Computational Linguistics and Chinese Language Processing
Vol. 16, No. 1-2, March/June 2011, pp. 1-14

Performance Evaluation of Speaker-Identification Systems for Singing Voice Data

Wei-Ho Tsai* **and Hsin-Chieh Lee***

Abstract

Automatic speaker-identification (SID) has long been an important research topic. It is aimed at identifying who among a set of enrolled persons spoke a given utterance. This study extends the conventional SID problem to examining if an SID system trained using speech data can identify the singing voices of the enrolled persons. Our experiment found that a standard SID system fails to identify most singing data, due to the significant differences between singing and speaking for a majority of people. In order for an SID system to handle both speech and singing data, we examine the feasibility of using model-adaptation strategy to enhance the generalization of a standard SID. Our experiments show that a majority of the singing clips can be correctly identified after adapting speech-derived voice models with some singing data.

Keywords: Model Adaptation, Singing, Speaker Identification.

1. Introduction

As an independent capability in biometric applications or as part of speech-recognition systems, automatic speaker-identification (SID) (Rosenberg, 1976; Reynolds & Rose, 1995; Reynolds, 1995; Campbell, 1997; Reynolds *et al*., 2000; Bimbot *et al*., 2004; Nakagawa *et al*., 2004, 2006; Murty & Yegnanarayana, 2006; Matusi & Tanabe, 2006; Beigi, 2011) has been an attractive research topic for more than three decades. It is aimed at identifying who among a set of enrolled persons spoke a given utterance. Currently, existing SID systems operate in two phases, training and testing, where the former models each person's voice characteristics using his/her spoken data and the latter determines unknown speech utterances based on some comparisons between models and utterances. As the purpose of SID is distinguishing one

*Department of Electronic Engineering & Graduate Institute of Computer and Communication Engineering, National Taipei University of Technology, Taipei, Taiwan
Tel: +886-2-27712171 ext. 2257 Fax: +886-2-27317120
E-mail: whtsai@ntut.edu.tw
The author for correspondence is Wei-Ho Tsai.

person's voice from another's, it is worth investigating if an SID system can not only identify speech voices but also singing voices.

There are a number of real applications where an SID system may need to deal with singing voices. For example, if we record the sounds from TV, it is very likely that the recording contains performers speaking then singing or singing then speaking. In such a case, an SID system capable of handling both speech and singing voices would be very useful to index the recording. Another example is when people gather to sing at a Karaoke. It would be helpful to record everyone's performance onto CDs or DVDs to capture memories of the pleasant time. For the audio in CDs or DVDs to be searchable, audio data would preferably be written in separate tracks, each labeled with the respective person. In this case, an SID system capable of identifying both speech and singing voices will be helpful to automate the labeling process.

To the best of our knowledge, there is no prior literature devoted to the problem of using an SID system to identify singing voices. Most related work (Rosenau, 1999; Gerhard, 2004, 2003) has investigated the differences between singing and speech. Some studies have developed methods for singing voice synthesis (Bonada & Serra, 2007; Kenmochi & Ohshita, 2007; Saino *et al*., 2006; Saitou *et al*., 2005), and some have discussed how to convert speech into singing (Saitou *et al*., 2007) according to the specified melody. In this paper, we begin our investigation by evaluating the performance of an SID system trained using speech voices when the testing samples are changed from speech to singing voices. Then, a well-studied model-adaptation strategy is applied to improve the system's capability in handling singing voices. Our final experiments show that a majority of the singing clips can be correctly identified after adapting speech-derived voice models with some singing data.

The rest of this paper is organized as follows. Section 2 reviews a prevalent SID system. Section 3 describes an improved SID system using some singing data to adapt speech-derived voice models. Then, Section 4 discusses our experiment results. In Section 5, we present our concluding remarks.

2. A Popular Speaker-Identification (SID) System

Figure 1 shows the most prevalent SID system currently, stemming from (Reynolds & Rose, 1995). The system operates in two phases: training and testing. During training, a group of N persons is represented by N Gaussian mixture models (GMMs), λ_1, λ_2, …, λ_N. It is found that GMMs provide good approximations of arbitrarily shaped densities of a spectrum over a long span of time (Murty & Yegnanarayana, 2006); hence, they can reflect the vocal tract configurations of individual persons. The parameters of GMM λ_i, composed of means, covariances, and mixture weights, are estimated using the speech utterances of the i-th person. The estimation consists of k-means initialization and Expectation-Maximization (EM)

(Dempster *et al*., 1977).

