2018 ICSEE International Conference on the Science of Electrical Engineering

Effect of Reverberation in Speech-based Emotion Recognition

Shujie Zhao, Yan Yang*, and Jingdong Chen
School of Marine Science and Technology
Northwestern Polytechnical University
Xi'an, Shaanxi, China

***Abstract*—In room environment, echo, reverberation, interference and additive noise cast the major challenges for emotional speech recognition due to degradation in quality and reliability of recorded speech signals. In this paper, we investigate effects of reverberation and noise on speech-based emotion recognition by comparing clean speech signal, adding simulated reverberant data, de-reverberant data and signal with added noise. First, we develop an emotional speech corpus of these four kinds of emotional speech data sources. Then we apply GMM-UBM framework to evaluate the performance of emotion recognition based on them. Results show that reverberation reduces emotion recognition accuracy by 5.87%, and a process of de-reverberation can largely cover this reduction.**

I. Introduction

The emotional aspect of speech is an important factor in human communication. Human-machine interaction (HMI) involves not only the interchange of explicit information, such as the linguistic information transmitted by speech, but also the implicit one, such as the information about the affective states of the interlocutor. Therefore, the Human-Robot Interaction (HRI) systems with the ability of emotional speech expression, can largely improve natural interaction performance. Emotional speech recognition focuses on automatically identify the effective state of a person from speech signals, and has many available applications in HMI. Such as detecting potential problematic points causing customers expressing anger and frustration in call center [1], lie detection [2], intelligent virtual agents and spoken dialogue computer tutors [3], [4].

In the real-life environment, reverberation and noise become challenges for emotional speech recognition. The performance of existing emotion recognition systems are mostly conducted in laboratory environment assuming as a noise-free environments, degrade rapidly in the natural environment. However, in natural environments, source data for emotion recognition are usually collected by one single microphone installed on a robot. In addition to the desired speech, a number of sound sources also exist. Taking signal acquisition in room environment for example, there always exist echo, reverberant, interference and additive noise, which can lead to degradation on the quality and reliability of speech related recognition tasks [5], [6], [7], [8], [9], [10]. Based on two public databases, Schuller et al. first studied the influence of noise conditions for speech emotion recognition [11]. For Danish Emotional Speech Corpus (DES), SVM classifier obtained a maximum of 74.5% for clean speech, and 54.9% at -10dB SNR level. For Berlin Emotional Speech Database (EMO-DB), remarkable 87.5% can be achieved for clean speech, 71.11% at -10dB SNR level as a maximum. Some speech signal processing methods are also employed in the speech emotion recognition systems, and the results of the performance with and without speech pre-processing techniques show that noise processing can greatly improve the emotion recognition performance. These methods are generally referred to either speech enhancement or noise reduction, such as the wavelet-based noise reduction method [12], wavelet-based adaptive thresholding technique [13], spectral subtraction and masking properties [14], wiener filter and minimum mean square error approaches [15], audio signal de-noising methods in cepstral and log-spectral domains [16].

Besides, only a few studies began to address the influence of reverberated noise on speech emotion recognition and the reverberant data are artificially created by convoluting the clean data with impulse responses recorded in real environment of different reverberation times [17]. Zhang et al. conducted the feature enhancement methods to address the environmental non-stationary additive and convolutional noise issue in speech emotion recognition. Based on a memory-enhanced recurrent Denoising Autoencoder (rDA) with Long Short-Term Memory neural networks, the proposed feature enhancement approach outperforms the baseline with non-enhancement methods, and can remarkably improve the performance of spontaneous emotion recognition from speech signal with added additive and convolutional noise [18]. Moreover, based on the human peripheral hearing system, some novel features are extracted in order to achieve a robust performance in noise and reverberation scenarios, such as the supervised Nonnegative Matrix Factorization (NMF) based features [19], damped oscillator cepstral coefficients (DOCCs) [20], Teager Energy Cepstrum Coefficient (TECC) based features [21], Cochlear filterbank Coefficients with zero crossing based features [22], pooling scheme based modulation spectral features [23].

