SDBM-Based Speaker Recognition for Speaking Style Variations

Linlin Wang and Mingxing Xu
Department of Computer Science and Technology, Tsinghua University; State Key Laboratory of Intelligent Technology and Systems; Center for Speech and Language Technologies, Tsinghua National Laboratory for Information Science and Technology, Beijing 100084
E-mail: wangll07@mails.tsinghua.edu.cn, xumx@tsinghua.edu.cn Tel: +86-10-62790810-808

***Abstract*— There are many factors corresponding to performance degradation of an actual speaker recognition system. Mismatch in speaking style of a target speaker during training and testing is an important one. When a client enrolls in a system, it is natural for him/her to speak in a spontaneous way. However, it is difficult to maintain the same speaking style throughout test phases. In view of this situation, this paper, based on a database with multiple speaking styles, proposes the concept of Speaking-style-Dependent Background Model (SDBM). The SDBM-based system is presented to train speaking style featured speaker models aiming to alleviate the speaking style mismatch between training and testing. Experimental results show that EER can be reduced by 35.40%.**

I. INTRODUCTION

Speaker recognition is one kind of biometric authentication technology that automatically recognizes a speaker's identity by using speaker-specific information present in speech waves. Performance degradation of a speaker recognition system is due to many factors. For example, environment, recording, and channel conditions, speaker traits (e.g., dialect/accent, stress, speaking style), and spoken language can be considered as different dimensions in the acoustic space. Mismatch between training and testing in any of these acoustic dimensions results in performance degradation in speaker recognition applications [1]. All these mismatches can be divided into two categories. One category is speaker-independent mismatches which originate from voice transmission outside the speaker itself. Environmental noise, echo, recording, and channel mismatches are of this category, which can be named as extrinsic variations. The other category is mismatches in the speaking behavior of the same speaker (e.g., speaking style, time-related variability), which can be named as intrinsic variations.

Most of the researches on speaker recognition have been concentrated on extrinsic variations rather than intrinsic variations. However, in recent years, more and more efforts have been exerted on the study of intrinsic variations in speaker recognition [2]. Among them, speaking style variations constitute an important part, since in practice, it is rather difficult for a speaker to maintain the same speaking style throughout testing phases as in the training phase. Many researchers have studied performance degradation of speaker recognition on emotional speech. [3] is the first study to collectively consider the characteristics of the five speech modes: whispered, soft, neutral, loud and shouted, and their impact on a speaker identification system. Change in speaking rate (normal, fast, and slow) is also a case of speaking style variations [4]. Besides, much attention has been paid to language mismatches [5]. Each of these researches only corresponds to a specific aspect of speaking style variations and the performance degradation brought by such variation has been presented.

Methods to alleviate performance degradation are explored. Wu et al. [6] investigated the applicability of feature modifications of duration, pitch and amplitude parameters for the robustness of speaker recognition over affective speech. A natural-emotion GMM transformation algorithm [7] and an emotion compensation method called emotion attribute projection [8] are also proposed to alleviate the negative effect of emotion mismatch. However, it seems that best recognition results still come from structural or mixed training approach [9]. Scherer et al. [9][10] achieved a better performance on emotional speech by using both neutral and emotional speech to train speaker models. This approach can also be applied to other kinds of speaking style variability, for example, multi-lingual applications [11]. But the approach is questioned by some researchers and they argue that, in many real applications, the training speech of a speaker can involve only one type of emotion (usually neutral), while the testing speech may be uttered in other different emotions [12]. It is unrealistic to make clients speak in all specified speaking style variations during enrollment.

For a sophisticated and human system, when a client enrolls in, it is natural for him/her to speak spontaneously, with normal speaking rate and volume, in a neutral emotional state, and in his/her mother tongue. Furthermore, it should accept utterances of any varied speaking style in the testing phase and identify the exact target client.

In consideration of these above, this paper tries to explore a unified solution for speaking style variation robust speaker recognition. Based on a multi-speaking-style database [13], this paper proposes the concept of Speaking-style-Dependent Background Model (SDBM) and SDBM-based methods are presented to train speaking style featured speaker models aiming to alleviate the speaking style mismatch between training and testing.

