Emotion Attribute Projection for Speaker Recognition on Emotional Speech

Huanjun Bao, Mingxing Xu, and Thomas Fang Zheng

Center for Speech Technology, Tsinghua National Laboratory for Information Science and
Technology, Department of Computer Science and Technology,
Tsinghua University, Beijing, 100084, China

`baohj@cst.cs.tsinghua.edu.cn, xumx@tsinghua.edu.cn, fzheng@tsinghua.edu.cn`

Abstract

Emotion is one of the important factors that cause the system performance degradation. By analyzing the similarity between channel effect and emotion effect on speaker recognition, an emotion compensation method called emotion attribute projection (EAP) is proposed to alleviate the intra-speaker emotion variability. The use of this method has achieved an equal error rate (EER) reduction of 11.7% with the EER reduced from 9.81% to 8.66%. When a linear fusion based on a GMM-UBM system with an EER of 9.38% and an SVM-EAP system with an EER of 8.66% is adopted, another EER reduction of 22.5% and 16.1% can be further achieved, respectively, and the final EER can be 7.27%.

Index Terms: speaker recognition, emotional speech, emotion attribute projection, fusion

1. Introduction

In widely used speaker recognition systems, there are many factors that cause performance degradation. These factors include the background noise, the channel effects, the speakers' health condition, and so on. Emotion is another factor that causes the speaker vocal variability.

In real applications, the training speech and the test speech are often not uttered in the same emotion. This kind of mismatch will lead to performance degradation very much [1]. However, there has not been so much work done in this area. In paper [1], an emotional dependent score normalization (E-Norm) method was proposed to solve such a problem in Gaussian mixture model-universal background model (GMM-UBM) systems [2] and the results were reasonable. Emotion-added models proposed in [3] and emotion-state conversion proposed in [4] for speaker recognition showed that the methods for channel compensation could be reasonable for emotion compensation.

This problem of emotion effect is somewhat similar to that of channel effect on speaker recognition. Considering such a similarity between channel effect and emotion effect, it is reasonable to borrow some ideas for the handling of channel effect to alleviate the negative effect of emotion mismatch for speaker recognition. Nowadays, the well-known nuisance attribute projection (NAP) [5, 6] has been proven to be a successful channel compensation method. Is it possible to perform emotion compensation for speaker recognition on emotional speech? The answer is yes. In this paper, an EAP method is proposed for speaker recognition on emotional speech, which was derived from the idea of NAP. The basic idea here is to remove from the SVM expansion dimensions that are irrelevant to the speaker recognition problem on emotional speech. Experimental results will show that the recognition accuracy could be improved after using this proposed method.

Considering the promising results achieved in our previous research [1], a GMM-UBM system will be used in this paper as the baseline. Since the support vector machine (SVM, [7]) with the input of GMM-supervectors [8] has also been very popular, comparison experiments will also be done on GMM-supervector based SVM system using EAP. Both the GMM-UBM system and the SVM system have been proved effective for speaker recognition, experiments will also be done to check the effect when a linear fusion performed on the score level over the GMM-UBM system, and the SVM system combined with EAP.

The rest of this paper is organized as follows. In Section 2, the systems to be used in the experiments in this paper will be introduced. In Section 3, the proposed EAP will be presented in details. The method for fusion will be given in Section 4. The emotional speech corpus, the system description, and experimental results will be presented in Section 5. Conclusion will be drawn in Section 6.

2. Systems of speaker recognition on emotional speech

GMM-UBM is one of the most widely used speaker recognition systems. Studies [1, 3, 4] on speaker recognition on emotional speech show that the GMM-UBM system can achieve promising results. Therefore in this paper, the GMM-UBM system will be taken as the baseline.

The SVM system has also been proven to be an effective method for speaker recognition when incorporated with GMM supervectors, so the SVM system will be used for comparison purpose. In this paper, we will take the output of the GMM-UBM adaptation procedure, i.e., GMM-UBM supervectors, as the input of the SVM system. The GMM-supervector linear kernel introduced in [5] will be used for SVM systems in this paper.

