times CF$ hyper-covariance matrix $\mathbf{\Sigma}$ (the elements are on the diagonals) with the covariance matrices connected in series is obtained.

Based on factor analysis, a supervector $\mathbf{M}$ of speaker s can be expressed as Eq. (1)

$$\mathbf{M}(s) = \mathbf{m} + \mathbf{v} \cdot \mathbf{y}(s) + \mathbf{u} \cdot \mathbf{x}(s), \tag{1}$$

where $\mathbf{m}$ represents a speaker- and channel-free mean supervector obtained through UBM training; both $\mathbf{m}$ and the dimension of $\mathbf{M}(s)$ are $CF \times 1$. $\mathbf{v}$ is a low-rank rectangular matrix that represents the eigenvoice, $\mathbf{y}(s)$ is a speaker factor; $\mathbf{u}$ is a low-rank rectangular matrix that represents the eigenchannel, and $\mathbf{x}(s)$ is a channel factor. All these factors are assumed to be independent and satisfy standard normal distribution. JFA estimates the hyper-parameter set $\Lambda = (\mathbf{m}, \mathbf{u}, \mathbf{v})$ in Eq. (1). In the eigenvoice matrix $\mathbf{v}$ and eigenchannel matrix $\mathbf{u}$ estimations, the correlation among factors will be involved if joint estimation is adopted, resulting in large volume of data and high computational cost. The literature suggested that the effects of separate estimation are better than those of the joint estimation when data of each person under multiple channels exist (Kenny *et al.*, 2008). Therefore, the estimation was carried out separately for the two kinds of space. The process is presented in the next sections.

2.1. Estimation of eigenvoice

Step 1: With the UBM trained using the basic Expectation-Maximization (EM) algorithm, the mean supervector $\mathbf{m}$ and hyper-covariance matrix $\mathbf{\Sigma}$ are obtained. $\mathbf{m}$ is formed by the serial connection of m_c, where m_c represents the mean supervector of the UBM Gaussian element c. In the following formula, variables with the c suffix represent the statistical quantities corresponding to the Gaussian function c of GMM, whereas variables without the c suffix represent all the statistical quantities of GMM.

Step 2: Section h of the speech data of speaker s is calculated corresponding to the zeroth-order statistics of the UBM model. The posterior probability of frame t in a whispered speech signal corresponding to UBM is set as $\gamma_t(c)$; thus

$$\begin{aligned} N_{h,c}(s) &= \sum_t \gamma_t(c), \\ N_c(s) &= \sum_h N_{h,c}(s). \end{aligned} \tag{2}$$

Considering $N_c(s)$ as an element in the leading diagonal, a $CF \times CF$ diagonal matrix $N(s)$ is constructed.

Step 3: Section h of the speech data of speaker s is calculated corresponding to the first-order statistics of the UBM model, i.e.,

$$\begin{aligned} F_{h,c}(s) &= \sum_t \gamma_t(c)(Y_t - m_c), \\ F_c(s) &= \sum_h F_{h,c}(s), \end{aligned} \tag{3}$$

where Y_t is the characteristic vector of frame t in a whispered speech signal. With $F_c(s)$ connected in series, a $CF \times 1$ $F(s)$ is obtained.

Step 4: All speaker data are processed according to the following formula, and the expected values of the first- and second-order statistics of speaker factor $\mathbf{y}(s)$ are estimated:

$$\begin{aligned} \mathbf{l}(s) &= I + \mathbf{v}^{\mathrm{T}} \mathbf{\Sigma}^{-1} N(s) \mathbf{v}, \\ E(\mathbf{y}(s)) &= \mathbf{l}^{-1}(s) \mathbf{v}^{\mathrm{T}} \mathbf{\Sigma}^{-1} F(s), \\ E(\mathbf{y}(s)\mathbf{y}(s)^{\mathrm{T}}) &= E(\mathbf{y}(s)) E(\mathbf{y}(s)^{\mathrm{T}}) + \mathbf{l}^{-1}(s), \end{aligned} \tag{4}$$

where $E(\cdot)$ is a solution of the mean value. Here, the channel effect is averaged using averaging operation in all the speech sections of the speakers. Assumption can be made that only the speaker information exists (Kenny *et al.*, 2005).