It still kept unclear whether reverberation and noises from the room environment have same level of effects on speech emotion recognition, and what are the most effective approaches in speech signal enhancement for the purpose of emotional recognition. In this paper, we first investigate the influence of reverberation and noises from the room environment on emo-

978-1-5386-6378-3/18/$31.00

tion recognition. In the later speech enhancement, we apply de-reverberation technology based on microphone array, which fully use the time, frequency and space information carried by the multiple microphone signals to design and optimize the filter and beamforming algorithm. We choose multi-channel MINT method proposed by Miyoshi and Kaneda, which shown to obtain a nearly perfect speech de-reverberation performance in the absence of noise [26].

This paper is arranged as follows: Section II gives a detailed description of the experiments conducted to collect emotional speech data, including the clean data, noisy data, reverberant data and de-reverberant data. Section III contains the methodology including feature extraction from the emotional speech data and emotion recognition based on the GMM-UBM model. The experimental results demonstrating the effects of reverberation in the natural environment on emotion recognition are analyzed in Section IV. Conclusions and future work follow in Section V.

II. Data Collection

A. Emotional speech data

At first, we developed a multi-modal affective corpus in laboratory. The corpus includes four kinds of emotional data: speech, video, electroencephalogram and electrocardiogram collected at a sampling frequency of 44.1kHz in an anechoic chamber environment and a natural room environment, respectively. 30 graduates (15 male, average ages of 24) were recruited from Northwestern Polytechnical University and participated the data collection. Eight Video clips were used as stimuli for four types of emotions: Neutral, Sad, Angry and Happy. Video selection was based on the Internet and the general evaluation, with a certain degree of representation. In addition, considering the degree of difficulty induced by different emotions, each segment of the film was controlled at $2-5$ minutes during the experiment.

Only the emotional speech data were selected as the data source for our research. Considering that speech based emotion recognition is based on logically complete and independent statements, we preprocessed the speech passages into independent sentences. Every sentence was labeled with certain corresponding discrete emotion. A total of $13,249$ sentences were intercepted. Manual judgments were used to manually evaluate all sentences [24]. Finally, the label of all sentences were renewed according to the judgment results of two judges. For all $13,249$ sentences, there are 468 sentences that have been discarded, with a proportion of 3.53%; The re-categorized sentences are 563 sentences, accounting for 4.25% of the total; There are 92.22% sentence judgments which are consistent with the original mood. Therefore, the final speech database has $12,781$ sentences. Among them, 5682 sentences of pure emotional speech data collected in the anechoic chamber and 5557 sentences of noisy emotional speech data collected in the natural room environment were selected as the baseline data sources in this study.

B. Reverberant data

In order to simulate the reverberant environment, the impulse response of the room that reaches the microphone sensor from the sound source is generated by the mirror model method [25]. The mirror model method proposed by Allen and Bkrkley is a time domain model. It is suitable for the simulation of acoustic channel impulse response in rectangular structure space with advantages of simple ideas and simple operation. In the reverberant emotional speech data acquisition experiments, it is assumed that a mirror model is established for a rectangular room with plane reflection boundaries (four walls, ceiling, and ground). Each boundary of the room has the same reflection coefficient, which is irrelevant to the frequency of the signal and the angle of incidence of the signal on the boundary. Experimental parameters are set as follows: the size specification of the room is $5.0m \times 4.0m \times 2.9m$; the reflection coefficients of the walls range from 0 to 1; the position coordinates of an omnidirectional point sound source are $(1.0, 1.0, 0.8)$, and the unit is m; and the source signals are selected from the baseline data source; Also it has an omnidirectional microphone with the position coordinates of $(4.0, 3.0, 1.5)$ in m. At last, the gathered reverberant emotional speech data are in total of 5682 sentences.

C. Dereverberant data

Channel equalization method aims to design an inverse filter based on a channel impulse response to implement speech dereverberation. It is a deconvolution method that requires some prior information about the channel impulse response. According to the number of microphones used in the method, it is divided into single-channel method and multi-channel method. Here, we choose the multi-channel system channel equalization method-MINT for speech dereverberation. The MINT method is a multi-channel inversion method proposed by Miyoshi and Kaneda. By integrating the information from multiple channels, the minimum phase constraint of a single-channel system can be overcome and an accurate inverse filter can be obtained.