3. Emotion attribute projection (EAP)

It is known that the NAP method can achieve a good result because it removes the subspace that may cause the channel or session variability in the kernel of an SVM system. The idea of the proposed EAP method comes from this and therefore inherits this kind of feature which is expected to be capable of removing the emotion variability. The idea of the EAP will be described in detail as follows.

Let $M(s)$ denote the GMM supervector of speaker s with the neutral emotion. Suppose there are several utterances by the same speaker s with various emotions, $h=1, \dots, H(s)$. For each utterance h, considering the effect of the corresponding speaker and emotion, let $M_h(s)$ denote the supervector correspondingly. Similar to the analysis of channel effects, assume that the differences between $M_h(s)$ and $M(s)$ can be accounted for by a vector of emotion factors $x_h(s)$ having a

August 27–31, Antwerp, Belgium

standard normal distribution. That is to say, we assume that there is a rectangular matrix u of low rank such that

$$\begin{aligned} M_h(s) &= M(s) + M_h(E) \\ &= M(s) + ux_h(s) \end{aligned} \quad (1)$$

for each utterance h=1, ... , $H(s)$, where $M_h(E)$ is the emotion-GMM supervector.

In other words, the emotion supervectors are assumed to be contained in a low-dimensional subspace of the whole supervector space, namely the range of uu^T, which we refer to as the emotion space. See Figure 1 for illustration.
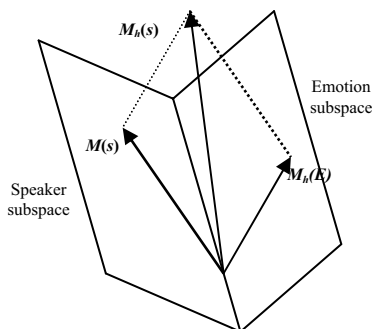

Figure 1: *illustration for decomposition of emotion-dependent speaker model*

The SVM-EAP method works in the way of removing the emotion attributes in the emotion subspace that may cause the variability in the kernel and the EAP constructs a new kernel K,

$$\begin{aligned} K(m^1, m^2) &= [Pb(m^1)]^T [Pb(m^2)] \\ &= b(m^1)^T Pb(m^2) \\ &= b(m^1)^T (I - vv^T) b(m^2) \end{aligned} \quad (2)$$

where m^i is the i-th input vector, $b(\cdot)$ is the SVM expansion, P is a projection ($P^2=P$), v is the direction being removed from the SVM expansion space, and $\|v\|_2=1$. The calculation of v is mainly based on the PCA analysis [6].

4. System fusion

Score fusion was constructed on two subsystems: the GMM-UBM based system and the SVM-EAP based system. For each hypothesized speaker, speaker models were trained separately by the two subsystems. For a given speech utterance, pattern matching was performed in these two subsystems, respectively. A final score was then derived from the score vector through a classifier, which was an SVM used in this paper. Since the results were dependent on the emotion of models and test utterances, it was not appropriate to perform score normalization with zero mean and unit variance. The scores from the subsystems were applied with E-Norm instead.

The SVM classifier used in the score fusion was trained with a part of the developing data which uttered by the same speakers in the SVM imposter cohort dataset (which will be introduced in the next section) without overlapped utterances. And the imposter cohort for training the SVM model was slightly different from the cohort described in the next section by wiping off the utterances of the speaker of the current training speech from the imposter dataset. Afterwards, we can avoid the speaker of training utterances from appearing in the imposter cohort set, which will induce a disaster. Although it was slightly decrease the accuracy as a result of different cohort set, the precision would be assured by using the imposter cohort with a number of utterances by other speakers.

5. Experiments

5.1. Experimental Corpus

A total of 25 male speakers and 25 female speakers were employed to utter sentences in a quiet environment with 5 emotion types: anger, fear, happiness, sadness, and neutral. All of the speakers were native Chinese speakers and they were selected non-professional to avoid exaggerated expression. In this dataset, the utterances for each speaker per emotion contained 30-50 seconds' pure speech of one paragraph, and 20 segments of 2-10 seconds' pure speech of commands or short phrases for each emotion.

