Journal of Signal and Information Processing, **2011, 2, 165-169**
doi:10.4236/jsip.2011.23021 Published Online August 2011 (http://www.SciRP.org/journal/jsip)

MAP-Based Audio Coding Compensation for Speaker Recognition*

Tao Jiang, Jiqing Han

School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China.
Email: {taojiang, jqhan}@hit.edu.cn

Received March 8th, 2011; revised April 20th, 2011; accepted April 27th, 2011.

ABSTRACT

The performance of the speaker recognition system declines when training and testing audio codecs are mismatched. In this paper, based on analyzing the effect of mismatched audio codecs in the linear prediction cepstrum coefficients, a method of MAP-based audio coding compensation for speaker recognition is proposed. The proposed method firstly sets a standard codec as a reference and trains the speaker models in this codec format, then learns the deviation distributions between the standard codec format and the other ones, next gets the current bias via using a small number adaptive data and the MAP-based adaptive technique, and then adjusts the model parameters by the type of coming audio codec format and its related bias. During the test, the features of the coming speaker are used to match with the adjusted model. The experimental result shows that the accuracy reached 82.4*% with just one second adaptive data, which is higher* 5.5*% than that in the baseline system.*

Keywords*: *Audio Coding Compensation, Speaker Recognition, MAP-Based

1. Introduction

Speaker recognition is a technology which extracts speaker information from speech signals to identify the speaker's identity. Most of the former speaker recognition systems are directed for the speech at the high bit rate coding [1,2]. In recent years, with the rapid development of network technology, the speech is encoded with compression for effective transmit, and the bit rate is relatively low, which results in the distortion of speech signal and the decline of speaker recognition performance. Particularly, when the condition of training and testing is codec mismatch, the performance is even worse [3-5]. The compensation method of audio coding influence has been attracting more attentions of a variety of researchers. Many techniques of compensating the degradation caused by this mismatch have been developed. They are roughly grouped into two categories, namely 1) feature compensation, in which the process of feature extraction is modified and 2) model adaptation, in which the parameters of recognition models are adjusted. In [6], four standard speech coding algorithms, *i.e.* GSM (12.2 kbps), G.729 (8 kbps), G.723 (5.3 kbps) and MELP (2.4 kbps) were used for testing the mismatch influence for speaker recognition, and also discussed the effect of score normalization. In [4], two approaches were proposed to improve the performance of Gaussian mixture model (GMM) speaker recognition, which were obtained from the G.729 resynthesized speech. The first one explicitly uses G.729 spectral parameters as a feature vector, and the second one calculates Mel-filter bank energies of speech spectra built up from G.729 parameters. In [7], the effect of the codec in GSM cellular telephone networks was investigated, in which the performance of the text-dependent speaker verification system trained with A-law coded speech and tested with GSM coded speech, as well as that of the system trained with GSM coded speech and tested with GSM coded speech were compared. Several parameter representations which were derived from fast Fourier transform and linear prediction cepstrum coefficients (LPCC) [8] estimates were compared.

Although various researches of mismatch effects caused by different codecs in speaker recognition have been investigated, most of them were related with speech codecs. So far, there are little works on mismatch effects caused by stream media codecs in speaker recognition. In this paper, we study the speaker recognition under stream media codecs, and select four popular known coding or unknown coding algorithm in stream media codecs on

*This work was supported by the National Basic Research Program of China (973 Program, No. 2007CB311100).

the Internet: mp3 (192 kbps, known coding algorithm), rm (64 kbps, unknown coding algorithm), wma (128 kbps, unknown coding algorithm) and ogg (128 kbps, known coding algorithm). We analyze the influence of parameters caused by these codecs, and compensate the distortions in the feature domain. We propose a method of MAP-based audio coding compensation for speaker recognition, which is a model adaptation method. The proposed method first sets a standard codec as a reference and trains the speaker models in this codec format, then learns the deviation distributions between the standard codec format and the other ones, next gets the current bias via using a small number adaptive data and MAP-based adaptive technique, and then adjusts the model parameters by the type of coming audio coding format and the related bias. During the test, the features of coming speaker are used to match with the adjusted model, so as to effectively solve the codec mismatch problem.

