ICIC Express Letters
Part B: Applications
Volume **7**, Number **2**, February **2016**

ICIC International ©2016 ISSN 2185-2766
pp. **265–270**

DISTANT SPEECH EMOTION RECOGNITION BASED ON FEATURE ENHANCEMENT

Xiaoping Jiang[1,2], Dingzhou Wang[1,2], Chenhua Li[1,2] and Hao Ding[1,2]

[1]College of Electronics and Information Engineering
[2]Hubei Key Laboratory of Intelligent Wireless Communications
South-Central University for Nationalities
No. 182, Minyuan Road, Hongshan District, Wuhan 430074, P. R. China
arkage@qq.com

Received July 2015; accepted October 2015

Abstract. *In this paper we study the emotion recognition problem in distant speech. The practical situations require effective methods dealing with distant speech signal. First, the basic speech emotion recognition procedure is investigated. Artificial neural networks are adopted to construct an effective recognizer. Second, the emotional features are analyzed under reverberant environment. Enhancement algorithm of the speech feature is studied based on linear filtering. Third, a novel optimization method is used to optimize the dereverberation algorithm parameters. Finally, experiments are carried out to verify the effectiveness of our method. The result shows that the distant speech emotion recognition is successful and the enhancement algorithm constantly improves the emotion recognition accuracy.*
Keywords: Emotion recognition, Shuffled frog leaping algorithm, Feature enhancement

1. **Introduction.** Speech emotion recognition is an important research topic in intelligent human-computer interaction (HCI) [1, 2, 3]. Steidl [1] studied the natural communication between an intelligent robot and children. Wang et al. [3] applied emotion recognition system in manned space mission. Recent developments in signal processing and machine learning have innovated some interesting HCI applications based on speech emotion recognizer [4, 5, 6]. Huang et al. [4] studied the emotions related to cognitive process. In their work, both off-line training and on-line training methods are considered [6]. However, in some real world applications, the speaker is away from the microphone. When the user is talking to an intelligent robot in an indoor environment or an outdoor environment, it is very likely that the speech signal is processed in a distance. The distant speech signal causes an obvious performance drop in both automatic speech recognition and automatic speech emotion recognition.

In this paper we focus on the indoor environment and study the enhancement technique to improve the speech emotion recognition performance. In distant speech processing in a room, the main reason of the signal quality decrease is the room reverberation. Several previous works have addressed this problem [7, 8, 9]. Nakatani et al. [7] proposed to use variance normalization to improve the speech dereverberation. Kinoshita et al. [8] presented a database for reverberant speech. Schwarz et al. [9] applied independent component analysis to speech dereverberation with post filtering. However, none of them considered the application of speech emotion recognition.

In order to better enhance the speech signal, the dereverberation algorithm parameters need to be set accurately. Various optimization algorithms can be used for parameter optimization. Among them shuffled frog leaping algorithm (SFLA) is a powerful meta-heuristic optimization algorithm. Yu et al. [10], applied SFLA to the parameter optimization in neural network and achieved good results in speech emotion recognition.

Zhang et al. [11], proposed an improved version of SFLA and solved constraint problems more efficiently. In this paper, we show that SFLA can be used for successful parameter estimation in the feature enhancement algorithm.

The rest of the paper is organized as follows: Section 2 gives the general description of our baseline speech emotion recognition system; Section 3 discussed the influence of distant speech on the emotional features and the enhancement method; Section 4 gives the parameter optimization based on SFLA; Section 5 provides the detailed experimental results, and finally, conclusions are given in Section 6.

2. **The Baseline Speech Emotion Recognizer.** Various acoustic features can be used in the speech emotion recognizer. In this section, we will introduce the commonly used speech emotion features and the neural network classifier trained on these features.