Among these 50 speakers, 10 male speakers and 10 female speakers were taken to form the evaluation dataset, in which the paragraph-related speech was used to build the speaker model and the commands or short phrases were utilized as testing utterances. The other 15 male speakers and 15 female speakers utilized as the development dataset were also used to calculate the emotion attribute projection matrix. 7 male speakers and 7 female speakers were utilized as the development dataset for E-Norm. The rest 8 male speakers and 8 female speakers, the paragraph speech and 5 segments of commands or short phrases selected randomly, were used as the development dataset for the SVM imposter cohort.

5.2. System description

In the GMM-UBM system, 16-dimensional MFCC plus delta was taken as a feature vector, where the MFCC feature vector was computed with 20ms frame length every 10ms. Cepstrum mean subtraction (CMS) and cepstrum variance normalization (CVN) were performed over the whole utterance. The UBM was trained with speech from 50 male speakers and 50 female speakers (without overlap with the speakers in the emotional speech corpus). The UBM consisted of 1,024 Gaussian mixtures. Speaker models were adapted from the UBM with Maximum *a Posterior* (MAP) estimation [9] by adapting means only.

The SVM system used the adapted means as input vectors, and utilized the covariance matrix of the UBM to construct the linear kernel.

5.3. Experimental results and analysis

There were 5 sets of experiments concerning the influence of emotion:
■ GMM-UBM based speaker recognition system(as a baseline);
■ SVM system on speaker recognition;
■ SVM system with the proposed EAP method integrated;
■ Fusion between the GMM-UBM system and the SVM-EAP system ; and
■ Performance measuring of most applications, which was trained with neutral speech and tested against the other four emotions'.
E-Norm was used for all of the five sets of experiments.

For all the first 4 sets of experiments, speaker models were trained using speech utterances with anger, fear, happiness, neutral, or sadness, respectively, and then testing utterances with any one emotion were tested against these models. These experiments were designed to study the recognition performances when the training and testing speech were in different emotions.

5.3.1. Baseline system

Table 1 shows the experimental results in EER for the first set of experiments which reached an average EER of 11.02%. Three conclusions could be made as follows.

1) It almost achieved a best result when the emotion of training matched with that of the testing utterances. It can be seen from the results that the mismatched emotion between training and testing speech could be one reason for performance degradation on speaker recognition on emotional speech. Evidences can also be found in Table 2 (organized from data in [10]). However, there is one exception for the emotion of anger, the reason of this exception possibly lies in that the speech speed was slightly faster and the pitch changer was abrupt on stressed.

2) Even when training and testing utterances were uttered with the same emotion, the system also performed differently for different types of emotional speech. This phenomenon can be attributed to different level of vocal variability when speakers were in different emotions.

3) The performance when the models or the testing utterances were with either the neutral or the happiness emotion are better than that when with any of the other 3 emotions. Possibly it is because that the pitch changer of happiness is very smooth and the articulation is normal, both of which are obviously different from the other 3 emotions.

Table 1: *EERs of speaker recognition systems with training and testing speech in varied emotions based on GMM-UBM with E-Norm (%)*

model \ speech	Neutral	Anger	Fear	Happi-ness	Sadness
Neutral	**2.61**	11.25	14.50	7.75	6.64
Anger	10.86	**15.25**	16.42	9.25	12.61
Fear	10.25	12.00	**10.36**	9.50	14.53
Happiness	8.00	7.75	8.00	**6.75**	11.67
Sadness	7.42	14.50	17.36	11.00	**9.06**

Table 2: *Comparison of emotions and speech parameters*

	Anger	Fear	Happi-ness	Sadness
Speech rate	Slightly faster	Much faster	Faster or slower	Slightly slower
Pitch average	Very much higher	Very much higher	Much higher	Slightly lower
Pitch range	Much wider	Much wider	Much wider	Slightly narrower
Intensity	Higher	Normal	Higher	Lower
Voice quality	Breathy, chest	Irregular voicing	Breathy, blaring tone	Resonant
Pitch changer	Abrupt on stressed	Normal	Smooth, upward inflections	Downward inflections
Articulation	Tense	Precise	Normal	Slurring