Step 5: Eigenvoice $\mathbf{v}$ is estimated according to the statistics acquired previously. Let

$$\begin{aligned} \mathbf{\Phi}_c &= \sum_s N_c(s) E(\mathbf{y}(s)\mathbf{y}(s)^{\mathrm{T}}), \\ \mathbf{\Gamma} &= \sum_s F(s) E(\mathbf{y}(s)^{\mathrm{T}}). \end{aligned} \tag{5}$$

For each Gaussian element $c = 1, \ldots, C$ and for each $f = 1, \ldots, F$, set $i = (c-1)F + f$. Let v_i denote the

i-th row of $\mathbf{v}$ and Γ_i denote the i-th row of $\mathbf{\Gamma}$. Then, $\mathbf{v}$ is updated through the following matrix equation:

$$v_i \Phi_c = \Gamma_i (i = 1, \ldots, CF). \tag{6}$$

Steps 4 and 5 are repeated until space $\mathbf{v}$ converges.

2.2. Estimation of eigenchannel

The steps involved in estimating the eigenchannel are the same as those involved in estimating the eigenvoice, but with two differences.

1. In the calculation of the first-order statistics, m is not considered as the center, whereas $\mathbf{m} + \mathbf{v} \cdot \mathbf{y}(s)$ is used as the basis. In other words, the corresponding speaker factor is supposed to be extracted first in the estimation of the eigenchannel so that the next calculation step can be carried out. The purpose of this procedure is to eliminate the corresponding speaker information.

2. The eigenvoice in Eq. (1) is based on the speaker. Theoretically, the space of a person can be estimated according to one section of his speech, but the eigenchannel needs to be estimated with multiple speech sections of a person under various channel conditions. Accordingly, Eq. (4) transforms into

$$\begin{aligned} \mathbf{l}(s) &= I + \mathbf{u}^{\mathrm{T}} \mathbf{\Sigma}^{-1} N_h(s) \mathbf{u}, \\ E(\mathbf{x}_h(s)) &= \mathbf{l}^{-1}(s) \mathbf{u}^{\mathrm{T}} \mathbf{\Sigma}^{-1} F_h(s), \\ E(\mathbf{x}_h(s) \mathbf{x}_h(s)^{\mathrm{T}}) &= E(\mathbf{x}_h(s)) E(\mathbf{x}_h(s)^{\mathrm{T}}) + \mathbf{l}^{-1}(s). \end{aligned} \tag{7}$$

3. Training and testing of the whispering speaker model

3.1. Training of model

The training of the speaker was conducted when the hyper-parameter set $\Lambda = (\mathbf{m}, \mathbf{u}, \mathbf{v})$ of the system was obtained. For a characteristic vector X_t $(t = 1, \ldots, T)$ of an arbitrary speech section, the speaker factor $y(s)$ was estimated according to Eq. (4), and the speaker model was obtained consequently. However, tests showed that the effects of the method were ineffective in recognizing whispering speakers, as shown in Table 2. Therefore, the estimation method, which combined the speaker and the channel factors, was proposed in this study to improve the recognition effect. Specifically, matrices $\mathbf{v}$ and $\mathbf{u}$ are merged, and factors $\mathbf{y}(s)$ and $\mathbf{x}(s)$ are merged. Let

$$\mathbf{w} = \begin{bmatrix} \mathbf{u} & \mathbf{v} \end{bmatrix}, \qquad \mathbf{z} = \begin{bmatrix} \mathbf{y}(s) \\ \mathbf{x}(s) \end{bmatrix}. \tag{8}$$

With the introduction of $\mathbf{w}$ and $\mathbf{z}$ into Eq. (4) for estimation, factor $\mathbf{x}(s)$ of the joint factor $\mathbf{z}$ obtained from the estimation was deleted finally, and the speaker model was obtained.

3.2. Testing of the model

The model testing process uses the integration of the channel factor, i.e.,

$$P(\chi | m_s) = \int P(\chi | m_s, x) N(x | 0, I) \, \mathrm{d}x. \tag{9}$$

In this formula, χ is the characteristic vector of whispered speech, m_s represents the supervector $\mathbf{m}+\mathbf{v}\cdot\mathbf{y}(s)$ of the speaker obtained from the training, and $\mathbf{x} = (\mathbf{x}_1(s), \ldots, \mathbf{x}_h(s))$ is a set of all channel factors.

