Emotional Adaptive Training for Speaker Verification

Fanhu Bie*,Dong Wang*,Thomas Fang Zheng*,Javier Tejedor† and Ruxin Chen‡
*Center for Speech and Language Technologies, Tsinghua University, China
E-mail: biefh@cslt.riit.tsinghua.edu.cn, wangdong99@mails.tsinghua.edu.cn, fzheng@tsinghua.edu.cn, Tel: +86-010-6279-6393
†Human Computer Technology Laboratory, Universidad Autónoma de Madrid, Spain
E-mail: javier.tejedor@uam.es
‡R&D, Sony Computer Entertainment America, Foster City, CA, USA
E-mail: Ruxin_Chen@PlayStation.Sony.com

***Abstract*—Speaker verification suffers from significant performance degradation with emotion variation. In a previous study, we have demonstrated that an adaptation approach based on MLLR/CMLLR can provide a significant performance improvement for verification on emotional speech. This paper follows this direction and presents an emotional adaptive training (EAT) approach. This approach iteratively estimates the emotion-dependent CMLLR transformations and re-trains the speaker models with the transformed speech, which therefore can make use of emotional enrollment speech to train a stronger speaker model. This is similar to the speaker adaptive training (SAT) in speech recognition.**

The experiments are conducted on an emotional speech database which involves speech recordings of 30 speakers in 5 emotions. The results demonstrate that the EAT approach provides significant performance improvements over the baseline system where the neutral enrollment data are used to train the speaker models and the emotional test utterances are verified directly. The EAT also outperforms another two emotion-adaptation approaches in a significant way: (1) the CMLLR-based approach where the speaker models are trained with the neutral enrollment speech and the emotional test utterances are transformed by CMLLR in verification; (2) the MAP-based approach where the emotional enrollment data are used to train emotion-dependent speaker models and the emotional utterances are verified based on the emotion-matched models.

I. Introduction

Speaker verification, or voiceprint recognition (VPR), is widely used in applications such as personal information security check. Despite the significant advance achieved in recent decades, many issues remain unaddressed yet, e.g., the serious performance degradation with adverse environments, dynamic channels and high intra-speaker variations. This paper focuses on a particular intra-speaker variation, i.e., the variation caused by human emotion.

Emotion is an intrinsic nature of human-beings and may change the rendering forms of speech signals significantly. Compared to other intra-speaker variations such as the speech rate and the tones, emotion tends to cause more substantial changes on speech properties, such as harmonic forms, formant structures and the entire temporal-spectral patterns. For speaker verification systems based on the Gaussian mixture model-universal background model (GMM-UBM) architecture [1], these changes lead to considerable difficulties for the UBM/GMMs trained with neutral speech to handle test utterances in different emotions, which in turn results in a serious performance degradation.

A multitude of researches have been conducted to address the emotion variation. The first category involves analysis of various emotion-related acoustic factors, e.g., prosody and voice quality [2], pitch [3], [4], duration and sound intensity [4]. The second category involves various emotion-compensation methods for models and scores. An early investigation was supported by the European VERIVOX project [5], [6], where the researchers proposed a 'structured training' which elicits enrollment speech in various speaking styles. By training the speaker models with the elicited multi-emotional speech, the authors reported reasonable performance improvements. This method, however, is unfriendly and unacceptable in practice. Wu et al. [3] presented an emotion-added model training, where a few amount of emotional data were used to train emotion-dependent models. In addition, [7] compared three types of speaker models (HMM, circular HMM and suprasegmental HMM), and concluded that the suprasegmental HMM is the most robust against emotion changes. Finally, some score normalization and transformation approaches have been proposed to improve emotional speaker verification [8], [9].