5.3.2. SVM system

From the second set of experiments based on SVM, we can find that the EERs in this set of experiments are in a much smaller range of 5.69% to 15.50%, compared with the EERs based on the GMM-UBM system whose ERRs ranged from 2.61% to 17.36%, especially when the utterances were tested against the models trained using speech with the anger emotion. It can also be found that it is successful to alleviate the effect anger emotion with SVM system. The reason possibly lies in that the training procedure of model was against number of imposter cohort in which the utterances varied with five kinds of emotions. However, the results of the rest emotions were suffering from slight performance degradation with an average EER of 11.67%.

Table 3: *EERs of speaker recognition systems with training and testing speech in varied emotions based on SVM with E-Norm (%)*

model \ speech	Neutral	Anger	Fear	Happi-ness	Sadness
Neutral	**5.69**	8.00	14.69	10.83	7.44
Anger	11.03	**10.00**	15.61	10.50	13.53
Fear	10.14	9.03	**12.83**	11.72	14.42
Happiness	8.64	9.25	11.75	**10.50**	12.22
Sadness	9.67	10.25	15.50	13.25	**9.50**

5.3.3. SVM-EAP system

In order to alleviate the emotion variability on speaker recognition, EAP was used to remove the corresponding emotion subspace for SVM system with E-Norm. The results are presented in Table 4. It is shown that the EAP works well by comparing the performances based on SVM for almost each section of performance, and finally an average EER of 10.37% can be achieved. Compared with the system based on GMM-UBM, some results were not as good as in the GMM-UBM system. The reason might be that the imposter cohort used when training SVM model was emotion-mixed so that the results will be varied in a smaller range.

Table 4: *EERs of speaker recognition systems with training and testing speech in varied emotions based on SVM with EAP and E-Norm (%)*

model \ speech	Neutral	Anger	Fear	Happi-ness	Sadness
Neutral	**5.50**	8.50	11.67	9.75	8.08
Anger	10.78	**10.11**	13.08	9.00	12.28
Fear	7.92	8.56	**10.25**	10.94	11.53
Happiness	7.25	8.42	10.25	**9.50**	10.44
Sadness	8.75	11.50	13.00	13.00	**7.86**

5.3.4. Systems fusion

A system fusion was also performed on the subsystems, the GMM-UBM based system, and the SVM-EAP based system. For both of these two subsystems, speaker models were trained with utterances of 5 kinds of emotions respectively, and the test utterances of each speaker were in anger, fear, happiness, sadness , or neutral, respectively. The performance is shown as table 5 with an average EER of 9.26%.

Table 5：*EERs of speaker recognition systems with training and testing speech in varied emotions with SVM strategy fusion (%)*

model speech	Neutral	Anger	Fear	Happiness	Sadness
Neutral	**3.06**	7.92	11.72	6.75	6.58
Anger	9.03	**11.00**	13.81	7.75	11.00
Fear	6.75	8.11	**8.64**	8.17	11.83
Happiness	6.03	6.00	7.61	**5.50**	10.00
Sadness	7.00	10.33	13.06	10.17	**7.81**

5.3.5. *Which is the best?*

From the results shown above, we cannot see which system is the best. That was why this set of experiments was designed. In this set of experiments, the speaker models were trained using speech utterances with neutral emotion, while the test utterances varied in the other 4 emotions, anger, fear, happiness, and sadness. This is the situation in many real applications, where speakers were enrolled in the system with neutral speech but might be in varied emotions during verification. The performance is shown in Figure 2.
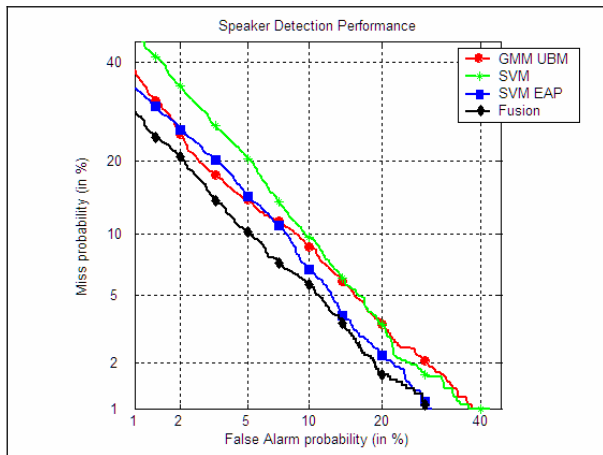