Prior to Gaussian mixture modeling, audio waveforms are converted, frame-by-frame, into Mel-scale frequency cepstral coefficients (MFCCs) (Davis & Mermelstein, 1980). The merit of MFCCs lies in the auditory modeling, which has been shown to be superior to other speech-production-based features in numerous studies. Given a test voice sample, the system computes its MFCCs $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_T\}$ and the likelihood probability $\Pr(\mathbf{Y}|\lambda_i)$ for each model λ_i:

$$\Pr(\mathbf{Y} \mid \lambda_i) = \prod_{t=1}^{T} \sum_{k=1}^{K} w_i^{(k)} \cdot \mathcal{N}(\mathbf{y}_t; \boldsymbol{\mu}_i^{(k)}, \mathbf{C}_i^{(k)}) , \tag{1}$$

$$\mathcal{N}(\mathbf{y}_t; \boldsymbol{\mu}_i^{(k)}, \mathbf{C}_i^{(k)}) = \frac{1}{\pi^N \mid \mathbf{C}_i^{(k)} \mid} \exp\left\{ -\left(\mathbf{y}_t - \boldsymbol{\mu}_i^{(k)}\right)' \mathbf{C}_i^{(k)-1} \left(\mathbf{y}_t - \boldsymbol{\mu}_i^{(k)}\right) \right\} \tag{2}$$

where K is the number of mixture Gaussian components; $w_i^{(k)}, \boldsymbol{\mu}_i^{(k)}$, and $\mathbf{C}_i^{(k)}$ are the k-th mixture weight, mean, and covariance of model λ_i, respectively; and prime ($'$) denotes the vector transpose. According to the maximum likelihood (ML) decision rule, the system decides in favor of person I^* when the condition in Eq. (3) is satisfied:

$$I^* = \arg\max_{1 \le i \le N} \Pr(\mathbf{Y} \mid \lambda_i) . \tag{3}$$

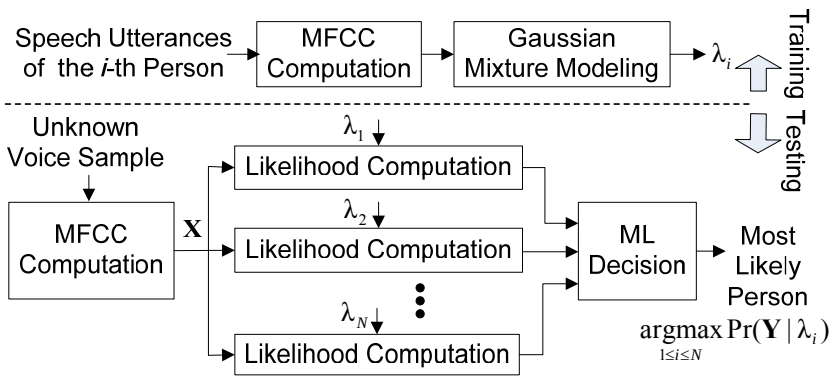


Figure 1. The most prevalent SID system.

3. An SID System Based on Model Adaptation for Singing Voices

Our experiments, discussed in detail in Section 4, find that the above-described SID system performs rather poorly in identifying singing voices of enrolled persons, since a person's singing voice can be significantly different from his/her speech voice. To see if the system can be improved, we apply a well-studied model-adaptation strategy to adapt each person's GMM using some of his/her singing voice data. The adaptation is based on the Maximum A Posterior (MAP) estimation of GMM parameters (Reynolds *et al*., 2000). We assume that the amount of

available singing data for adaptation is very limited; hence, only the mean vectors of GMMs are adapted. For the *i*-th person's GMM, the mean vector of the *k*-th mixture is updated using

$$\hat{\boldsymbol{\mu}}_i^{(k)} = \frac{\tau_i^{(k)}}{\tau_i^{(k)} + \gamma} \bar{\boldsymbol{\mu}}_i^{(k)} + \frac{\gamma}{\tau_i^{(k)} + \gamma} \boldsymbol{\mu}_i^{(k)}, \tag{4}$$

$$\tau_i^{(k)} = \sum_{\ell=1}^{L} \Pr(k \mid \mathbf{x}_\ell, \lambda_i), \tag{5}$$

$$\bar{\boldsymbol{\mu}}_i^{(k)} = \frac{1}{\tau_i^{(k)}} \sum_{\ell=1}^{L} \Pr(k \mid \mathbf{x}_\ell, \lambda_i) \mathbf{x}_\ell, \tag{6}$$

$$\Pr(k \mid \mathbf{x}_\ell, \lambda_i) = \frac{w_i^{(k)} \mathcal{N}(\mathbf{x}_\ell; \boldsymbol{\mu}_i^{(k)}, \mathbf{C}_i^{(k)})}{\sum_{n=1}^{K} w_i^{(n)} \mathcal{N}(\mathbf{x}_\ell; \boldsymbol{\mu}_i^{(n)}, \mathbf{C}_i^{(n)})}, \tag{7}$$

where $\mathbf{x}_\ell$, $1 \le \ell \le L$, are the MFCCs of the available adaptation (singing) data, $\hat{\boldsymbol{\mu}}_i^{(k)}$ is the resulting mean vector after the adaptation, $\mathcal{N}(\cdot)$ is a multivariate Gaussian density function, and γ is a weighting factor of the *a priori* knowledge to the adaptation data. The block diagram of the system based on MAP adaptation is shown in Figure 2.

