Study on Speaker Verification on Emotional Speech

Wei Wu, Thomas Fang Zheng, Ming-Xing Xu, and Huan-Jun Bao

Center for Speech Technology, Tsinghua National Lab for Information Science and Technology, Tsinghua University, Beijing, 100084, China

{wuwei, baohj}@cst.cs.tsinghua.edu.cn, {fzheng, xumx}@tsinghua.edu.cn

ABSTRACT

Besides background noise, channel effect and speaker's health condition, emotion is another factor which may influence the performance of a speaker verification system. In this paper, the performance of a GMM-UBM based speaker verification system on emotional speech is studied. It is found that speech with various emotions aggravates the verification performance. Two reasons for the performance aggravation are analyzed, they are mismatched emotions between the speaker models and the test utterances, and the articulating styles of certain emotions which create intense intra-speaker vocal variability. In response to the first reason, an emotion-dependent score normalization method is proposed, which is borrowed from the idea of Hnorm.

Index Terms: speaker verification, emotional speech

1. INTRODUCTION

The performance of speaker verification is affected by many factors, such as background noise, channel effect, and speaker's health condition. Any of these factors, either from external sources (background noise and channel effect) or from internal source (speaker's health condition), can bring about negative influence on speaker verification by inducing extra intra-speaker vocal variability. In addition, emotion is another internal source which can induce intra-speaker vocal variability. The study in [1] has shown that different emotions, such as anger, despair, sadness, and happiness, can induce different intra-speaker vocal variability to speaker's voice, and thus might aggravate the performance of speaker verification. In recent studies on affective computing, Mel-scale frequency cepstral coefficients (MFCC), linear predictive cepstral coefficients (LPCC), and prosodic features were used to perform speech emotion recognition, and promising results have been achieved [2][3][4]. These features for speech emotion recognition are also the low- and high-level features which are widely utilized in speaker verification [5], and it therefore proves indirectly that emotion might affect the results of speaker verification. In this paper, the influence of emotion on the performance of a GMM-UBM [6] based speaker verification system is studied. Specifically, verification performance of mismatched emotions between training and testing speech is compared and analyzed. Experimental results show that emotion, similar to background noise and channel effect, also has negative effects on the performance of GMM-UBM based speaker verification systems. And it is discovered that two reasons are responsible for the decline of verification performance on emotional speech: mismatched emotions between the speaker models and the test utterances, and the articulating styles of certain emotions which create intense intra-speaker vocal variability

To alleviate negative effects of emotion on speaker verification systems, [7] introduced a strategy to involve speech with mixed emotions in model training, and it achieved a better verification performance on emotional speech. However, 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. This situation is similar to the problem of channel effect, in which case the training speech is usually recorded through one channel, while the testing speech may come from other different channels. Because of the similarity between channel effect and emotion effect, it is possible to borrow some ideas for handling the channel effect to alleviate the negative influence of emotion. In this paper, an emotion-dependent score normalization for speaker verification on emotional speech is proposed. It is derived from Hnorm [8], which was designed to alleviate channel effect on speaker verification. Experimental results show that the verification performance can be improved after this emotion-dependent score normalization is performed.

The remainder of this paper is organized as follows. In Section 2, the emotional speech corpus and the GMM-UBM based speaker verification system are introduced. In Section 3, the performance of this system on emotional speech is compared and analyzed. In Section 4, the emotion-dependent score normalization is proposed and the experimental results are given. In Section 5, conclusions are drawn and future research focuses are suggested.

2. CORPUS AND SYSTEM DESCRIPTION

2.1 Emotional Speech Corpus

This emotional speech corpus includes 5 types of acted emotions: anger, fear, happiness, sadness, and neutral. Non-broadcasting speakers were selected to avoid exaggerated expression. A total of 25 male and 25 female standard Chinese speakers were employed to utter the 5 emotion types in a quiet environment.

Speech from 15 male and 15 female speakers composes the evaluation dataset. In this dataset, each

This work is supported by National Natural Science Foundation of China under grant 60433030

speaker utters a short passage and 20 command phrases in each type of emotions. The context of the short passages is specifically designed to help elicit corresponding emotions. The utterance of the short passage for each type of emotions is used for training the speaker model of corresponding emotion. Each of these utterances contains 30 to 50 seconds of pure speech (after silence elimination). Utterances of the 20 command phrases in each type of emotions are used as test samples for verification, each of which contains 2 to 10 seconds of pure speech.