2. Influence Analysis of Audio Codecs in LPCC Domain

LPCC is the dominant feature which is frequently used in speaker recognition and speech recognition. In order to implement the compensation of audio codec influence, a statistics analysis of audio codec influence is first conducted in the LPCC domain.

Under a variety of audio codecs, LPCC parameters of distortion deviation h_i is defined as: $h_i = o_i - o_0$, $i \in \{0,1,2,3,4\}$, where o_0 is the standard codec feature, o_i is feature of ith codec, and h_i is the bias between ith codec and the standard one.

The 12 dimensional average characteristics parameters are extracted from four types of coded corpus, each of which includes 20 males and 20 females, with two minutes per speaker. **Figure 1** gives the average deviation of the LPCC parameters.

From **Figure 1**, it can be seen that the low-dimensional deviations are larger, while the high-dimensional ones are smaller.

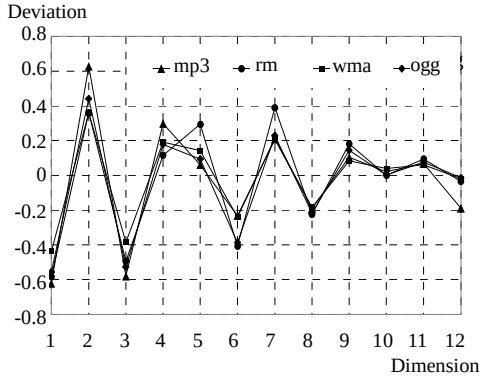

Figure 1. The average deviation of LPCC parameters.

The average values of 12-dimensional characteristics are calculated and the deviations of them are obtained from a speech set. In the speech set, the speaker number is 200 (100 males, 100 females), the time range is from 2 seconds to 30 minutes and there are four types codec. **Figure 2** is the third-dimensional deviation for the mp3 coding, and **Figure 3** is the eleven-dimensional deviation for the wma coding.

From **Figures 2** and **3**, it can be seen that the deviations h change in the vicinity of a certain value. The deviations of low-dimensional characteristics are scattered, and those of high-dimensional ones are small and relatively concentrated. In short, the deviation distribution of LPCC can be described with a single-Gaussian distribution.

3. MAP-Based Coding Compensation

Maximum a posteriori(MAP) [9] is an adaptation approach, which has been widely adopted and successfully applied to speaker adaptation. In this technique, the parameters of the model are regards as random variables, which have an assumed joint prior probability density function (p.d.f.). The MAP estimation of the parameter vectors is defined as the mode of the posterior p.d.f.

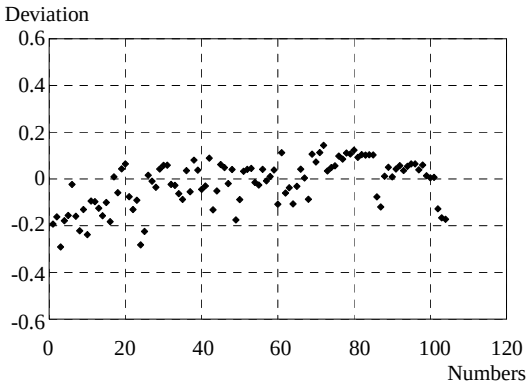

Figure 2. The third-dimensional deviation of mp3.

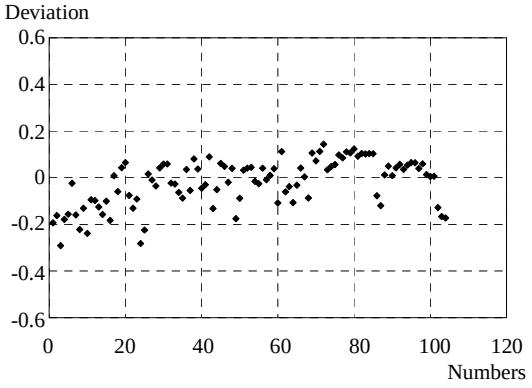

Figure 3. The eleventh-dimensional deviation of wma.

given the adaptation. The MAP-based compensation method is based on two assumptions [6]:

1) Hypothesis that there is the codec bias between the testing and training codecs.