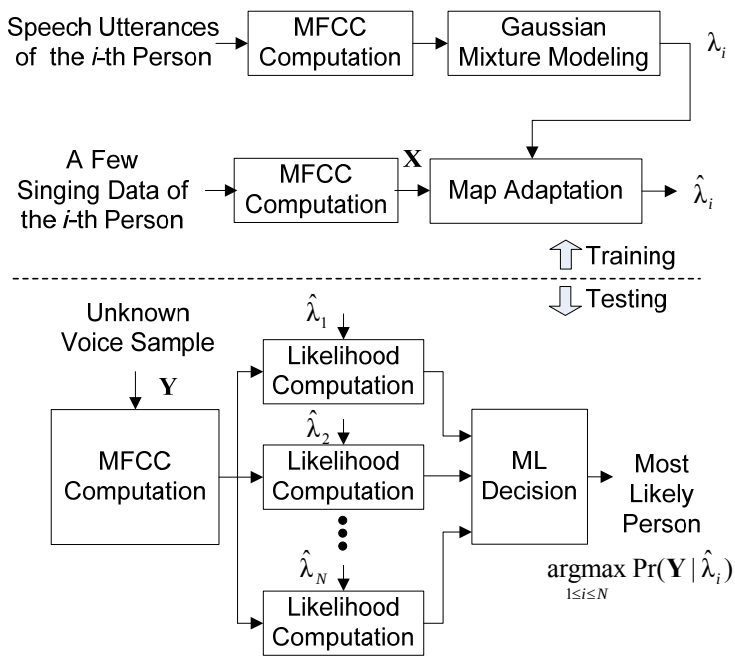


Figure 2. An SID system based on MAP adaptation of a speaker GMM to a singer GMM.

4. Experiments

4.1 Voice Data

We created a database of test recordings ourselves, since no public corpus of voice data currently meets the specific criteria we set up for this study. The database contains vocal recordings by twenty male participants between the ages of 20 and 39. We asked each person to perform 30 passages of Mandarin pop songs using a karaoke machine in a quiet room. All of the passages were recorded at 22.05 kHz, 16 bits, in mono PCM wave. The karaoke accompaniments were output to a headset and were not captured in the recordings. The duration of each passage ranges from 17 to 26 seconds. We denoted the resulting 600 recordings by DB-Singing. Next, we asked each person to read the lyrics of the 30 song passages at a normal speed. All of the read utterances were recorded using the same conditions as those in DB-Singing. The resulting 600 utterances were denoted as DB-Speech.

For ease of discussion in the following sections, we use a term "parallel" to represent the association between a speech utterance and singing recording that are based on the same texts. For example, when the texts are in turn spoken and sung by a person, the speech utterance is referred to as the "parallel" speech utterance of the resulting singing recording, and *vice-versa*. In addition, for use in different purposes, we divided DB-Singing into two subsets, DB-Singing-1 and DB-Singing-2, where the former contains the first 15 recordings per person and the latter contains the last 15 recordings per person. Similarly, DB-Speech was divided into subsets DB-Speech-1 and DB-Speech-2, where the former contains the first 15 speech utterances per person and the latter contains the last 15 speech utterances per person.

4.2 Experiment Results

We used the 15 speech utterances per person in DB-Speech-1 to train each person-specific GMM, and tested the singing recordings in DB-Singing-2. To obtain a statistically-significant experimental result, we repeated the experiment using the 15 speech utterances in DB-Speech-2 to train each person-specific GMM and tested the singing recordings in DB-Singing-1. The number of Gaussian components used in each GMM was tuned to optimum according to the amount of training data. The SID performance was assessed with the accuracy:

$$\text{SID Accuracy (in \%)} = \frac{\#\text{correctly-identified recordings}}{\#\text{ testing recordings}} \times 100\% \,.$$

In addition, to make sure if the system could work well for the conventional SID task, we also evaluated the SID performance using DB-Speech-1 to train each person-specific GMM and tested the speech utterances in DB-Speech 2. Also, in order for the result to be statistically

significant, the experiments were repeated using DB-Speech-2 to train each person-specific GMM before testing the speech utterances in DB-Speech-1. Table 1 shows the SID results. We can see from Table 1 (a) and (b) that the system trained using a set of speech data can perfectly identify the speakers of another set of speech data. Nevertheless, the system fails to identify most persons' voices in DB-Singing-1 and DB-Singing-2. Such poor results indicate the significant differences between most people's speaking and singing voices.