The remainder of this paper is organized as follows. Section II introduces the entire SDBM-based speaker recognition system in detail. A brief description of the multi-speaking-style database is shown in Section III. Experimental

results are presented in Section IV and we draw our conclusions in Section V.

II. THE SDBM-BASED SYSTEM

As mentioned above, a typical scenario for common cases is: clients speak in a natural way when training their models; while in testing, there may exist a lot of speaking style variations. Then a critical question is how to narrow the gap between speaking styles of training and testing utterances. The SDBM (Speaking-style-Dependent Background Model) based system aims to narrow this gap by means of speaker models.

A. Principle of The Proposed System

In order to find methods alleviating speaking style mismatches, the state-of-the-art GMM-UBM framework is examined from a systematic point of view. The framework can be represented by a triple: the background model (UBM), client models (through MAP adaptation from UBM), and the scoring strategy (LLR calculation).

Let Λ_B, λ_C and $Score(O_{tst}, C)$ denote the background model, speaker model for client C by adapting O_{trn} from the background model, and the recognition score of O_{tst} against client C, respectively. The representation of this framework is:

$$\begin{cases} \Lambda = \{\Lambda_{BM}, \lambda_C\} = \{\Lambda_{BM}, MAP(O_{trn}, BM)\}, \\ Score(O_{tst}, C). \end{cases} \tag{1}$$

For a classic GMM-UBM system:

$$\begin{cases} \Lambda_{BM} = UBM, \lambda_C = MAP(O_{trn}, UBM), \\ Score(O_{tst}, C) = LLR(O_{tst} \mid \lambda_C) - LLR(O_{tst} \mid UBM). \end{cases} \tag{2}$$

Therefore, mismatch alleviating methods should be explored from the three aspects mentioned above: the background model, client models, and the scoring strategy. Transforming the speaking style independent UBM into speaking style dependent background models (SDBM) is an obviously direct way to cope with speaking style problems. Similarly, client models can be featured by various speaking styles in some approach, which results in a set of speaker models for each client. Finally the scoring strategy, to some extent, deals with how to estimate the speaking style of test utterances, since client models that are of the same speaking style as test utterances should be chosen. The three aspects are illustrated in detail in the following three sections.

B. The Speaking Style Dependent Background Model

Transformation of speaking style independent UBM into speaking style dependent background models (SDBMs) is an obvious way. It is generally believed that the UBM can describe the speech space of the general public. By dividing the UBM space into speaking style dependent subspaces, the SDBM comes into being which attempts to describe the speech subspace related to a certain speaking style.

Suppose N different speaking styles are considered in the speaker recognition system. The relationship between UBM and SDBM can be illustrated by the expression below:

$$UBM = \bigcup_{k=1}^{N} SDBM_k . \tag{3}$$

The SDBM is trained using large amounts of utterances of a certain speaking style. The two models – UBM and SDBM – are isomorphic and can be trained with the same training approach, e.g. the EM algorithm.

C. Speaking Style Featured Modeling Approaches

A set of speaker models for client C corresponding to each speaking style is the target:

$$\lambda_C = \{\lambda_C^k \mid k = 1, 2, \cdots, N\}. \tag{4}$$

How to obtain speaking style featured client models? The presupposition is that training utterances are of the natural speaking style which is denoted by speaking style *1* in the following discussion. The traditional adaptation method is shown below:

$$\Lambda_{BM} = UBM, \lambda_C = MAP(O_1, UBM). \tag{5}$$

In view of this, speaking style featured client models can be obtained from two aspects: the background model and the training utterance. Three modeling approaches are explored:

- Approach One:

$$\Lambda_{BM} = SDBM_k, \lambda_C^k = MAP(O_1, SDBM_k). \tag{6}$$

Client models are directly adapted from SDBMs.