The integration is difficult to solve; hence, the following formula is generally used for an approximate solution (Kenny *et al.*, 2007):

$$\begin{aligned} \log P(\chi | m_s) \approx{}& \mathrm{tr}\left(\mathbf{\Sigma}^{-1} \mathrm{diag}(F \cdot m_s)\right) \\ & - \frac{1}{2} \mathrm{tr}\left(\Sigma^{-1} \mathrm{diag}(N \cdot m_s \cdot m_s^{\mathrm{T}})\right) \\ & + \frac{1}{2} \left\| \mathbf{l}^{-1/2} \mathbf{u}^{\mathrm{T}} \mathbf{\Sigma}^{-1} F \right\|^2, \end{aligned} \tag{10}$$

where N and F represent the zeroth- and the first-order statistics of the tested speech relative to UBM, $\mathrm{tr}(\cdot)$ represents the solution of the matrix trace, $\mathrm{diag}(\cdot)$ represents the solution of the selected diagonal elements, and $\mathbf{l}^{-1/2}$ represents the selection of the upper triangular matrix.

4. Test results

4.1. Whispered speech database

The test data were obtained from the whispered speech database in Soochow University, composed of whispered speeches recorded from 100 students (80 males and 20 females), aged from 18 years to 27 years, under eight channel environments. The eight recording channels included hand microphone (HM), headset (HS), desktop microphone (DT), earphone-type microphone (EP), recorder pen (RP1), recorder pen+hand microphone (RP2), cellphone recording (CR), and cellphone conversation recording (CC). The corpus was constructed using different types of sentences, including representation of all vowels and consonants. The sampling frequency was 8 KHz, and the quantization precision was 16 bit. A total of 800 different records of whispered speeches were available, and the duration of each record was approximately 2 min. For each record, the first 90 s was cut off to compose database 1, and the last 30 s was cut off to compose database 2; thus, each database has 800 sections.

4.2. Characteristic parameter extraction of acoustics

First, the whispered speech signals were pre-emphasized using a weight factor of 0.97. Then, the beginning and the end were cut off through endpoint

detection. DC removal and normalization were conducted on the preserved parts to guarantee identical ranges of the fluctuation amplitudes under various channels, and a Hamming window with a frame width of 20 ms and frame shift of 6 ms was added. Finally, the characteristic coefficients of 12-dimensional MFCC were extracted from each frame for framing. With the corresponding difference coefficients taken into account, a total of 24 dimensions were obtained.

4.3. UBM and space estimation

We used database1 to train the UBM model. The UBM models were trained separately because of the different proportions of male and female speeches, and the Gaussian element of each UBM was set to 256. A UBM with 512 Gaussian elements was formed after the combination. The linguistic data for the eigenvoice and the eigenchannel were the same as above. The factor number of the speaker was set to 50, and the factor number of the channel was set to 20. From the tests, the system was saturated when these two values were set. Even if the system factor number was increased further, the system performance did not improve substantially; on the contrary, the storage space increased accordingly.

4.4. Test results

The whispered speech recorded through HM channels with high SNR in database2 was considered as the training data. The whispered speeches recorded with the other seven channels in database2 were used as the test data set. For each record in the test data set, we randomly selected a fixed section as test data. The fixed section times selected were 1, 2, and 6 s. We repeated the test 30 times at 1 s fixed time, 15 times at 2 s fixed time, and 5 times at 6 s fixed time. The mean of the test results are shown in Tables 1–3.