In our experiments, the uniform linear microphone array composed of five omnidirectional microphones is placed in a simulated reverberant environment. The experimental parameters in the reverberation room are set the same as that in the environment for the collection of the reverberant emotional speech data. The center of the linear array is at $(4.0, 3.0, 1.5)$. The spacing between two adjacent array elements in the array is $8.5cm$. The dereverberant data collection process is described as follows. First, the pure emotional speech signal is played by a speaker and received by five microphones via the room reverberant environment. Then the received multichannel speech signals are dereverberated through the MINT algorithm. Last, all the speech signals from the output of the MINT algorithm constitute the dereverberant emotion speech data source. The results of the MINT algorithm are shown in Fig.1.

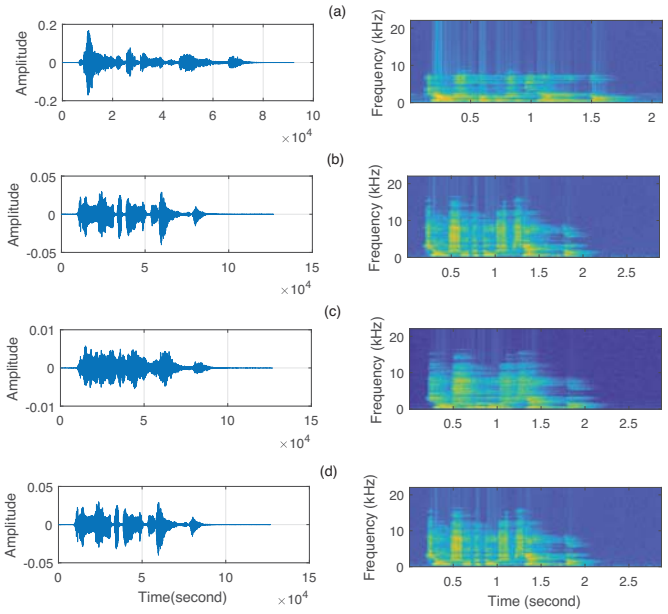

Fig. 1. Time series of the emotional speech signal (left) and the corresponding time spectrum (right). (a) The original noisy speech signal collected in natural environment; (b) The original pure speech signal gathered in anechoic chamber environment; (c) Signals received by microphone in simulated reverberant environment; (d) Output speech signal processed by MINT method.

TABLE I
EXTRACTED FEATURES FROM SPEECH SIGNALS

Types	Features	Dimensionality
Prosody	Energy	1
	Average amplitude	1
	Average zero crossing rate	1
	Maximum amplitude	1
	autocorrelation coefficient	1
	Energy and entropy ratio	1
Voice Quality	Pitch	2
	Reflection coefficients	13
	Format	8
Spectrum	Spectral centroid	1
	Spectral entropy	1
	LPCs	13
	LPCCs	12
	MFCCs	37
Total		93

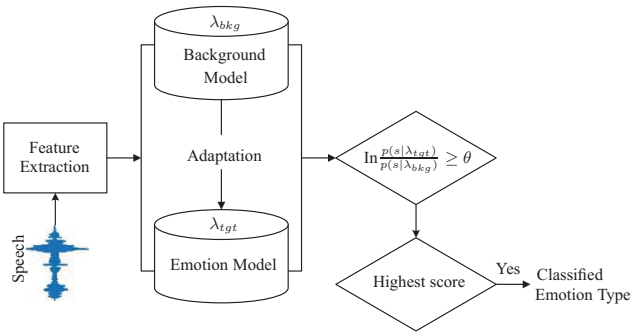

Fig. 2. GMM-UBM based emotion recognition system.

III. METHODOLOGY

This section is split into two subsections. To begin with, the feature extraction procedure from emotional speech data is introduced, which serves as a stepping-stone for emotion recognition. The other subsection will show the GMM-UBM based emotion recognition model in detail.

A. Feature Extraction

We first downsampled all of the speech signals into 16kHz, then spilt emotional speech signal into frames and extracted frame-level features. Each frame contains 1536 sample points, with duration of about 34 ms. The frame shift is 256 sample points. Three types of ninety-three basic features including prosody features, voice quality features, and spectrum features are extracted for further analysis. The extracted speech features are shown in TABLE I.