- Approach Two:

$$\Lambda_{BM} = UBM, \lambda_C^k = MAP(O_k, UBM). \tag{7}$$

Since training utterances of speaking style k are not available, an approximation method is needed to substitute their effect. Considering that the MAP algorithm is linear, the following expression is used to approximate the effect:

$$\lambda_C^k \approx MAP(O_1, UBM) + \Delta_k^{UBM}, \tag{8}$$

where Δ_k^{UBM} stands for the difference between training utterances of the natural speaking style and speaking style k in the model level. Suppose there is an M-speaker development set and every speaker has utterances of all N speaking styles. Let $\theta_D^{UBM}(k)$ stand for the model of speaker D trained from the utterance of speaking style k and adapted from the UBM, and the difference is:

$$\delta_D^{UBM}(k) = \theta_D^{UBM}(k) - \theta_D^{UBM}(1). \tag{9}$$

A proper assumption here is that this difference is mainly speaker independent and speaking style dependent. Therefore, Δ_k^{UBM} can be calculated as:

$$\Delta_k^{UBM} = \frac{1}{M} \sum_{j=1}^{M} \delta_j^{UBM}(k). \tag{10}$$

- Approach Three:

$$\Lambda_{BM} = SDBM_k, \lambda_C^k = MAP(O_k, SDBM_k). \tag{11}$$

This is a combination of the above two approaches. It replaces the UBM in the second approach with the SDBMs in the first approach. Some critical expressions are :

$$\lambda_C^k \approx MAP(O_1, SDBM_k) + \Delta_k^{SDBM}, \tag{12}$$

$$\Delta_k^{SDBM} = \frac{1}{M} \sum_{j=1}^{M} \delta_j^{SDBM}(k), \tag{13}$$

$$\delta_D^{SDBM}(k) = \theta_D^{SDBM}(k) - \theta_D^{SDBM}(1). \quad (14)$$

D. Speaking Style Estimation of Test Utterances

Now every client has N speaker models corresponding to each speaking style and the question comes to how to estimate the speaking style of test utterances. Two strategies are explored as follows.

In the blind estimation strategy, the test utterance O_{tst} scores against all these speaker models and a maximum score is chosen as the final recognition score on the target client C. Then the blind estimation can be expressed as equation 15 shows:

$$Score(O_{tst}, C) = \max_{k=1,2,\cdots,N} \left[LLR(O_{tst} \mid \lambda_C^k) - LLR(O_{tst} \mid \Lambda_{BM}) \right]. \quad (15)$$

Unlike blind estimation, in the SDBM-based estimation, the test utterance O_{tst} first scores against all these SDBMs and it is estimated that the test utterance is of the speaking style which gives the maximum score.

$$k = \arg\max_{j=1,2,\cdots,N} LLR(O_{tst} \mid SDBM_j). \quad (16)$$

Then, the final recognition score on the target client C is obtained by scoring against the client model of the corresponding speaking style. This procedure can be written as equation (17):

$$Score(O_{tst}, C) = LLR(O_{tst} \mid \lambda_C^k) - LLR(O_{tst} \mid \Lambda_{BM}), \quad (17)$$

III. The Multi-speaking-style Database

In our previous work [13], a multi-speaking-style database was created. In this database, six aspects of common speaking style variations are taken into consideration, including speaking manner, rate, and volume, emotional and physical state, and language in speaking. Hence, each utterance can be represented by a 6-tuple. Neutral spontaneous speech at normal rate and volume in Chinese is selected as the base scenario. It is the natural speaking style because it covers everyday conversation among the general public in China. There are also 11 other scenarios derived from this principal scenario with only one of the six aspects varies as Fig. 1 shows.

For example, the *fast* scenario is neutral spontaneous speech at normal volume in Chinese, and particularly, at fast rate. Only in the *reading* scenario, newspaper articles are provided. Speech contents are free for other scenarios. Some scenarios, like *reading*, *fast*, *slow*, *English*, *loud*, *soft*, and *whispered*, are easy for participants to finish. For the *denasalized* and *mumbled* scenarios, we provide props: a nose clip to simulate a stuffy nose in bad cold, and sugar candies to simulate talking with things in mouth, respectively. The *angry* and *happy* scenarios are challenging, therefore, we set some daily life scenes and staff are asked to irritate or amuse the participants.

Most of the participants can fulfill the 12 scenarios successfully in an hour. Currently there are 110 persons enrolled in the database, and each person has recorded for about 3 minutes in each scenario. All the recording is done through the same headset, so there is no cross-channel problem.