2) Assumption that the deviation distribution can be described with a single-Gaussian distribution (μ,Σ), in which μ is the mean vector and Σ is the covariance matrix.

The formula derivation of MAP codec estimation is described as follows: let h denote the codec bias between the testing and training feature vectors. The codec bias h is characterized by a multivariate Gaussian.

Based on MAP criterion, the codec bias h is estimated by maximizing a posterior probability $p(h|X,\lambda)$, *i.e.*

$$\bar{h}_{MAP} = \arg\max_{h} \{ p(h|X,\lambda) \} \tag{1}$$

where λ is the speaker model, and $X = \{x_1, \cdots, x_T\}$ is the vector sequence of testing speech feature.

This equation is also equivalent to

$$\bar{h}_{MAP} = \arg\max_{h} \{ \log p(X|h,\lambda) + \log p(h) \} \tag{2}$$

where $p(h)$ is the prior knowledge of codec bias.

Thus, the maximization problem is transferred to maximize the sum of log likelihood $\log p(X|h,\lambda)$ and logarithm of a prior p.d.f. $\log p(h)$. This motivates us to introduce a scale factor α into Equation (2) for evaluating the weights of these two terms. Thus, we generalize the MAP estimation as follows,

$$\begin{aligned} \bar{h}_{MAP} = \arg\max_{h} \{ & \alpha \log p(X|h,\lambda) \\ & + (1-\alpha) \log p(h) \} \end{aligned} \tag{3}$$

where $p(X \mid h,\lambda)$ is a mixed Gaussian distribution, *i.e.*

$$\begin{aligned} p(X|h,\lambda) &= \sum_{i=1}^{M} p(X,i|h,\lambda) \\ &= \sum_{i=1}^{M} c_i p_i (X|h,\lambda) \end{aligned} \tag{4}$$

where M is the number of mixture components, and c_i is the mixture weight.

For Equation (3), the current codec bias is estimated using expectation maximum (EM) [10] algorithm in the T-frame data. Function Q can be written as follows:

$$\begin{aligned} Q(h,\bar{h}) = & \alpha \sum_{t=1}^{T} \sum_{i=1}^{M} \frac{p(x_t,i|h,\lambda)}{p(x_t|h,\lambda)} \log p(x_t,i|\bar{h},\lambda) \\ & + (1-\alpha) T \log p(\bar{h}) \end{aligned} \tag{5}$$

where h is the previous iteration result, $\bar{h}$ is the current iteration result, $p(x_t|h,\lambda)$ is the Gaussian mixture density after the adjustment of vector x_t by the bias h, and $p(x_t,i|h,\lambda)$ is the component density.

Assumption that the covariance is a diagonal matrix, let $\partial Q / \partial \bar{h}_j = 0$, we can find

$$\bar{h}_j = \frac{\alpha \sum_{t=1}^{T} \left[\sum_{i=1}^{M} \frac{c_i p_i (x_t|h,\lambda)}{p(x_t|h,\lambda)} \times \frac{(x_{tj} - \mu_{ij})}{\sigma_{ij}^2} \right] + (1-\alpha) \frac{T \mu_{ij}}{\sigma_{ij}^2}}{\alpha \sum_{t=1}^{T} \left[\sum_{i=1}^{M} \frac{c_i p_i (x_t \mid h,\lambda)}{p(x_t|h,\lambda)} \times \frac{1}{\sigma_{ij}^2} \right] + (1-\alpha) \frac{T}{\sigma_{ij}^2}} \tag{6}$$

where $\bar{h}_j$ is the jth component of the current iteration result, x_{tj} is the jth component of vector x_t, μ_{ij} is the mean and σ_{ij}^2 is the covariance of the speaker. μ_{hj} is the mean and σ_{hj}^2 is the covariance of the codec bias.