Human emotions can be modelled into a continuous dimensional model, namely the arousal-valence model. The arousal dimension is generally considered to have a close relationship with the prosodic features in speech. The valence dimension, on the other hand, is closely related to the voice quality features. In our baseline speech emotion recognition system, we include both prosodic features and voice quality features. Intensity, as one of the most commonly used prosodic features, is adopted as an emotion feature. Pitch frequency extracted from speech signal and its global statistics, including maximum, minimum, standard deviation and range, are constructed for emotion recognition. The first three formant frequencies are also taken as the voice quality features. The global statistics are also constructed. In total, we have twenty dimensions of emotional features for the modelling.

A three-layer artificial neural network is adopted, as shown in Figure 1. The input layer contains twenty neurons, and each corresponds to one feature dimension. The hidden layer consists of eight neurons. The output layer consists of six neurons, and each represents the possibility of one specific emotion type. The neural network is trained on emotional speech utterances with back-propagation (BP) algorithm.

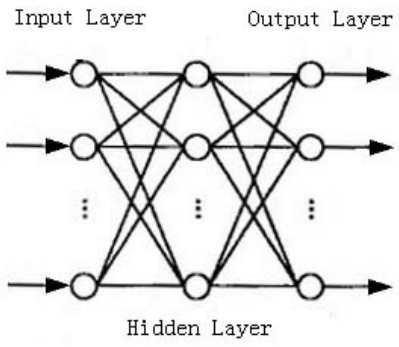
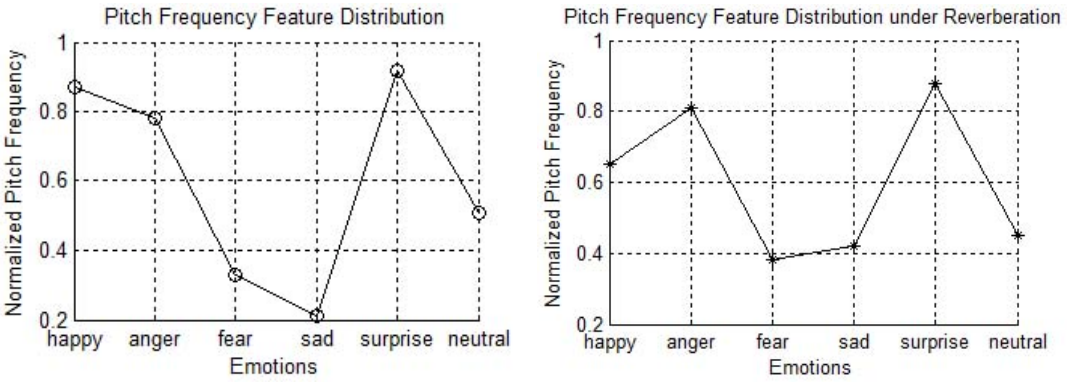


FIGURE 1. A depiction of the three-layer artificial neural network

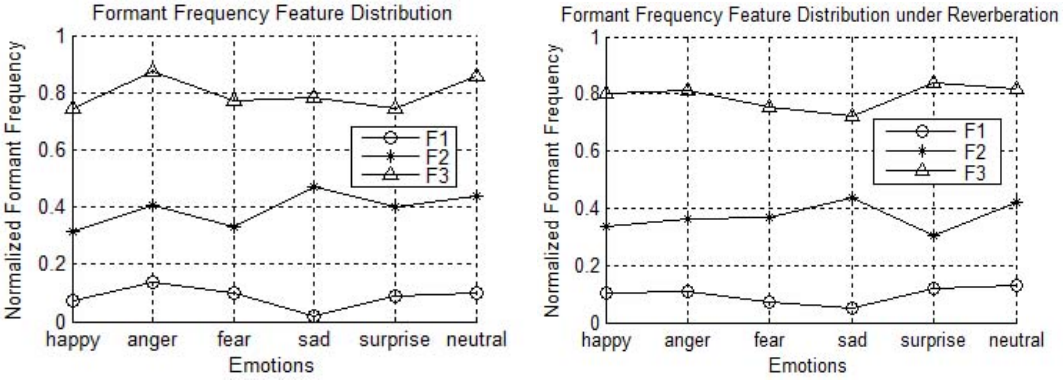
3. **The Emotional Feature Enhancement.** In distant speech emotion recognition, we focus on the indoor scenario. The reverberation is one of the most important factors in this scenario. The room can be modelled as an acoustic channel, which is determined by the location of the speaker, the location of the microphone, and the condition of the room. Most of room impulse responses in real world flow the exponential fading rule. The strong echo is not commonly seen, and the interferences caused by multiple acoustic paths in a room can be compensated by direct spectral subtraction in frequency-time domain.