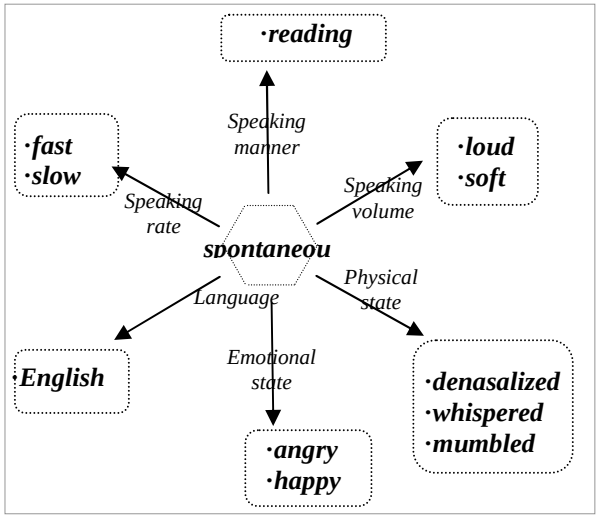

Fig. 1 12 Speech scenarios in the database

IV. Experimental Results

A. Experimental Setup

32-dimensional MFCCs with 16-dim coefficients and 16-dim first derivatives are used as acoustic features, and all the models, such as, UBM, SDBMs, and client models, are of 1024 mixtures. Speech data of 30 persons (around 1080 minutes) in the database are for UBM and SDBMs training, while speech data of 20 persons (around 720 minutes) are for synthesis parameter estimation and others are for verification performance testing. All client models are trained using only utterances of *spontaneous* scenario, and tested against utterances of all the 12 scenarios. All training and testing utterances are around 90 seconds in length.

The classic GMM-UBM system is chosen as the baseline. It does not deal with speaking style mismatches and the UBM is trained using speech data of all 12 scenarios.

B. The SDBM-Based System

In the proposed SDBM-based system, 12 SDBMs are trained separately using the same data as the UBM in the baseline. As can be seen from Section 2, by combining three modeling approaches and two style estimation strategies, there are altogether 6 different workflow combinations under the SDBM-based framework. Table 1 illustrates performance of the seven experiments in EER (equal error rate, %) and EER reduction (%).

TABLE I PERFORMANCE OF THESE SYSTEMS

System		EER	Reduction
Baseline		22.91	----
Modeling	*Style Estimation*	----	----
Approach One	*Blind*	18.33	19.99
	SDBM-based	16.67	27.24
Approach Two	*Blind*	17.99	21.48
	SDBM-based	16.25	29.07
Approach Three	*Blind*	16.54	27.80
	SDBM-based	14.80	35.40

It can be seen from these statistics that all the proposed modeling approaches in the SDBM-based system have achieved significant improvements compared to the baseline. Theoretically speaking, combining the effect of both background models and training utterances, the third modeling approach makes best use of speaking style features. There is no doubt that it yields the best performance, no matter what style estimation strategies are chosen.

C. Some Discussion

No doubt the proposed framework increases computational complexity. However, much computation is done offline, such as, SDBM training and client models adaptation and synthesis. Therefore, the framework is not a considerable challenge to the online process.

It is intended that the 12 speaking styles in the database are independent of each other, while in reality, it is a little difficult for a person to speak angrily with normal rate and normal volume. Therefore, although dependence of speaking style variations in theory does not influence SDBM modeling, data-driven speaking style subspace modeling is perhaps a better way in the proposed SDBM-based framework, and it can also reduce the human annotation burden. Theoretically, the proposed system can also deal with other speaking style variations that are not considered in the database, like time-related variability, in a similar way.

Performance of each separate scenario is also examined. See Table II below (in EER, %).