Speech from the remaining 10 male and 10 female speakers composes the development dataset for emotion-dependent score normalization. Each of the speakers in this dataset has 20 utterances in each type of emotions. These utterances are short command phrases, containing 2 to 10 seconds of pure speech.

2.2 System Description

The speaker verification system used for experiments in this paper was based on traditional GMM-UBM [6]. Features used were 16-dimensional MFCCs plus delta, computed with 20 ms frame length every 10 ms. Cepstral mean subtraction (CMS) was performed over each whole utterance. The universal background model was trained with neutral speech from 50 male and 50 female speakers different from those in the emotional speech corpus, and consisted of 1,024 Gaussian mixtures. Speaker models were adapted from the universal background model with MAP by adapting means only.

3. SPEAKER VERIFICATION RESULTS ON EMOTIONAL SPEECH

In this section, two sets of experiments concerning the influence of emotion on a GMM-UBM based speaker verification system are presented and analyzed.

In the first set of experiments, speaker models were trained with speech in neutral emotion, and the test utterances of each speaker were in anger, fear, happiness, sadness and neutral, respectively. All the speakers and test utterances in the evaluation dataset were used in this set of experiments, each test utterance was verified against models of the same gender. Figure 3.1(a) gives the verification results on speech in neutral and mixed emotion. It can be seen that the verification performance will decline greatly when the test utterances are involved with various emotions. Figure 3.1(b) gives the verification results on speech in five types of emotions respectively. The verification performance for speech in anger, fear or happiness declines more than that for speech in sadness. A possible reason is that when speakers are in the emotion of anger, fear or happiness, their mood is in a much higher arousal level [9] than that of sadness. Hence, the greater discrepancy between the training and testing speech leads to a greater decline in verification performance.

The second set of experiments was designed to study the verification performance with training and testing speech in different emotions. Five speaker models were trained with speech of anger, fear, happiness, sadness and neutral for each speaker in the evaluation dataset, respectively. For each group of speaker models of the same emotion, test utterances in the five types of emotions were verified respectively. The equal error rate (EER) under each verification condition is presented in Table 3.2.

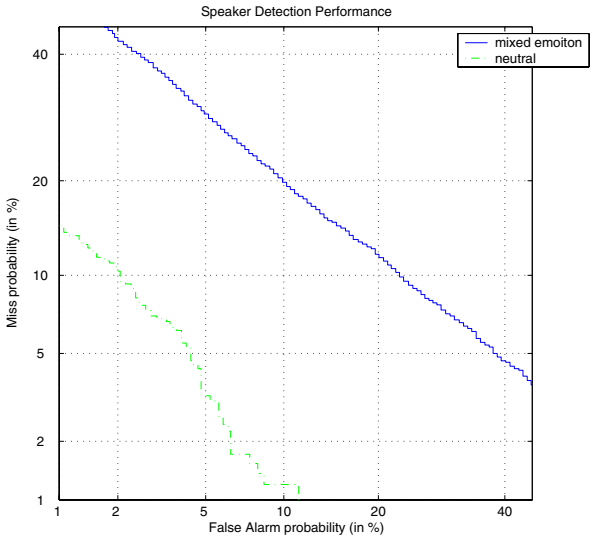

Figure 3.1 (a): DET curve when models are trained with neutral speech and tested with neutral and mixed emotional speech, respectively

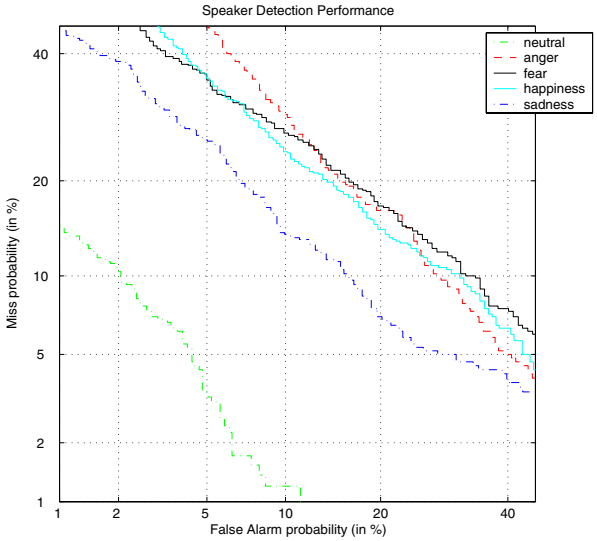

Figure 3.1(b): DET curve when models are trained with neutral speech and tested with anger, fear, happiness, sadness and neutral speech, respectively

From the table it can be seen that the verification system tends to achieve a better performance when training and testing speech is in the same emotion. This phenomenon proves that mismatched emotions between training and testing speech are one of the important reasons for the aggravation of verification performance. Proofs can also be found in Table 3.1 (taken from data in [9]), which compared the vocal feature characteristics of speech in each type of emotions. It can be seen that speech in different emotions shows different characteristics in various vocal features.