In a previous study [10], we proposed an adaptation approach based on the maximum likelihood linear regression (MLLR) [11] and its feature-space variant, the constrained MLLR (CMLLR) [12]. The basic idea is to project the emotional test utterances to neutral utterances by a linear transformation, so that they can be verified with the neutral-trained speaker models. We demonstrated that the MLLR/CMLLR-based adaptation can provide significant performance improvements on emotional test speech, and that the CMLLR-based approach is more effective than the MLLR-based approach. This paper follows this direction and presents a novel emotion adaptive training (EAT) approach. This approach iteratively estimates the emotion-dependent CMLLR transformations and re-trains the speaker models with the transformed speech, which therefore can make use of emotional enrollment speech to train a stronger speaker model. A major difference between the EAT approach and the previously proposed CMLLR adaptation approach is that the former applies the transformation on both the training and the test data, while the latter applies the transformation on the test data only.

The rest of the paper is organized as follows: we first give a brief analysis for emotional speech signals and review the CMLLR technique in Section II. Section III presents the EAT approach. The experiment is reported in Section IV, followed by the conclusions and the future work in Section V.

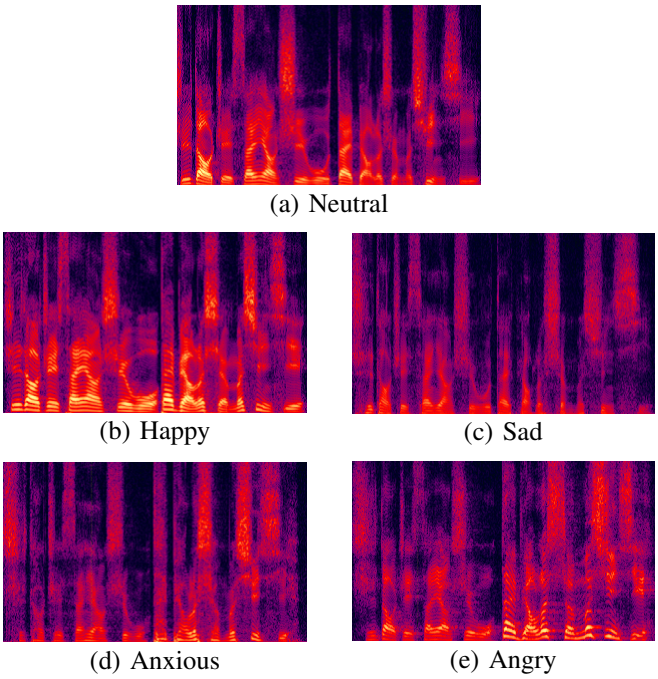
II. Emotion variation and CMLLR

This section presents a brief analysis on spectral characteristics of emotional speech signals, and then gives a quick review of the CMLLR technique. Note that a comprehensive analysis on emotional speech is not the intention of the paper; instead, we just give a rough and intuitive idea about how emotion

change impacts on speech signals, and so motivates the linear adaptation approach. Dedicated study about the impact of emotion change on speech signals can be found in [2], [3], [4], [5], [6], [13].

A. Emotion impact on speech signals

To have an intuitive idea about how emotion changes impact on speech signals, we choose a group of speech segments which were recorded with the same person. The segments are identical in text but are different in emotions. We studied five emotions: neutral, happy, sad, anxious and angry. The spectrograms of these segments are plotted in Figure 1.



(a) Neutral

(b) Happy

(c) Sad

(d) Anxious

(e) Angry

Fig. 1. Spectrograms of speech segments in five emotions.

We observe clear difference among the five spectrograms: the happy and the angry show clearer energy and formant structures than the neutral, while the anxious and the sad go to the opposite. Comparing the happy and the angry, it seems that the happy tends to extend the energy and formant patterns to the high frequency area, while the anger constrains the energy within a low frequency area. In addition, the happy tends to exhibit more significant contrast between speech and non-speech signals. Another pair of emotions, the anxious and the sad, are similar in energy distributions and formant structures, though the latter exhibits more vague speech patterns. These changes in spectrograms have been observed in other segment groups and can be regarded to be general.