Table 1. The recognition rate of whispering speaker in the GMM–UBM system.

test time	recognition rate %						
	HS	DT	EP	RP1	RP2	CR	CC
1 s	46.58	42.53	23.05	32.74	37.63	17.80	6.11
2 s	52.83	52.47	27.94	42.94	46.18	22.29	8.53
6 s	60.78	62.78	32.00	49.78	54.78	30.22	10.33

Table 2. The recognition rate of whispering speaker using JFA with separate estimation of channels.

test time	recognition rate %						
	HS	DT	EP	RP1	RP2	CR	CC
1 s	59.33	72.17	57.17	54.33	60.50	52.17	37.83
2 s	75.67	86.17	69.67	65.17	75.83	66.83	51.50
6 s	90.33	93.00	79.83	78.67	86.00	84.00	68.50

Table 3. The recognition rate of whispered speaker using the JFA with the joint estimation of channels.

test time	recognition rate %						
	HS	DT	EP	RP1	RP2	CR	CC
1 s	66.50	82.00	74.50	65.67	73.17	63.17	49.50
2 s	85.50	95.17	88.83	80.67	86.50	81.00	64.00
6 s	96.00	99.67	95.00	92.17	95.33	96.33	86.50

The results acquired using the GMM–UBM system with no channel compensation added and based on the classical MAP algorithm (Reynolds *et al.*, 2000) are shown in Table 1. The results acquired through JFA with channels estimated separately and JFA with channels estimated jointly are shown in Tables 2 and 3, respectively. Inferences were made from the comparison of the three tables. 1) Compared with the GMM–UBM system, the performance of the system using JFA after channel factor compensation improved significantly in all test channels. In particular, the recognition rate for the channel using cellphone conversation drastically increased. 2) The performance of JFA using joint estimation of channels was superior to that using separate channel estimation. 3) As the test time increased, all recognition rates of the three systems improved gradually, but that of the system using JFA improved faster.

5. Conclusions

In this study, JFA was applied for recognition of whispered speakers. Based on the specific conditions of databases, the advantages of the eigenvoice and the eigenchannel were fully applied to compensate the channels in the characteristic domain of speakers. Compared with the GMM–UBM system, the performance improved significantly. In the JFA application, the following improvements were attained according to the characteristics of the whispered speech databases: 1) the UBM model was obtained directly using the whispered speech databases in the estimation of hyperparameter sets. Thus, conducting iterative updating for m_c in Eq. (3) was unnecessary, and the computational burden was reduced. 2) the JFA approach was adopted: the eigenvoice and the eigenchannel, as well as the speaker and the channel factors, were respectively merged before the speaker model was trained.

From the tests, the recognition rate was observed to increase as the test time lengthened. However, this condition failed to meet the real-time requirement of the system. Therefore, the focus of further research is that, based on the acoustic characteristics of whispering speakers, the system should be compensated simultaneously in the model and in the characteristic domains; thus, higher recognition rates can be attained in the testing of short-time whispered speech signals.

Acknowledgment

This work is supported by the National Natural Science Foundation of China, under Grant No. 61071215 and Canadian Center of Science and Education under Grant No. B2009-122.

References

1. Fan X., Hansen J.H.L. (2009), *Speaker identification with whispered speech based on modified LFCC parameters and feature mapping*, ICASSP, 4553–4556.
2. Fan X., Hansen J.H.L. (2010), *Acoustic analysis for speaker identification of whispered speech*, ICASSP, 5046–5049.
3. Gang L., Heming Z. (2009), *Formant frequency estimations of whispered speech in Chinese*, Archives of Acoustics, **34**, 127–135.
4. Jin Q., Jou S.S., Schultz T. (2007), *Whispering speaker identification*, IEEE International Conference on Multimedia and Expo, 1027–1030.
5. Kenny P., Boulianne G., Dumouchel P. (2005), *Eigenvoice modeling with sparse training data*, IEEE Transactions on Audio, Speech and Language Processing, **13**, 345–354.
6. Kenny P., Boulianne G., Ouellet P., Dumouchel P. (2007), *Joint factor analysis versus eigenchannels in speaker recognition*, IEEE Transactions on Audio, Speech and Language Processing, **15**, 1435–1447.
7. Kenny P., Queller P., Dehak N., Gupta V., Dunouchel P. (2008), *A study of inter-speaker variability in speaker verification*, IEEE Transactions on Audio, Speech and Language Processing, **16**, 980–988.
8. Reynolds D.A., Quatieri T.F., Dumm R.B. (2000), *Speaker verification using adapted Gaussian mixture models*, Digital Signal Processing, **10**, 19–41.