Table 3.1: Comparison of emotion and speech parameters (taken from data in [9])

	Anger	**Fear**	**Happiness**	**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 change	Abrupt on stressed	Normal	Smooth, upward inflections	Downward inflections
Articulation	Tense	Precise	Normal	Slurring

Table 3.2: Equal error rate of speaker verification system with training and testing speech in varied emotions (N: neural; A: anger; F: fear; H: happiness; S: sadness)

EER (%)		**Emotion type of test utterances**				
		N	**A**	**F**	**H**	**S**
Emotion type of speaker models	**N**	**4.48**	17.93	18.62	17.24	12.59
	A	17.76	**17.24**	20.17	17.24	21.38
	F	20.52	18.28	**17.59**	15.00	22.93
	H	14.48	16.55	12.07	**11.21**	19.83
	S	4.48	19.83	16.55	19.66	**6.90**

A phenomenon can be seen from Table 3.2 that even when training and testing speech are of the same emotion, the system could still perform differently on speech in different emotions. The verification results for speech in neutral or sadness greatly outperforms that in anger, fear or happiness. This might be attributed to different levels of intra-speaker vocal variability when speakers are in different emotions. Similarly in Table 3.1, it can be seen that when speakers are in the emotion of anger, fear or happiness, the pitch has a much wider range than that in sadness or neutral, which indicates that when speakers are in these three types of emotions, their articulating styles tend to create much greater intra-speaker vocal variability than they are in the emotion of sadness or neutral. So the articulating style of a certain type of emotions, which creates greater intra-speaker vocal variability, is another reason of the performance decline of speaker verification on emotional speech.

The above two sets of experiments prove that the performance of GMM-UBM based speaker verification systems is greatly affected by emotional speech. Two reasons for this are the mismatched emotions between training and testing speech, and the articulating styles of certain emotions which tend to create intense intra-speaker vocal variability.

4. EMOTION-DEPENDENT SCORE NORMALIZATION

The first reason for the performance decline analyzed in Section 3 is similar to the situation of channel effect, which is also induced by mismatched training and testing conditions. Hence, it might be helpful to utilize some ideas for handling channel effect to alleviate the emotion-induced negative effects on speaker verification. In this section, an emotion-dependent score normalization, named as Enorm, is proposed. This algorithm comes from Hnorm which was originally designed to alleviate channel effect in cross-channel speaker verification.

When training and testing speech are in different emotions, the discrepancy between the speaker models and the test utterances will induce biases in verification scores. The Enorm is designed to estimate from development data these emotion-dependent biases, and then remove them from verification scores. According to this algorithm, for each type of emotions, a set of impostor utterances in this emotion is preserved as the development data. For a speaker model, the set of impostor utterances in a specific type of emotions is scored. These scores are assumed to have a Gaussian distribution, and the mean and standard deviation are computed. Therefore each speaker model has an additional set of parameters describing its response to impostor utterances in each type of emotions, $\{\mu(E), \sigma(E)\}$, where E is *anger*, *fear*, *happiness*, *sadness*, or *natural*. To avoid bimodal distributions, the impostor utterances should be of the same gender as the speaker models. In the verification stage, the score of test utterance X is normalized using the following equation,

$$S_{Enorm}(X) = \frac{S(X) - \mu(E(X))}{\sigma(E(X))} \quad (1)$$

where $S(X)$ is the original verification score, and $E(X)$ is the emotion label of test utterance X. This normalization aims at transferring verification scores of impostor utterances in various emotions into standard Gaussian distributions.

In the experiments that examine the effect of Enorm, all speakers and test utterances in the evaluation dataset were utilized; and the development dataset was used for computing Enorm parameters. Each speaker model was trained with neutral speech, and the test utterances were mixed with speech in the five types of emotions. Two verification systems were involved in the experiments. The baseline system was the traditional GMM-UBM speaker verification system described in Section 2.2. For the Enorm, since we have not developed an effective emotion classifier yet, the emotion label of each test utterance was assumed to be available in the verification stage, thus here performed was an oracle test of Enorm. Verification results are shown in Figure 4.1, it is shown that the Enorm outperforms the baseline system as a whole.