TABLE II PERFORMANCE OF EACH SEPARATE SCENARIO

Test Scenario	*Baseline*	*Approach Three +SDBM-based*	*EER Reduction*
spontaneous	2.51	1.40	44.22
reading	10.00	5.10	49.00
fast	35.00	25.75	26.43
slow	11.25	10.05	10.67
loud	36.25	28.75	20.69
soft	15.00	10.00	33.33
angry	36.25	29.75	17.93
happy	23.75	18.75	21.05
denasalized	13.75	10.24	25.53
whispered	47.50	30.00	36.84
mumbled	14.97	11.92	20.37
English	7.43	3.28	55.85

Table II shows that, although application of the SDBM-based framework alleviates the speaking style mismatches, the absolute EERs of separate scenarios are still a little higher. Especially the 4 scenarios – *fast*, *loud*, *angry*, and *whispered* – give much worse performance than others. In those scenarios, pronunciation has gone through a dramatic change. Perhaps traditional MFCC features we use are not suitable for this situation and efforts in the model level only lead to limited improvements. Further efforts should be made to find better acoustic features in order to better characterize client identity.

V. CONCLUSIONS

This paper presents an SDBM-based system for speaking style variations. The three proposed speaking style featured modeling approaches have accomplished significant improvements in alleviating speaking style mismatches in training and testing, among which the third approach gives the best performance with the overall EER reduced by 35.40%. For speaking style estimation, the SDBM-based estimation strategy outperforms the blind estimation one. Besides, exploring suitable acoustic features for some special speaking styles need more efforts.

The proposed SDBM-based system deals with the typical situation of training with neutral speech and testing on speech of all speaking styles. A more sophisticated framework that can deal with situations of speaking style-independent training and testing is also a direction of our efforts.

ACKNOWLEDGMENT

This study is supported by Nokia Research Center (China).

REFERENCES

[1] Douglas A. Reynolds, “An Overview of Automatic Speaker Recognition Technology”, Proc. of ICASSP2002, Orlando, pp. 4072--4075, May 2002.

[2] Lei Zhu, Rong Zheng, Bo Xu, “Simplied Residual Factor Analysis for Text-independent Speaker Verification”, Proc. of ICASSP2010, Dallas, 2008.

[3] Chi Zhang, and John H.L. Hansen, “Analysis and Classification of Speech Mode: Whispered through Shouted”, Proc. of InterSpeech2007, Antwerp, pp. 2289--2292, August 2007.

[4] Seiichi Nakagawa, Wei Zhang, and Mitsuo Takahashi., “Text-Independent Speaker Recognition by Combining Speaker-Specific GMM with Speaker Adapted Syllable-based HMM”, Proc. of ICASSP2004, Montreal, pp. 81—84, May 2004.

[5] Murat Akbacak, and John H.L. Hansen, “Language Normalization for Bilingual Speaker Recognition Systems”, Proc. of ICASSP2007, Hawaii, pp. 257--260, April 2007.

[6] Zhouhui Wu, Dongdong Li and Yingchun Yang, “Rules Based Feature Modification for Affective Speaker Recognition”, Proc. of ICASSP2006, Toulouse, 2006.

[7] Zhenyu Shan, Yingchun Yang, Ye Ruizhi, “Natural-Emotion GMM Transformation Algorithm for Emotional Speaker Recognition”, Proc. of InterSpeech2007, Antwerp, 2007.

[8] Huanjun Bao, Mingxing Xu, and Thomas Fang Zheng, “Emotion Attribute Projection for Speaker Recognition on Emotional Speech”, Proc. of InterSpeech2007, Antwerp, 2007.

[9] K.R.Scherer, D.Grandjean, T.Johnstone, G.Klasmeyer, and T.Bänziger, “A Statistical Approach To Assessing Speech And Voice Variability In Speaker Verification”, Proc. of EuroSpeech2003, Geneva, 2003.

[10] K. R. Scherer, T. Johnstone, G. Klasmeyer, “Can Automatic Speaker Verification be Improved by Training the Algorithms on Emotional Speech?”, Proc. of ICSLP2000, Beijing, China.

[11] Bin Ma and Helen Meng, “English-Chinese Bilingual Text-Independent Speaker Verification”, Proc. of ICASSP2004, Montreal, 2004.

[12] Wei Wu, Thomas Fang Zheng, Ming-Xing Xu, and Huan-Jun Bao, “Study on Speaker Verification on Emotional Speech”, Proc. of InterSpeech2006, Pittsburgh, 2006.

[13] Mingxing Xu, Lipeng Zhang, and Linlin Wang, “Database Collection for Study on Speech Variation Robust Speaker Recognition”, Proc. of O-COCOSDA2008, Kyoto, 2008.