Figure 2: *DET curves of GMM-UBM based system, SVM system, SVM system with EAP, fusion with GMM-UBM system and SVM system with EAP*

The EER of the baseline, i.e. the GMM-UBM based system, was 9.38%, and the system with SVM was 9.81%, which was slightly worse than the baseline. When it was incorporated with the EAP method, it achieved an EER reduction of 11.7% with the EER reduced from 9.81% to 8.66%; when a linear fusion based on the GMM-UBM system and the SVM-EAP system was adopted, another EER reduction of 22.5% and 16.1% could be further achieved over the two subsystems, respectively, with the final EER of 7.27%.

6. Conclusion

In this paper, two methods are used for alleviating the emotion effects on speaker recognition on emotional speech. One of which is the emotion compensation method called EAP which has been proved successful with an EER reduction of 11.7%. The other one is the linear fusion over two subsystems, the GMM-UBM based system and the SVM with EAP system, with the final EER of 7.27%. The influence of the emotion involved in the speaker recognition tasks is

also studied. First, different level of vocal, pitch, articulation variability will exist when speech is uttered in different emotions. As a result, the performance will degrade in speaker recognition even when the emotion matches between the training and the testing speech. Second, the mismatch of emotions between the training and testing speech will reduce the recognition performance further. Third, considering the similarity between the channel effect and emotion effect, it is reasonable to borrow the methods from channel compensation to alleviating the emotion effects.

In our future work, more methods for alleviating the channel effects should be introduced to the speaker recognition on emotional speech. The factors that might cause the performance degradation, such as pitch and articulation, should be further studied as well as emotion-state conversion. As the corpus used in this paper is in Chinese, further experiments can be applied at forensic databases to examine the EAP method.

7. Acknowledgement

8. References

[1] Wei Wu, Thomas Fang Zheng, Mingxing Xu, Huanjun Bao, "Study on Speaker Verification on Emotional Speech", Interspeech 2006-ICSLP, 2102-2105, Pittsburgh, Pennsylvania.

[2] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models", Digital Signal Processing, 10: 19-41, 2000

[3] Dongdong Li, Yingchun Yang, Zhaohi Wu, Tian Wu, "Emotion-state conversion for speaker recognition", Affective Computing and Intelligent Interaction, 2005, 3784: 403-410, Beijing, China.

[4] Tian Wu, Yingchun Yang, Zhaohui Wu, "Improving speaker recognition by training on emotion-added models", Affective Computing and Intelligent Interaction, 2005, 3784:382-389, Beijing, China.

[5] Campbell, W.M., Sturim, D.E., Reynolds, D.A, "SVM based speaker verification suing a GMM supervector kernel and NAP variability compensation", Signal Processing Letters, 2006, 13(5): 308-311.

[6] Solomonoff Alex, Campbell W. M., and Boardman I., "Advances in channel compensation for SVM speaker recognition", in Proceedings of ICASSP, 2005.

[7] Nello Cristianini and John Shawe-Taylor, "Support Vector Machines", Cambridge University Press, Cambridge, 2000.

[8] P. Kenny and P. Dumouchel, "Experiments in speaker verification using factor analysis likelihood ratios", in Proc. Odyssey04, 2004, 219–226.

[9] J.-L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains", IEEE Transction Speech and Audio Processing, 2(2):291–298, 1994

[10] Cowie R., Douglas-Cowie E., Tsapatsoulis N., et al, "Emotion recognition in human-computer interaction", IEEE Signal Processing Magazine, 18(1): 32-80, 2001.