Due to the significant change in spectral patterns with different emotions, it is clear suboptimal if we use a speaker model trained with neutral speech to verify test speech in other emotions. We therefore seek for an emotion-dependent model $M_{s,e}$ where s stands for the speaker and e stands for the emotion. Training $M_{s,e}$ directly requires enrollment data for each speaker s and each emotion e, which is often infeasible. We thus resort to an emotion-dependent transformation W_e which can be trained with a set of training speakers, and then applied to test speech of *any* speaker in verification. This can be formulated as follows:

$$p(X; M_{s,e}) \approx p(W_e(X); M_s)$$

where X denotes the test speech, and $p(X; M)$ denotes the probability of X modeled by M. Note that W_e is speaker independent, and therefore can be pre-trained at system design and then applied in system operation. For simplicity, we prefer a linear transformation, which motivates the CMLLR-based adaptation.

B. Review of CMLLR

The MLLR adaptation approach was first proposed by the Cambridge group to deal with channel mismatch and session variability [11], [14]. Its feature-space variant, the constrained MLLR, has been developed to learn transformation on feature vectors [12]. A major advantage of the CMLLR is that the covariance matrices are implicitly adapted besides the mean vectors without increasing the number of training parameters, which often leads to additional gains, as has been demonstrated in our previous study [10]. Another advantage of the feature-space transformation is that it can be used to formulate the EAT approach that we will present in the next section. We therefore focus on the CMLLR.

Define a transformation matrix $W = [b \quad A]$ that projects an input speech signal x_i as follows:

$$\hat{x}_i = Ax_i + b = W\xi_i$$

where A is a rotation matrix and b is a bias term. $\xi_i = [1 \quad x_i]^T$ is the extended observation vector. The optimal W can be attained by maximizing the following likelihood function

$$Q(W; X, M) = \sum_i log(p(W\xi_i; M)) \tag{1}$$

with respect to W, where $M = \{\mu_c, \sigma_c\}$ represents the GMM based on which the CMLLR is conducted. This leads to the following iterative solution:

$$W_l^T = G^{(l)-1}(\alpha p_l + k^{(l)}) \qquad l = 1, 2, 3, ..., L \tag{2}$$

where W_l is the l-th column of W, and p_l is the extended cofactor vector $[0 \; cof(A_{l,1}) \; ... \; cof(A_{l,L})]^T$. $G^{(l)}$ and $k^{(l)}$ are the accumulative statistics, defined by:

$$G^{(l)} = \sum_i \xi_i \xi_i^T \sum_c \frac{r_{c,i}}{\sigma_{c,l}}$$

$$k^{(l)} = \sum_i \xi_i \sum_c \frac{r_{c,i}\mu_{c,l}}{\sigma_{c,l}}$$

where c indexes the Gaussian components, and $r_{c,i}$ is the effective occurrence defined as follows:

$$r_{c,i} = \frac{\mathcal{N}(\boldsymbol{x}_i; \boldsymbol{\mu}_c, \boldsymbol{\sigma}_c)}{\sum_m \mathcal{N}(\boldsymbol{x}_i; \boldsymbol{\mu_m}, \boldsymbol{\sigma_m})}. \tag{3}$$

Finally, $\mu_{c,l}$ and $\sigma_{c,l}$ are the l-th dimension of the mean and diagonal variance vectors of the c-th Gaussian component, respectively. The factor α is solved from the following equation and the root that maximizes the likelihood function is selected:

$$\alpha^2 p_l^T G^{(l)-1} p_l = \alpha p_l^T G^{(l)-1} k^{(l)} - \beta = 0$$

where

$$\beta = \sum_{i,c} r_{c,i}.$$

III. Emotional adaptive training

A. CMLLR for speaker verification

In the previous study [10], we investigated the possibility of using the CMLLR to transform the emotional test utterances so that they can be verified by the neutral-trained speaker models. Specifically, we chose a training set which involves the training speakers $\{S_n\}$. For each speaker S_n in $\{S_n\}$, the speaker model M_n was trained with his/her neutral enrollment speech via the maximum *a posteriori* (MAP) estimation based on the UBM. For each emotion e, the test utterances in emotion e of every speaker S_n in $\{S_n\}$, denoted by $X_{n,e}$, were collected to learn a CMLLR transformation W_e by maximizing the following likelihood function:

$$Q(W_e) = \sum_n Q(W_e; X_{n,e}, M_n)$$

where $Q(W; X, M)$ has been defined in (1). Note that the likelihood function of each speaker is based on his/her own neutral-trained speaker model M_n. Once W_e is trained, it is straightforward to be applied to test utterances[1]. We have demonstrated that this CMLLR-based adaptation approach, although simple, can lead to significant performance improvement for emotional verification when compared to the baseline system where the emotional test utterances are verified with the neutral-trained speaker models directly [10].

B. Emotional adaptive training

An obvious limitation of the CMLLR-based adaptation approach is that the transformations are applied to the test utterances only. This means that the emotional data cannot participate in speaker model training, even though they are available at enrollment. A natural extension is that we can apply the pre-trained $\{W_e\}$ to adapt the emotional *enrollment* speech data if possible, and then use the transformed data to re-train the speaker models. This usually leads to a stronger speaker model due to the increased volume of training data.

In order to apply transformations onto the enrollment data, we need to estimate a set of transformations $\{W_e\}$ based on a single 'pseudo neutral' model for each speaker in $\{S_n\}$. We choose an iterative approach: first estimate $\{W_e\}$ on the neutral-trained speaker models $\{M_n\}$, and then apply the transformed training speech to re-train $\{M_n\}$ via MAP. The re-trained models are then used to re-estimate $\{M_n\}$. This process continues until the convergence criterion is reached. Note that to ensure that the obtained transformations work consistently with the neutral-trained speaker models (which are not iteratively trained with emotional speech) at the enrollment stage, the neutral speech data participate the re-training without transforming. We call this iterative adaptation approach the 'emotion adaptive training', or EAT. The same approach has been successfully employed in speech recognition to deal with speaker variability, in the name of cluster adaptive training (CAT) [15] or the SAT [16]. Note that the goal of the EAT is to estimate a set of transformations $\{M_n\}$ with the *training data* $\{S_n\}$, and the learned transformations are applied to adapt the enrollment speech at the *enrollment* stage. These transformations can certainly be applied to the test utterances at the verification stage as well.

For a clear presentation, we separate the EAT-based approach into three steps: transformation learning, enrollment training and speaker verification, as illustrated in Figure 2. The details are listed as follows:

[1]The emotions of the training and test utterances are assumed to be known in this study.

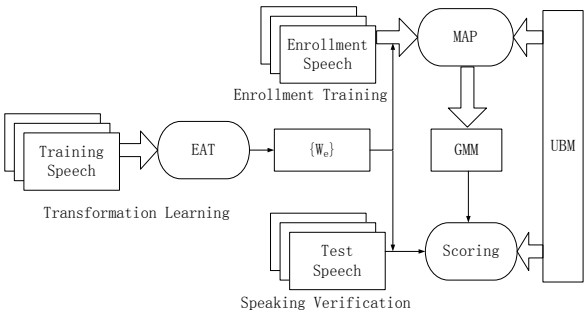


Fig. 2. The EAT-based framework.

- **Transformation Learning:** Define a training set, iteratively train the transformation $\{W_e\}$ and the speaker model $\{M_s\}$, following the EAT algorithm.
- **Enrollment Training:** For each new enrollment k, apply $\{W_e\}$ to transform all the enrollment speech and use the transformed speech data to train the speaker model M_k.
- **Speaker Verification:** For each test utterance in emotion e alleged to be speaker k, first apply W_e to transform the speech features, and then score the transformed speech on model M_k.