Table 1. Accuracies of the SID systems trained using speech data

(a) System trained using DB-Speeech-1

Testing Data	SID Accuracy (%)
DB-Speech-2	100.0
DB-Singing-2	17.7

(b) System trained using DB-Speeech-2

Testing Data	SID Accuracy (%)
DB-Speech-1	100.0
DB-Singing-1	16.3

Table 2 shows the confusion matrix of the SID results in Table 1. The columns of the matrix correspond to the ground-truth of the singing recording, while the rows indicate the hypotheses. It can be seen from Table 2 that there are a large number of persons whose voice recordings were completely mis-identified. There were only a few people, *e.g.*, #4 and #9, whose singing recordings mostly could be identified well. Further analysis found that persons #4 and #9 are not good at singing, and often cannot follow the tune. They cannot modify their voices properly to make the singing melodious either. Perhaps due to a lack of singing practice, persons #4 and #9 do not change their normal speech voices too much during singing; hence, the system trained using their speaking voices can identify their singing voices well.

To gain insight into the SID errors with respect to different persons, we analyzed the spectrograms of the singing recordings and their parallel speech utterances produced by persons #9 and #10. The waveforms were divided into segments of 512 samples with 50% overlap for the computation of short-term Fourier transform. We can see from Figure 3 (a) and (b) that the formant structure of #9's singing recording is relatively similar to that of his speech utterance, compared with the case of #10, shown in Figure 3 (c) and (d). There is almost no vibrato in #9's singing voice. This is consistent with the observation that #9's voice does not differ too much from speech to singing; thus, it can be handled with speech-derived GMM.

Table 2. Confusion matrix of the SID results in Table 1.

Actual Person Index	Hypothesized Person Index																				Accuracy (%)
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	
1	**3**	1	0	1	2	1	1	0	2	2	3	4	1	1	2	1	3	0	2	0	10.0
2	2	**1**	1	0	1	2	2	1	1	3	2	1	0	0	0	0	5	3	4	1	3.3
3	2	0	**2**	1	1	3	4	2	0	1	4	1	1	3	1	0	0	4	0	0	6.7
4	0	0	0	**25**	0	1	0	1	2	0	0	0	0	1	0	0	0	0	0	0	83.3
5	0	0	3	0	**3**	1	2	1	5	0	6	2	0	0	3	0	0	3	1	0	10.0
6	5	1	0	0	2	**3**	1	4	0	0	3	1	1	0	3	4	1	1	0	0	10.0
7	1	0	0	0	2	3	**2**	0	0	0	3	6	1	0	0	0	8	0	0	4	6.7
8	2	1	0	0	7	0	3	**4**	2	0	5	0	0	4	3	4	1	1	1	0	13.3
9	0	0	0	0	1	0	0	0	**28**	1	0	0	0	0	0	0	0	0	0	0	93.3
10	3	2	0	5	0	0	1	0	2	**2**	1	1	1	4	0	0	5	0	0	3	6.7
11	0	0	4	0	2	0	3	0	1	1	**3**	0	6	0	1	1	2	0	6	0	10.0
12	2	1	1	1	0	3	0	0	5	0	0	**2**	1	1	0	4	6	2	0	1	6.7
13	0	1	1	0	1	3	4	1	1	1	1	2	**3**	2	6	0	0	1	0	2	10.0
14	1	1	1	2	4	0	8	0	0	5	1	0	0	**5**	0	0	1	1	0	0	16.7
15	0	2	6	1	1	0	1	1	0	0	0	3	5	0	**2**	3	3	0	0	2	6.7
16	0	0	0	4	0	7	0	1	1	1	3	3	0	0	0	**3**	2	4	1	0	10.0
17	4	0	8	0	0	1	1	1	0	1	3	0	0	4	0	4	**2**	1	0	0	6.7
18	2	1	1	0	1	1	3	4	0	7	0	0	5	1	1	1	0	**2**	0	0	6.7
19	0	0	0	0	5	1	1	0	1	5	4	0	0	0	2	2	2	3	**4**	0	13.3
20	1	0	8	0	0	2	3	2	1	1	1	1	0	0	0	0	4	0	1	**5**	16.7