B. Emotion Recognition Model

A complete emotion recognition model mainly includes two components: a front-end and a back-end. The front-end is a feature extraction process converting the original acoustic waveform signal into a novel representation which is more compact and less redundant. The back-end is the core of the identification system where the specific model is designed for different types of emotions and verification trials are scored. Based on a statistical background model, the modeling phase aims to estimate the detailed acoustic space model for each emotion, resulting in a special model for determining the type of emotion.

We use the GMM-UBM framework provided by the MATLAB toolbox MSR Identity Toolbox as the emotion recognition model. GMM-UBM is also called a Gaussian mixture model-global background model. Initially, it is applied in speaker recognition systems, in order to obtain a speaker-independent feature distribution. In the designed emotion recognition system, the GMM-UBM model system is a background model that has nothing to do with emotion types. It uses all the training data from various emotions to be identified to obtain an emotional global background model, and then uses the maximum posterior probability estimation to regularize the parameters of the trained model adaptively. Finally, based on the established emotional model, the scores are computed as the log-likelihood ratio between the given emotion models and the UBM given the test data. It is assumed that the emotion type of the specific test trial corresponds to the emotion model with the highest scores. The GMM-UBM based emotion recognition system is shown in Fig.2.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

A. Features statistical analysis

Analysis of the significance of interaction between environment and emotion categories. In order to analyze whether there are interactions between the two experimental variables in this experiment (experimental environment and emotion category), we used the statistical analysis software SPSS to perform a multivariate analysis based on variance (confidence

is 0.95). None of the 93-dimensional features have significant differences (sig value greater than 0.05). Therefore, the interaction between the environment and the emotion category is not significant. In this case, we separately analyzed the impact of environmental differences and emotional categories on emotional recognition, without considering the cross-effects of the environment and emotional categories.

Variance analysis of the significance of speech signal features in different environments. In order to verify the impact of reverberation and other noises on emotion recognition performance, we performed a variance analysis with confidence of 0.95 for the 93 dimensional speech signal features extracted from the three data sources (pure data, reverberant data and noisy data), separately. Moreover, we analyzed whether each feature had some significant differences under different conditions. The specific results illustrate that there are 46-dimension features showing significant difference to the environment, as shown in TABLE II. From TABLE II, it can be seen that the prosody features screened out after significant analysis have the highest relative ratio to the original features, 100%; and the spectral features have the lowest relative proportion of 35.93%. It shows that the prosody features are significantly influenced by the environments, but the influence for spectral features is small.

TABLE II
FEATURES HAVING SIGNIFICANT DIFFERENCES IN DIFFERENT ENVIRONMENTS

Types	Dimensio--nality reserved	Features of significant differences	Dimens--ionality	Ratio
Prosody	6	Energy	1	100%
		Average amplitude	1	
		Zero crossing rate	1	
		Maximum amplitude	1	
		Energy and entropy ratio	1	
		autocorrelation coefficient	1	
Voice Quality	17	Pitch	2	73.91%
		Reflection coefficients	7	
		Format	8	
Spectrum	23	Spectral centroid	1	35.93%
		Spectral entropy	1	
		LPCs	9	
		LPCCs	7	
		MFCCs	5	
Total	46		46	49.46%

TABLE III
EMOTION RECOGNITION ON PURE DATA

True label \ Predict label	Neutral	Sad	Angry	Happy
Neutral	79	36	21	64
Sad	26	115	19	40
Angry	17	36	95	52
Happy	29	26	30	115
Accuracy(%)	39.5	57.5	47.5	57.5
Average accuracy(%)	50.50			

TABLE IV
EMOTION RECOGNITION ON NOISY DATA

True label \ Predict label	Neutral	Sad	Angry	Happy
Neutral	46	63	39	52
Sad	67	39	42	52
Angry	28	28	97	47
Happy	29	35	36	100
Accuracy(%)	23	19.5	48.5	50
Average accuracy(%)	35.25			

TABLE V
EMOTION RECOGNITION ON REVERBERANT DATA

True label \ Predict label	Neutral	Sad	Angry	Happy
Neutral	69	43	40	48
Sad	30	114	23	33
Angry	33	42	94	31
Happy	25	55	40	80
Accuracy(%)	34.5	57	47	40
Average accuracy(%)	44.63			

TABLE VI
EMOTION RECOGNITION ON DEREVERBERANT DATA

True label \ Predict label	Neutral	Sad	Angry	Happy
Neutral	74	35	33	58
Sad	30	114	23	33
Angry	27	29	95	49
Happy	28	45	38	89
Accuracy(%)	37	57	47.5	44.5
Average accuracy(%)	46.50			