The reverberation, therefore, is categorized into the early reverberation and the late reverberation. The early reverberation is quite common in natural speech and it may



(a) (b)

FIGURE 2. Pitch feature distribution in reverberated speech signal: (a) pitch frequency in clean speech, (b) pitch frequency in reverberated speech



(a) (b)

FIGURE 3. Formant feature distribution in reverberated speech signal: (a) formant frequency in clean speech, (b) formant frequency in reverberated speech

improve the perception of speech content in some cases. The late reverberation causes a significant drop in speech emotion recognition accuracy and it is the major concern in our feature enhancement algorithm.

The effects on the features caused by reverberation are shown in Figure 2 and Figure 3. We extract the first three formant frequencies which are denoted as F1, F2 and F3. We can see that the two important emotional features are both influenced greatly by the room acoustic channels. The farther the speaker departs from the microphone, the worse the received speech feature becomes.

There are two ways to compensate this effect on the features. One is based on the well-known spectral subtraction method. The late reverberation can be removed in spectrum based on the observation of past speech signals. The other method is based on linear filtering, and it is more efficient because it takes the phase information into account.

In this paper we adopt the delayed linear filtering proposed in [7], as shown in Equation (1).

$$s(n) = \sum_{t=L}^{D} g(t)x(n-t) \tag{1}$$

where $s(n)$ is the output of linear filtering, n is the discrete time sequence index, t is the index of filter taps, x is the observed signal sequence, and L and D define the time window

relating to the late reverberation. In this linear filtering model, the past speech between index L and D is delayed and used for the compensation of late reverberation. The late reverberation is usually considered after 50ms, and D can be set around $0.05 \times f_s$. f_s denotes the sampling rate.

In order to estimate the filter coefficients, we can use a maximum likelihood (ML) method [7]. The likelihood is shown in Equation (2).

$$L(\theta) = \sum_{n=1}^{N} \log P\left(\mathbf{x(n)}|\mathbf{x(n-D)};\theta\right) \tag{2}$$

where L is the likelihood, N is the length of the observation sequence, θ is the filter parameter, and P is the probability.

The observation signal vector $\mathbf{x}$ is determined by Equation (3)

$$\mathbf{x(n)} = [\mathbf{x(n)}, \mathbf{x(n-1)}, \ldots, \mathbf{x(n-L)}]^{\mathbf{T}} \tag{3}$$

Considering the linear filtering process in Equation (1), the likelihood can be rewritten as:

$$L(\theta) = \sum_{n=1}^{N} \log P\left(\mathbf{x(n)} - \mathbf{g_L x(n-D)};\theta\right) \tag{4}$$

The parameter of the filtering consists of the delay factor D and filter taps L. The proper setting of the filter parameter is key to the feature enhancement against reverberation. In the next section we will apply SFLA to solving this optimization problem.

4. **Parameter Optimization Based on SFLA.** In this paper, we adopt the powerful meta-heuristic optimization algorithm SFLA for the parameter searching. We initialize the frog individuals using the linear filtering parameter D and L, and then search for the global optimal for these parameters. The fitness function f is defined based on the emotion recognition false rates e.

$$f = -20 * \log\left(\frac{1}{2M}\sum_{i=0}^{M}\sum_{j=0}^{1} e_{i,j}\right) \tag{5}$$

where i is the index of emotion type, j is the index of false type, and M is the number of the emotion types.