In Equation (6), μ_{hj} and σ_{hj}^2 are unknown. The codec bias h should to be gotten firstly. Given the scale factor $\alpha = 1$, the Equation (6) is reduced to the maximum likelihood (ML) estimation, *i.e.*

$$\bar{h}_j = \frac{\alpha \sum_{t=1}^{T} \left[\sum_{i=1}^{M} \frac{c_i p_i (x_t|h,\lambda)}{p(x_t|h,\lambda)} \times \frac{(x_{tj} - \mu_{ij})}{\sigma_{ij}^2} \right]}{\alpha \sum_{t=1}^{T} \left[\sum_{i=1}^{M} \frac{c_i p_i (x_t|h,\lambda)}{p(x_t|h,\lambda)} \times \frac{1}{\sigma_{ij}^2} \right]} \tag{7}$$

If there are H types of audio codec, we can get a set of a prior code statistics $\{\bar{h}_{M1}, \bar{h}_{M2}, \cdots, \bar{h}_{MH}\}$, the mean μ_h and variance Σ_h can be estimated using Equations (8) and (9).

$$\mu_h = \frac{1}{H} \sum_{k=1}^{H} \bar{h}_{Mk} \tag{8}$$

$$\Sigma_h = \frac{1}{H} \sum_{k=1}^{H} (\bar{h}_{Mk} - \mu_h)^2 \tag{9}$$

In Equation (7), the initial h is

$$h_0 = \frac{1}{T} \sum_{t=1}^{T} \sum_{i=1}^{M} [c_i \times (x_t - \mu_i)] \tag{10}$$

where x_t is the feature vector in non-standard codec, μ_i is the mean of speaker model in standard codec.

The codec compensated vectors are given by

$$\bar{X} = X - \bar{h}_{MAP} \tag{11}$$

We obtain the final recognition result by finding the maximum a posterior probability for the compensated sequence in standard speaker models.

The overall framework with MAP coding compensation has the following steps: first, select a standard codec and assume that the deviation between the other codecs and the standard codec follows a Gaussian distribution; next, estimates the specific distribution with coding

changing corpus as a codec bias prior knowledge; then, gets current codec bias using a small amount of testing data by MAP algorithm and adjusts the testing data; finally, recognize in standard speaker models and obtains the final result.

4. Experiments and Discussions

In order to evaluate our proposed method, some experiments were designed to test it. The related works are also discussed in detail. The corpus of experiments was collected from Internet, which includes the data of a variety of codec types and speakers. There were 200 speakers, including 100 males and 100 females. The corpus contained news, talks, recitations, interviews and so on. The time duration of speech was from 2 seconds to 20 minutes. The speech from Internet is used as the original which is in the standard codec. We then obtain the mp3, rm, wma and ogg coded speech, which are named as follows: 0 - no codec, 1 - mp3 codec, 2 - rm codec, 3 - wma codec, 1 - mp3 codec.

The number of standard speaker GMM is 128, and the speech used for training per speaker is 5 minutes and that used for testing is from 1 second to 6 seconds. The contents between training and testing data are different. We select 12-dimensional characteristics of LPCC and its difference as the features. In the MAP estimation formula, α needs to be determined. We conducted a series of experiments to compare the performances of using different α values. **Figure 4** shows the accuracy comparisons when the value of α changes from 0.0 to 0.9 and the testing speech is 5 seconds.

From **Figure 4**, we can see that the recognition rate using MAP coding compensation is highest when the adjustment factor α is around 0.5. Under this condition, considering the Equation (6), we may find that the proportion of adaptive data and prior knowledge is close. In the following experiments, the value adjustment factor α is selected as 0.5.

We compared the performances of MAP-based method

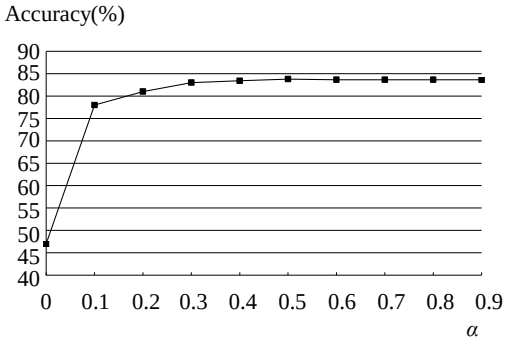

Figure 4. The recognition rate when the value of α is between 0.0 and 0.9.

Table 1. The recognition rates of MAP method when using different adaptive data.