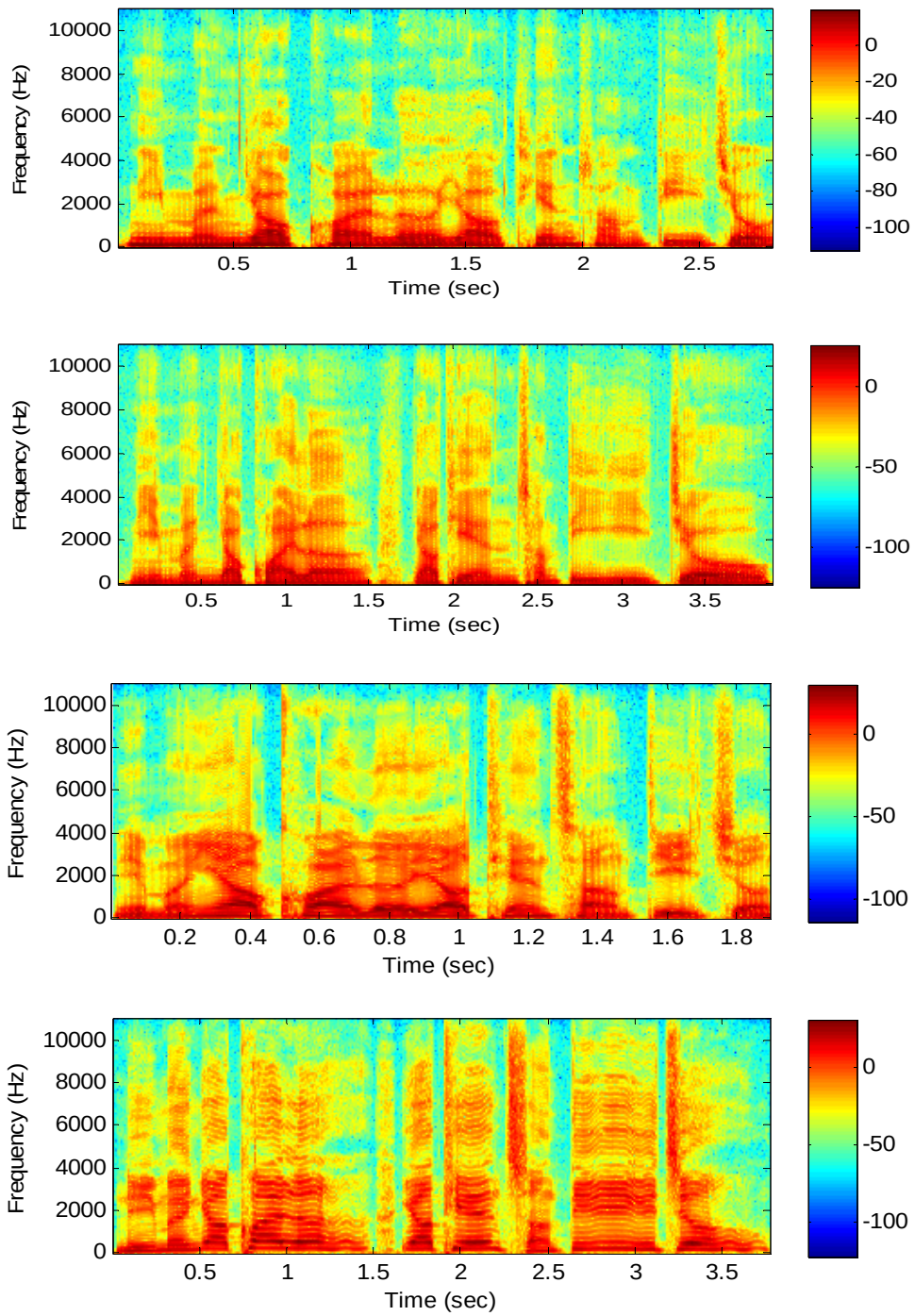
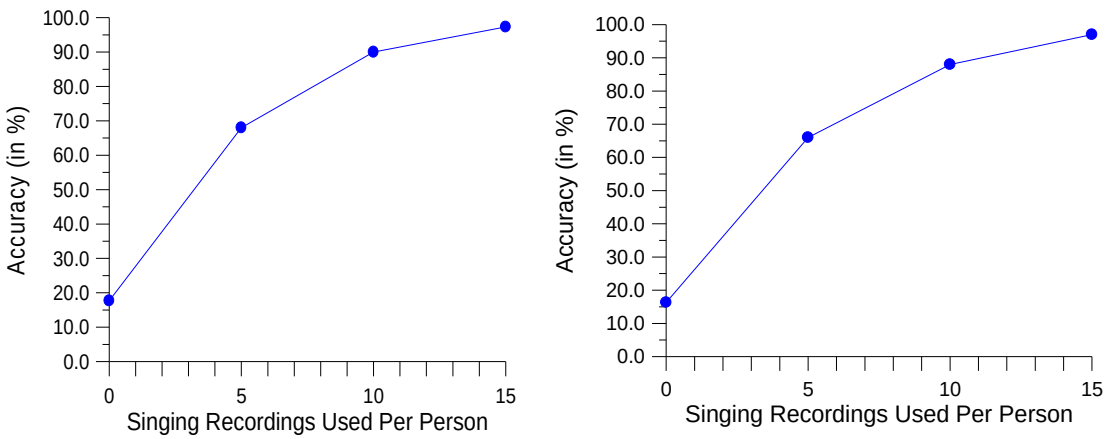


Figure 3. (a) spectrogram of a speech utterance produced by person #9, (b) spectrogram of a singing recording produced by person #9, (c) spectrogram of a speech utterance produced by person #10, and (d) spectrogram of a singing recording produced by person #10, where all the singing recordings and speech utterances are based on the same lyrics: "/ni/ /man/ /iau/ /kuai/ /le/ /iau/ /tian/ /chang/ /di/ /jiou/".