IV. Experiments

A. Database

We perform the experiments on an emotional speech database CSLT-ESDB which was recorded in CSLT, Tsinghua University. The recording was conducted with a carbon-button desktop microphone. The sampling rate is 16kHz and the sample size is 16 bits. There are 30 Chinese speakers (15 males and 15 females) in total. For each speaker, a speech segment of 60-90 seconds in the neutral style was recorded for enrollment; in addition, 100 test utterances were recorded in each of the five emotion states: neutral, happy, sad, anxious and angry. Every test utterance involves approximately 15 words and lasts about 5 seconds. These utterances were designed such that all the Chinese syllables are covered as many as possible.

B. Baseline systems

The speaker verification system was designed based on the GMM-UBM framework. The conventional 16-dimensional Mel frequency cepstral coefficients (MFCCs) plus the first order temporal derivatives are used as the acoustic features, and the utterance-based cepstral mean and variance normalization (CMVN) was applied to reduce the channel variation. The UBM was trained on 5 hours of neutral speech (30 males and 30 females). The speaker models are GMMs, and were trained based on the UBM via MAP. Both the UBM and GMMs consist of 1024 Gaussian components; the covariance matrices are set to be diagonal, and are shared by the corresponding Gaussian components of the UBM and the GMMs.

We set up three baseline systems. The first baseline (NMAP) uses the neutral speech to train the speaker model, and then verifies the emotional test speech directly; the second baseline (EMAP) assumes that emotional data are available at enrollment, and uses the emotional enrollment data to train emotion-dependent speaker models via MAP; the third baseline (CMLLR) trains the speaker models with the neutral speech, and employs the CMLLR to adapt the emotional test speech in verification, as presented in our previous work [10]. We choose speech data of 10 speakers from the CSLT-ESDB as the training set, which is used to learn the transformations

	EER%				
	Neutral	Happy	Sad	Anxious	Angry
NMAP	2.19	12.50	16.56	13.26	15.69
EMAP	-	8.06	6.74	6.20	9.57
CMLLR	-	10.50	14.94	12.39	14.20

TABLE I
THE BASELINE EER RESULTS.

	EER%			
	Happy	Sad	Anxious	Angry
EAT-NE	10.56	14.86	12.19	14.31
EAT-EE	5.38	6.36	5.37	7.88

TABLE II
THE EER RESULTS WITH EAT. 'NE' MEANS 'NON-EMOTIONAL ENROLLMENT' AND 'EE' MEANS 'EMOTIONAL ENROLLMENT'.

in the third baseline. The remaining 20 speakers are used for testing. The performance is evaluated in terms of the equal error rate (EER).

The results are reported in Table I. Note that we assume the neutral utterances do not need any special treatment in verification, so the result on the neutral speech is only reported for the NMAP system. We first observe that in the NMAP system, the mismatched emotions (i.e., happy, sad, anxious and angry) lead to tremendous performance degradation when conducting verification with the neutral model. This degradation is particularly serious for the angry emotion, which may partly be attributed to the considerable lost of speech formant patterns in the signal, as has been shown in Figure 1. With the EMAP or the CMLLR, the performance on the emotional test speech is significantly improved, due to the matched training and the emotion transformation respectively. Comparing the EMAP and the CMLLR approaches, we find that the former is more effective than the latter, probably due to the component-based adaptation with the EMAP. However, we notice that the assumption of the EMAP is that the emotional data are available at enrollment, which may not be the case in practice, and thus constrains its application.

C. EAT results

We employ the EAT to improve the speaker model training with the emotional enrollment speech. As in the CMLLR system, we choose 10 speakers to train the transformations in the EAT style, and then apply the obtained transformations at the enrollment and/or the verification stages. The test set involves 20 speakers. We evaluate two scenarios: in the first scenario, the emotional data are not available at enrollment, and so the speaker models are trained with the neutral speech and the transformations are applied at the verification stage only; in the second scenario, the emotional data are available at enrollment, and so the transformations are applied at both the enrollment stage and the verification stage.