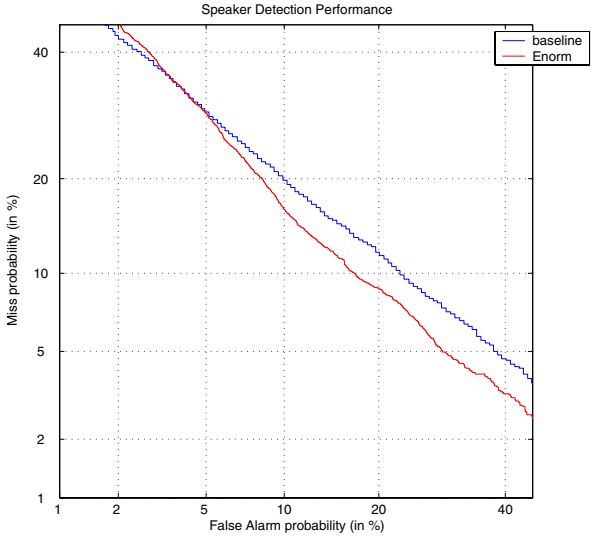

Figure 4.1: DET curves comparing Enorm and baseline system on mixed emotional speech

5. CONCLUSION

In this paper, the influence of emotion on speaker verification systems is studied. It is proved that emotion involved in the training or testing speech will aggravate the verification performance. Two reasons for this performance decline are mismatched emotions between the speaker models and the test utterances, and the articulating styles of certain emotions which create intense intra-speaker vocal variability. In response to the former reason for the performance decline, an emotion-dependent score normalization, the Enorm, is proposed, and an oracle test shows that Enorm can well alleviate the emotion effect on speaker verifications.

In future work, experiments combining emotion classifier and Enorm will be performed to examine the effect of Enorm in real applications. Since the performance decline of speaker verification on emotional speech is induced by the two reasons discussed above, future studies should be focused on developing emotion-robust algorithms in responding to these two reasons. Since the performance decline induced by mismatched emotions between speaker models and test utterances is similar to the situation of channel effect, which is also produced by mismatched training and testing conditions, some ideas of channel compensation algorithms, which are performed on the feature domain, model domain or score domain [10], can be borrowed to alleviate negative effects induced by emotion; For performance decline caused by certain emotions which produce intense intra-speaker vocal variability, algorithms which aim at reducing intra-speaker vocal variability should be studied. Research on the analysis of emotional speech [9][11] shown that pitch correlates closely with emotion. Hence it might be helpful to utilize pitch as a gauge to direct the normalization of feature vectors, and thus to alleviate or remove the intra-speaker vocal variability induced by emotions on feature vectors.

6. REFERENCES

[1] K. R. Scherer, "A cross-cultural investigation of emotion inferences from voice and speech: implication for speech technology," *in Proc. ICSLP 2000*

[2] T. L. Nwe, S. W. Foo, and L. C. De Silva, "Speech emotion recognition using hidden Markov models," *Speech Communication*, vol. 41, pp.603-623, 2003

[3] B. Schuller, G. Rigoll, and M. Lang, "Hidden Markov model-based speech emotion recognition," *in Proc. ICASSP 2003*

[4] A. Nogueiras, A. Moreno, A. Bonafonte, and J. B. Mari ň o, "Speech emotion recognition using hidden Markov models," *in Proc. Eurospeech 2001*

[5] D. A. Reynolds, *et al*, "The superSID project: exploiting high-level information for high-accuracy speaker recognition," *in Proc. ICASSP 2003*

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

[7] K. R. Scherer, T. Johnstone, G. Klasmeyer, and T. Bänziger, "Can automatic speaker verification be improved by training the algorithms on emotional speech?" *in Proc. ICSLP 2000*

[8] D. A. Reynolds, "Comparison of background normalization methods for text-independent speaker verification," *in Proc. Eurospeech 1997*

[9] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor, "Emotion recognition in human-computer interaction," *IEEE Signal Processing Magazine*, vol.18, no.1, pp.32-80, 2001

[10] D. A. Reynolds, "Channel robust speaker verification via feature mapping," *in Proc. ICASSP 2003*

[11] R. Cowei and E. Douglas-Cowie, "Automatic statistical analysis of the signal and prosodic signs of emotion in speech," *in Proc. ICSLP 1996*