In reference [12], an improved version of the SFLA is introduced and we apply this variant to our parameter optimization problem. The location of each frog is determined by time delay parameter D and filter tap length parameter L in the linear filtering algorithm. The fitness function in Equation (5) is calculated for finding the global optimal individual and the worst individual. The location of the worst individual is then updated in each iteration of SFLA. The updating strategy is based on the Velocity-Verlet algorithm which is a powerful solution to the motion equation in the molecular dynamics.

The advantage of this optimization algorithm is that the parameter of the linear filtering can be optimized efficiently. The algorithm can jump out of the local extreme and avoid ill-posed delay parameter or filter tap parameter. The constraint on the delay parameter is also considered in the optimization.

5. **Experimental Result.** In this experiment we combined the enhancement method with the baseline speech emotion recognizer.

The reverberant speech and the enhanced speech are shown in Figure 4. We can see that the linear filtering based on the SFLA optimization is effective. The filter tap length is optimized by SFLA, and the global optimal value is 25. Longer filter tap length will cause a wast of computational resource and it may also prevent the filter algorithm from

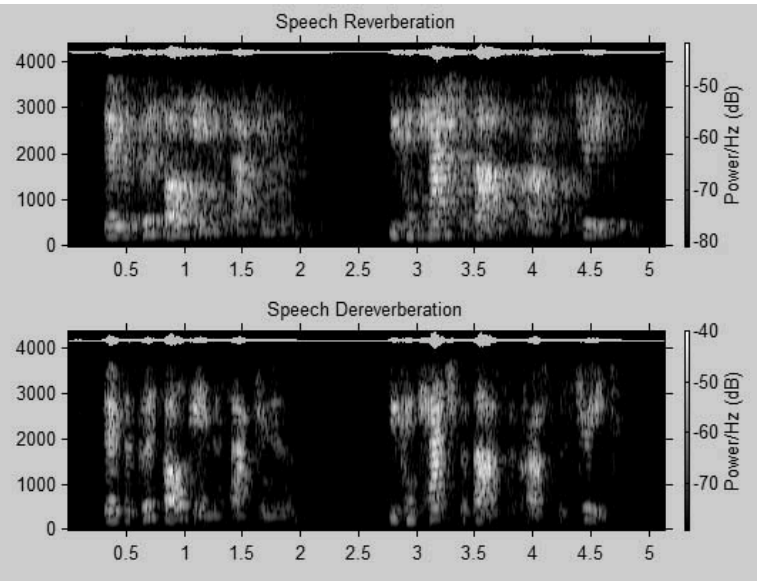


FIGURE 4. Speech enhancement under reverberation

TABLE 1. Improved signal quality by enhancement

Measurements	IS	SNR	WSS	LAR	LLR
Distant Speech	8.1	8.3	98	6.1	1.8
Enhanced Feature	5.4	15.1	74	4.5	1.3

TABLE 2. Speech emotion recognition rates before and after feature enhancement

Various Emotion Types	Happiness	Anger	Fear	Sadness	Surprise	Neutrality
False Acceptance in Distant Speech	15.1%	8.4%	20.1%	9.6%	10.7%	19.1%
False Rejection in Distant Speech	21.2%	10.1%	22.3%	8.7%	11.7%	10.5%
False Acceptance after Enhancement	13.4%	6.5%	13.2%	7.2%	8.2%	11.4%
False Rejection after Enhancement	18.2%	7.2%	18.3%	5.6%	6.5%	9.9%
Improvement (False Acceptance)	11.3%	22.6%	34.3%	25.0%	23.4%	40.3%
Improvement (False Rejection)	14.2%	28.7%	17.9%	35.6%	44.4%	5.7%

convergence. Shorter filter tap length will decrease the signal quality in the dereverberation. The time delay parameter is also optimized by SFLA, and the optimal value is 57. This parameter is key to the cancellation of reverberation.