B. Emotion recognition

In the experiment, 1000 sentences for each of the four emotions (neutral, sad, angry, and happy) were selected randomly from the corpus as training samples and 200 sentences as test samples. A total of 93 dimensional feature vectors were extracted from the selected samples. Based on the feature statistical analysis results, the GMM-UBM algorithm is used to classify four kinds of emotions and the statistical analysis of the classification results is performed according to the recognition results, as shown in TABLE III, TABLE IV, TABLE V, TABLE VI.

The results show that the emotion recognition model achieves the highest accuracy of 50.50% for four-class classification when applied on clean speech signal, and noisy speech showed the lowest accuracy of 35.25% (decrease by 15.25%). Reverberation caused a decrease in model performance to 44.63% (decrease by 5.87%). The MINT de-reverberant process helps to recover the reduction significantly and achieved an accuracy of 46.50%. In general, TABLE IV, TABLE V and TABLE VI show that the average accuracy of de-noising speech increased by 9.38% compared to noisy speech. While the average accuracy of the de-reverberant speech has a rise of 1.87% compared to reverberant speech. These results indicate that the effects of other noises are more significant than that of reverberation in noisy speech on the performance of

emotional speech recognition. Reverberation is generated by reflections of obstacles, such as walls, ceilings, and floors when acoustic waves propagate indoors. According to time series, reverberation is divided into early reverberation and late reverberation. Studies have shown that early reverberation has a positive influence on speech signals, while other noises are only considered an interference with speech signals. Thus reverberation may have less negative impact on the performance of speech emotion recognition.

V. Conclusion

In this paper, we first establish an emotional speech database containing four types of speech data sources (clean speech signals, reverberant speech, noisy speech, and de-reverberant speech) that correspond to four types of room environments. Then based on the GMM-UBM framework, the performance of emotion recognition under these four conditions were analyzed and compared. The results show that, the noises in the room environment have high impact on emotion recognition performance, compared to reverberation. Therefore, we can conclude that it is very important to perform noise reduction on speech signals for emotion recognition.

In future work, the crucial task is to extend and evaluate more noise reduction methods based on multi-channel microphone array on emotion recognition from noisy speech. Future research will also focus on establishing a joint optimization problem of speech enhancement and emotional classification and finding an efficient solution.

Acknowledgment

The authors would like to thank CIAIC group, Northwestern Polytechnical University, China for the support of the research. This project was supported in part by the Natural Science Foundation of China(NSFC) under grant no. 6177012290 (F011305).

References

[1] D. Morrison, R. Wang, and L. C. De Silva, *Ensemble methods for spoken emotion recognition in call-centres*, Speech Communication, vol. 49, no. 2, pp. 98-112, 2007.

[2] D. Tomotsune, M. Shirai, Y. Takihara, and K. Shimada, *Detecting deception: the promise and the reality of voice stress analysis*, Journal of Forensic Sciences, vol. 27, no. 2, pp. 340-51, 1982.

[3] C. Becker-Asano, T. Kanda, C. Ishi, and H. Ishiguro, *How about laughter? Perceived naturalness of two laughing humanoid robots*, pp. 1-6, 2009.

[4] K. Forbes-Riley, and D. Litman, *Benefits and challenges of real-time uncertainty detection and adaptation in a spoken dialogue computer tutor*, Speech Communication, vol. 53, no. 9-10, pp. 1115-1136, 2011.

[5] H. W. Loellmann, H. Barfuss, A. Deleforge, and S. Meier, *Challenges in Acoustic Signal Enhancement for Human-Robot Communication*, Speech Communication, pp. 1-4, 2014.

[6] M. S. Maucec, Z. Kacic, and A. Zgank, *Speech recognition for interaction with a robot in noisy environment*, Przeglad Elektrotechniczny, vol. 89, no. 5, pp. 232-236, 2013.

[7] J. Hansen, and D. Cairns, *ICARUS: source generator based real-time recognition of speech in noisy stressful and Lombard effect environments*, Speech Communication, vol. 16, no. 4, pp. 391-422, 1995.