The EER results are reported in Table II. We first observe that the performance of the EAT without emotional data at enrollment (non-emotional enrollment, NE) is comparable with the performance of the CMLLR baseline. This indicates that the transformations learned with the EAT are similar to those learned based on the conventional CMLLR. If the emotional data are available at enrollment (emotional enrollment, EE), however, the EAT leads to significant gains: the EERs are not only lower than those obtained with the NMAP system and the CMLLR system, but also lower than those obtained with the EMAP system. This indicates that the transformation-based adaptation may be even better than the emotion-matched model training. This advantage of the emotional enrollment EAT is probably due to the increased training data obtained by the MLLR transformations.

V. CONCLUSIONS

We presented an emotional adaptive training approach to address the emotion variation in speaker verification. By transforming the enrollment data, the EAT may learn stronger speaker models if emotional data are available. The experiments on a 5-emotion database demonstrated that the EAT approach provides highly significant performance improvement, and even outperforms the emotion-matched model training. Future work involves thorough investigation of the transformations on various emotions, and study of transformations on eigen voices. Particularly, we need to collect more emotional data to train robust transformations and ensure the statistical significance.

VI. ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China under Grant No. 61271389 and the National Basic Research Program (973 Program) of China under Grant No. 2013CB329302.

REFERENCES

[1] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1-3.

[2] E. Zetterholm, "Prosody and voice quality in the expression of emotions," in *Proc. ICSLP'98*, 1998, pp. 109–113.

[3] T. Wu, Y.-C. Yang, and Z.-H. Wu, "Improving speaker recognition by training on emotion-added models," in *Proc. Affective Computing and Intelligent Interaction*, 2005, pp. 382–389.

[4] C. Pereira and C. Watson, "Some acoustic characteristics of emotion," in *Proc. ICSLP'98*, 1998, pp. 927–930.

[5] K. R. Scherer, T. Johnstone, G. Klasmeyer, and T. Banziger, "Can automatic speaker verification be improved by training the algorithms on emotional speech?" in *Proc. ICSLP'00*, 2000, pp. 807–810.

[6] K. R. Scherer, D. Grandjean, T. Johnstone, G. Klasmeyer, and T. Banziger, "Acoustic correlates of task load and stress," in *Proc. ICSLP'02*, 2002, pp. 2017–2020.

[7] I. Shahin, "Speaker identification in emotional environments," *Iranian Journal of Electrical and Computer Engineering*, vol. 8, no. 1, pp. 41–46, 2009.

[8] W. Wu, T. F. Zheng, M.-X. Xu, and H.-J. Bao, "Study on speaker verification on emotional speech," in *Proc. Interspeech'06*, 2006, pp. 2102–2105.

[9] Z.-Y. Shan and Y.-C. Yang, "Learning polynomial function based neutral-emotion GMM transformation for emotional speaker recognition," in *Proc. ICPR'08*, 2008, pp. 1–4.

[10] F.-H. Bie, D. Wang, T. F. Zheng, and R.-X. Chen, "Emotional speaker verification with linear adaptation," in *Submitted to ChinaSIP'13*, 2013.

[11] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech and Language*, vol. 9, no. 2, pp. 171–185, 1995.

[12] M. J. F. Gales and P. C. Woodland, "Mean and variance adaptation within the MLLR framework," *Computer Speech and Language*, vol. 10, no. 4, pp. 249–264, 1996.

[13] T. Johnstone and K. R. Scherer, "The effects of emotions on voice quality," in *Proc. 14th International Conference of Phonetic Sciences*, 1999, pp. 2029–2032.

[14] A. Stolcke, S. S. Kajarekar, L. Ferrer, and E. Shrinberg, "Speaker recognition with session variability normalization based on MLLR adaptation transforms," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 15, no. 7, pp. 1987–1998, 2007.

[15] M. Gales, "Cluster adaptive training of hidden Markov models," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 8, no. 4, pp. 417–428, 2000.

[16] T. Anastasakos, J. McDonough, R. Schwartz, and J. Makhoul, "A compact model for speaker adaptive training," in *Proc. ICSLP'96*, 1996, pp. 1137–1140.