Five measurements are adopted to show the result of the enhancement, as shown in Table 1. IS stands for Itakura-Saito; SNR stands for Signal-to-Noise Ratio; WSS stands for Weighted Spectral Slope; LAR stands for Log-Area Ratio; LLR stands for Log Likelihood Ratio. These objective measurements are used to verify the results of distant speech feature enhancement. As shown in Table 1, the speech qualities are significantly improved. The emotional features based on the enhanced speech signal are also enhanced. The lower IS, WSS, LAR, and LLR become, the better the signal quality is. The higher SNR gets, the better the signal quality is.

We further verified the enhancement framework in the emotion recognition test. The recognition rates are shown in Table 2. We can see that over the six emotions, namely happiness, anger, fear, sadness, surprise and neutrality, the false acceptance rates are decreased. So are the false rejection rates.

6. **Conclusions.** In this paper we analyze the application of SFLA algorithm in speech dereverberation framework. We apply the proposed algorithm in the parameter optimization problem in dereverberation algorithm. The fitness function is designed based on the

error rate of the speaker emotion identification system. The improvement over the baseline artificial neural network system shows that the optimization is successful. In future work, we will further explore the possibility of combining expectation-maximization (EM) algorithm to improve the efficiency of the speaker emotion identification system.

Acknowledgement. This work was supported by the General Program of the Natural Science Fund of Hubei Province (2014CFB916) and the Key Project of the Fundamental Research Funds for the Central Universities (CZZ13001 and CZW15043).

REFERENCES

[1] S. Steidl, *Automatic Classification of Emotion-Related User States in Spontaneous Children's Speech*, Ph.D. Thesis, FAU Erlangen-Nuremberg, Logos Verlag, Berlin, Germany, 2009.

[2] M. A. Nicolaou, H. Gunes and M. Pantic, Output-associative rvm regression for dimensional and continuous emotion prediction, *Image and Vision Computing*, vol.30, no.3, pp.186-196, 2012.

[3] J. Wang, B. Wu, C. Huang, H. Qin, C. Zha and L. Zhao, Segment-based static feature analysis and recognition of emotional speech for manned space mission, *ICIC Express Letters*, vol.8, no.6, pp.1541-1546, 2014.

[4] C. Huang, Y. Zhao, Y. Jin, Y. Yu and L. Zhao, A study on feature analysis and recognition for practical speech emotion, *Journal of Electronics & Information Technology*, vol.33, no.1, pp.112-116, 2011.

[5] C. Huang, D. Han, Y. Bao, H. Yu and L. Zhao, Cross-language speech emotion recognition in German and Chinese, *ICIC Express Letters*, vol.6, no.8, pp.2141-2146, 2012.

[6] C. Huang, R. Liang, Q. Wang, J. Xi, C. Zha and L. Zhao, Practical speech emotion recognition based on online learning: From acted data to elicited data, *Mathematical Problems in Engineering*, pp.1-8, 2013.

[7] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi and B. H. Juang, Speech dereverberation based on variance-normalized delayed linear prediction, *IEEE Trans. Audio, Speech and Language Processing*, vol.18, no.7, pp.1717-1732, 2010.

[8] K. Kinoshita, M. Delcroix, T. Yoshioka et al., The reverb challenge: A common evaluation framework for dereverberation and recognition of reverberant speech, *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, pp.13-17, 2013.

[9] A. Schwarz, K. Reindl and W. Kellermann, Two-channel reverberation suppression scheme based on blind signal separation and Wiener filtering, *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp.1-4, 2012.

[10] H. Yu, C. Huang and X. Zhang, Shuffled frog-leaping algorithm based neural network and its application in speech emotion recognition, *Journal of Nanjing University of Science and Technology*, vol.5, no.14, pp.13-17, 2011.

[11] X. Zhang, L. Zhao and C. Zou, An improved shuffled frog leaping algorithm for solving constrained optimization problems, *Journal of Shandong University*, vol.43, no.1, pp.1-8, 2013.

[12] X. Zhang, F. Hu, L. Zhao and C. Zou, Improved shuffled frog leaping algorithm based on molecular dynamics simulations, *Journal of Data Acquisition and Processing*, vol.27, no.3, pp.327-332, 2012.