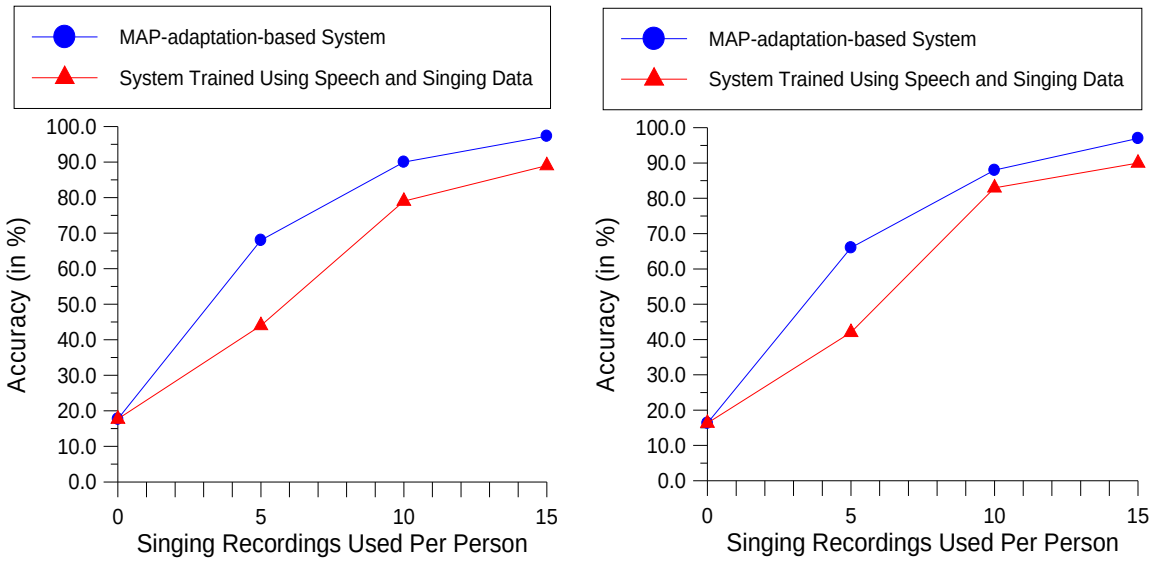
Next, the SID performance of the "MAP-adaptation-based system" described in Sec. 3 was evaluated. We used the 15 speech utterances per person in DB-Speech-1 to train the person-specific GMMs. Each GMM then was adapted using J randomly-selected singing recordings per person in DB-Singing-1, where J = 5, 10, and 15. Based on the adapted GMMs, the system identified the persons of the singing recordings in DB-Singing-2. In addition, to obtain statistically-significant experiment results, we repeated the experiment by using DB-Speech-2 as the training data, DB-Singing-2 as the adaptation data, and DB-Singing-1 as the testing data. The identification accuracy then was computed as the percentage of the correctly-identified recordings. Figure 4 shows the SID accuracies obtained with the MAP-adaptation-based system. It can be seen from Figure 4 that, as expected, the SID accuracies increase with the increase in the amount of singing data used.



(a)Testing DB-Singing-1 ***(b) Testing DB-Singing-2***

Figure 4. SID accuracies obtained with the MAP-adaptation-based System.

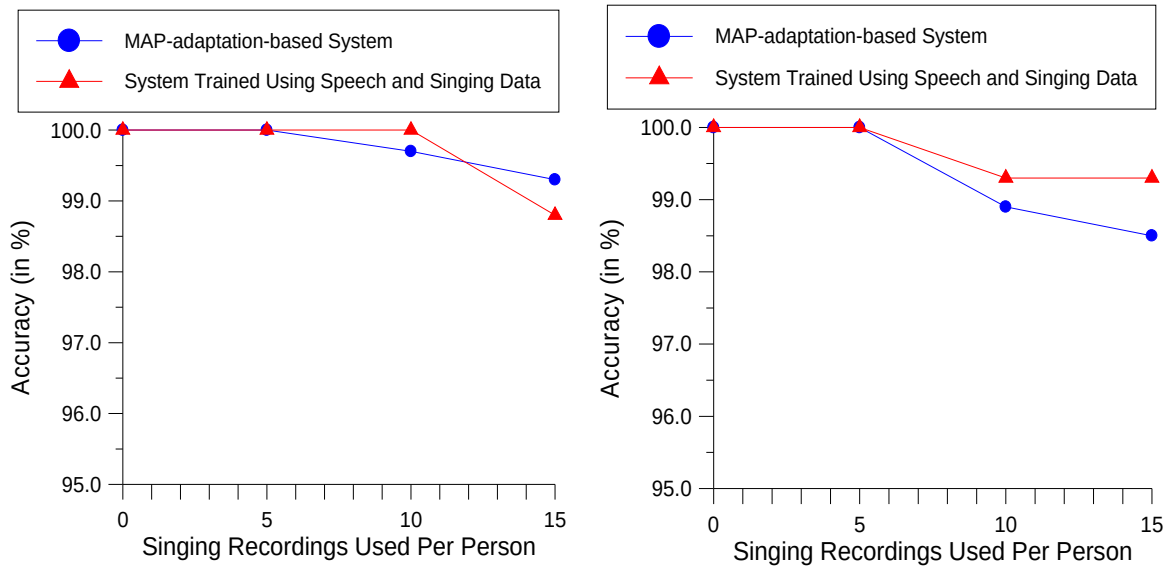
As the MAP-adaptation-based system uses more voice data than the system using speech data only, it is worth comparing the SID performance of the MAP-adaptation-based system with that of the system trained using both speech data and singing data. We thus generated an SID system using 15 utterances plus J singing recordings per person in Gaussian mixture modeling. Figure 5 shows our experiment results. We can see from Figure 5 that the system trained using both speech data and singing data cannot achieve comparable performance to the MAP-adaptation-based system, especially when the amount of singing data is small. This may be because a GMM trained using a mix of speech and singing data tends to model the common voice characteristics of speech and singing, but overlooks their individual differences.