[8] J. Hansen, *Analysis and compensation of speech under stress and noise for environmental robustness in speech recognition*, Speech Communication, vol. 20, no. 1-2, pp. 151-173, 1996.

[9] K. Scherer, T. Johnstone, G. Klasmeyer, and T. Banziger, *Can autuomatic speaker verification be improved by training the algorithmes on emotional speech?*, pp. 807-810, 2000.

[10] S. Karimi, and M. H. Sedaaghi, *Best features for emotional speech classification in the presence of babble noise*, in Electrical Engineering, 2012.

[11] B. Schuller, D. Arsic, F. Wallhoff, and G. Rigoll, *Emotion Recognition in the Noise Applying Large Acoustic Feature Sets*, in Proc. Speech Prosody, 2006.

[12] L. He, L. He, *Stress and emotion recognition in natural speech in the work and family environments*, 2010.

[13] A. Tawari, and M. M. Trivedi, *Speech Emotion Analysis in Noisy Real-World Environment*, in International Conference on Pattern Recognition, 2010.

[14] C. Huang, G. Chen, H. Yu, Y. Bao, and L. Zhao, *Speech Emotion Recognition under White Noise*, Archives of Acoustics, vol. 38, no. 4, pp. 457-463, 2013.

[15] F. Chenchah, and Z. Lachiri, *Speech emotion recognition in noisy environment*, in International Conference on Advanced Technologies for Signal and Image Processing, 2016.

[16] J. Pohjalainen, F. Ringeval, Z. Zhang, B. Schuller, *Spectral and Cepstral Audio Noise Reduction Techniques in Speech Emotion Recognition*, Proceedings of the 2016 Acm Multimedia Conference, pp. 670-674, 2016.

[17] F. Eyben, F. Weninger, and B. Schuller, *Affect recognition in real-life acoustic conditions - A new perspective on feature selection*, 14th Annual Conference of the International Speech Communication Association, Interspeech, F. Bimbot, C. Cerisara, C. Fougeron, G. Gravier, L. Lamel, F. Pellegrino and P. Perrier, eds., pp. 2043-2047, Baixas: Isca-Int Speech Communication Assoc, 2013.

[18] Z. Zhang, F. Ringeval, J. Han, J. Deng, E. Marchi, B. Schuller, *Facing Realism in Spontaneous Emotion Recognition from Speech: Feature Enhancement by Autoencoder with LSTM Neural Networks*, 17th Annual Conference of the International Speech Communication Association, Interspeech, pp. 3593-3597, 2016.

[19] F. Weninger, B. Schuller, A. Batliner, S. Steidl, and D. Seppi, *Recognition of Nonprototypical Emotions in Reverberated and Noisy Speech by Nonnegative Matrix Factorization*, Eurasip Journal on Advances in Signal Processing, vol. 2011, no. 1, pp. 1-16, 2011.

[20] V. Mitra, A. Tsiartas, and E. Shriberg, *Noise and reverberation effects on depression detection from speech*, in IEEE International Conference on Acoustics, Speech and Signal Processing, 2016.

[21] R. Sun, and I. I E. Moore, *Investigating the robustness of teager energy cepstrum coefficients for emotion recognition in noisy conditions*, 2012.

[22] P. K. Aher, S. D. Daphal, A. N. Cheeran, *Analysis of Feature Extraction Techniques for Improved Emotion Recognition in Presence of Additive Noise*, 2016.

[23] A. R. Avila, J. Monteiro, O. Douglas, and T. H. Falk, *Speech emotion recognition on mobile devices based on modulation spectral feature pooling and deep neural networks*, in IEEE International Symposium on Signal Processing and Information Technology, 2017.

[24] T. Bänziger, M. Mortillaro, and K. R. Scherer, *Introducing the Geneva Multimodal expression corpus for experimental research on emotion perception*, Emotion, vol. 12, no. 5, pp. 1161-1179, 2012.

[25] J. B. Allen, and D. A. Berkley, *Image method for efficiently simulating small-room acoustics*, Journal of the Acoustical Society of America, vol. 65, no. S1, pp. 943-950, 2016.

[26] M. Miyoshi, and Y. Kaneda, *Inverse filtering of room acoustics*, Acoustics Speech & Signal Processing IEEE Transactions on, vol. 36, no. 2, pp. 145-152, 1988.