Duration of adaptive speech	Baseline	MAP
0 s	76.9%	–
1 s	–	82.4%
2 s	–	82.3%
3 s	–	82.6%
4 s	–	82.4%
5 s	–	83.1%
6 s	–	82.7%

when using different lengths of adaptive data. **Table 1** gives the results of baseline system and the system using MAP-based method when adaptive data time is from 1 second to 6 seconds.

From the above experimental results, it can be seen that the influence of codec mismatch is very large. The recognition rate of the baseline system is only 76.9%. With using MAP-based compensation method, the system performance was improved effectively under codec mismatch condition. When just using 1 second adaptive data, the accuracy reaches 82.4%, which was 5.5% higher than the baseline system; when the adaptive data is 5 seconds, the recognition rate reaches 83.1%. With the adaptive data increasing, the performance of MAP-based compensation method is gradually close to the former. The performance of the system with 6 seconds adaptive data decreases a little comparing with that of the system with 5 seconds adaptive data. This shows that the codec prior knowledge is useful to improve the system performance when the adaptive speech is little. With the adaptive speech increasing, the effect of the codec prior knowledge would gradually reduce.

5. Conclusions

This paper analyses the effect of the audio coding on speaker recognition parameters LPCC, and introduce MAP technique to compensate the codec mismatch. The proposed method can reduce the influence of training and testing codec mismatch. Experimental results show that with one second adaptive data and using proposed method, an increase of 5.5% in accuracy is obtained comparing with the baseline system. Thus, the proposed method could effectively reduce the influence of training and testing codec mismatch.

REFERENCES

[1] F. Bimbot, J. F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-García, D. P. Delacrétaz and D. A. Reynolds, “A Tutorial

on Text-Independent Speaker Verification," *EURASIP Journal on Applied Signal Processing*, Vol. 4, 2004, pp. 430-451. doi:10.1155/S1110865704310024

[2] T. Kinnunen and H. Li, "An Overview of Text-Independent Speaker Recognition: From Features to Supervectors," *Speech Communication*, Vol. 52, No. 1, 2010, pp. 12-40. doi:10.1016/j.specom.2009.08.009

[3] M. Phythian, J. Ingram and S. Sridharan, "Effects of Speech Coding on Text-Dependent Speaker Recognition," *Proceedings of IEEE Conference Speech and Image Technologies for Computing and Telecommunications*, Vol. 1, Brisbane, December 1997, pp. 137-140.

[4] R. B. Dunn, T. F. Quatieri, D. A. Reynolds and J. P. Campbell, "Speaker Recognition from Coded Speech and the Effects of Score Normalization," 35*th Asilomar Conference on Signals, System s and Computers*, Vol. 2, Pacific Grove, November 2001, pp. 1562-1567.

[5] T. Jiang, B. Y. Gao and J. Q. Han, "Speaker Identification and Verification from Audio Coded Speech in Matched and Mismatched Conditions," *IEEE International Conference on Robotics and Biomimetics*, Guilin, December 2009, pp. 2199-2204.

[6] T. F. Quatieri, R. B. Dunn, D. A. Reynolds, J. P. Campbell and E. Singer, "Speaker Recognition Using G. 729 Speech Codec Parameters," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Istanbul, Vol. 2, June 2000, pp. 1089-1092.

[7] M. G. Kuitert and L. Boves, "Speaker Verification with GSM Coded Telephone Speech," *Proceedings EUROSPEECH* 1997, Vol. 2, Rhodes, September 1997, pp. 975-978.

[8] B. S. Atal, "Effectiveness of Linear Prediction Characteristics of the Speech Wave for Automatic Speaker Identification and Verification," *Journal of the Acoustical Society of America*, Vol. 55, No. 6, 1974, pp. 1304-1312. doi:10.1121/1.1914702

[9] G. L. Gauvain and C. H. Lee, "Maximum a Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains," *IEEE Transactions on Speech and Audio Processing*, Vol. 2, No. 2, 1994, pp. 291-298. doi:10.1109/89.279278

[10] J. Bilmes, "A Gentle Tutorial on the EM Algorithm and Its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models," Technical Report ICSI-TR-97-021, 1997.