(a)Testing DB-Singing-1 ***(b) Testing DB-Singing-2***

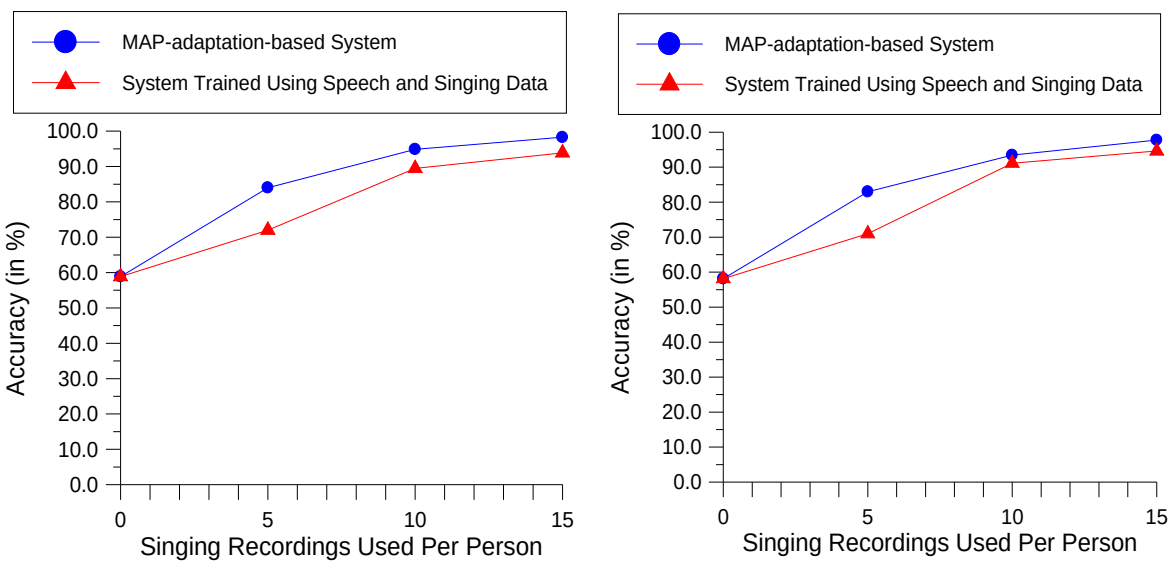
Figure 5. Comparison of the SID performance of the MAP-adaptation-based system with that of the system trained using both speech data and singing data.

In addition, it is worth examining if the MAP-adaptation-based system is still capable of identifying speech data, since its models have been adapted to handle singing data. Figure 6 shows the SID accuracies of testing speech utterances using the MAP-adaptation-based system. For the purpose of comparison, we also evaluated the SID accuracies obtained with the system trained using both speech and singing data. It can be seen from Figure 6 that both of the systems work well in identifying speech utterances. This indicates that the GMMs in the MAP-adaptation-based system do not lose the essence of covering the speaking voice characteristics after they are adapted to cover the singing voice characteristics. Figure 7 presents the accuracies of identifying all of the speech utterances and singing recordings in our database. We can see from Figure 7 that the MAP-adaptation-based system performs better overall than the system trained using both speech and singing data.



(a) Testing DB-Speech-1 ***(b) Testing DB-Speech-2***

Figure 6. Accuracies of identifying speech utterances



(a) Testing DB-Speech-1 and DB-Singing-1 (b) Testing DB-Speech-2 and DB-Singing-2

Figure 7. Accuracies of identifying both speech utterances and singing recordings.

5. Conclusion

In this study, the problem of speaker identification has been extended from identifying a person's speech utterances to identifying a person's singing recordings. Our experiment found that a standard SID system trained using speech utterances fails to identify most singing data, due to the significant differences between singing and speaking for a majority of people. In order for an SID system to handle both speech and singing data, we examine the feasibility of applying a well-known model-adaptation strategy to enhance the generalization of a standard SID. The basic strategy is to use a small sample of the singing voice to adapt each speech-derived GMM based on MAP estimation. The experiments show that, after the model adaptation, the system can identify a majority of the singing clips, while retaining the capability of identifying speech utterances.

Although this study shows that a speech-derived SID system can be improved significantly through the use of a model-adaptation strategy, the system pays the cost of acquiring the singing voice data from each person. In realistic applications, acquiring singing voice data in the training phase may not be feasible. As a result, further investigation on robust audio features invariant to speech and singing would be needed. Our future work will focus on this topic and extend our voice database to a larger scale.

Acknowledgement

This research was partially supported by the National Science Council, Taiwan, ROC, under Grant NSC 98-2622-E-027-035-CC3.

References

Beigi, H. (2011). *Fundamentals of Speaker Recognition*. New York: Springer. ISBN 978-0-387-77591-3, 2011.

Bimbot, F. J., Bonastre, F., Fredouille, C., Gravier, G., Magrin-Chagnolleau, I., Meignier, S., Merlin, T., Ortega-Garcia, J., Petrovska-Delacretaz, D., & Reynolds, D. A. (2004). A tutorial on text-independent speaker verification. *EURASIP J. Appl. Signal Process.*, 430-451.

Bonada, J., & Serra, X. (2007). Synthesis of the singing voice by performance sampling and spectral models. *IEEE Signal Processing Magazine*, 24(2), 67-79.

Campbell, J. P. (1997). Speaker recognition: a tutorial. *Proc. IEEE*, 85(9), 1437-1462.

Davis, S. B., & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Trans. Acoust., Sspeech, Signal Process.*, 28, 357-366.

Dempster, A., Laird, N., & Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. R. Statist. Soc.*, 39, 1-38.

Gerhard, D. (2002). Pitch-based acoustic feature analysis for the discrimination of speech and monophonic singing. *Journal of the Canadian Acoustical Association*, 30(3), 152-153.

Gerhard, D. (2003). *Computationally measurable differences between speech and song*. Ph.D. dissertation, Simon Fraser University.

Kenmochi, H., & Ohshita, H. (2007). VOCALO-ID – commercial singing synthesizer based on sample concatenation. In *Proc. Interspeech*, 4011-4010.

Matusi, T., & Tanabe, K. (2006). Comparative study of speaker identification methods: DPLRM, SVM and GMM. *IEICE Trans. on Information and Systems*, E89-D(3), 1066-1073.

Murty, K. S. R., & Yegnanarayana, B. (2006). Combining evidence from residual phase and MFCC features for speaker verification. *IEEE Signal Process. Lett.*, 13(1), 52-55.

Nakagawa, S., Zhang, W., & Takahashi, M. (2004). Text-independent speaker recognition by combining speaker specific GMM with speaker adapted syllable-based HMM. In *Proc. ICASSP*, I, 81-84.

Nakagawa, S., Zhang, W., & Takahashi, M. (2006). Text-independnt/text-prompted speaker recognition by combining speaker-specific GMM with speaker adapted syllable-based HMM. *IEICE Trans. on Information and Systems*, E89-D(3), 1058-1064.

Reynolds, D. A. (1995). Speaker identification and verification using Gaussian mixture speaker models. *Speech Commun.*, 17(1-2), 91-108.

Reynolds, D., & Rose, R. (1995). Robust text-independent speaker identification using Gaussian mixture speaker models. *IEEE Trans. Speech Audio Process.*, 3(1), 72-83.

Reynolds, D. A., Quatieri, T. F., & Dunn, R. (2000). Speaker verification using adapted Gaussian mixture models. *Dig. Signal Process.*, 10(1-3), 19-41.

Rosenau, S. (1999). An analysis of phonetic differences between German singing and speaking voices. In *Proc. 14th Int. Congress of Phonetic Sciences* (ICPhS).

Rosenberg, A. E. (1976). Automatic speaker verification: A review. In *Proc. IEEE*, 64(4), 475-487.

Saitou, T., Goto, M., Unoki, M., & Akagi, M. (2007). Speech-to-singing synthesis: vocal conversion from speaking voices to singing voices by controlling acoustic features unique to singing voices. In *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA2007)*, 215-218.

Saitou, T., Unoki, M., & Akagi, M. (2005). Development of an F0 control model based on F0 dynamic characteristics for singing-voice synthesis. *Speech Comm.*, 46, 405-417.

Saino, K., Zen, H., Nankaku, Y., Lee, A., & Tokuda, K. (2006). HMM-based singing voice synthesis system. In *Proc. Int. Conf. Spoken Lang. Process.* (*ICSLP*), 1141-1144.