ICIC Express Letters
Part B: Applications
Volume **8**, Number **1**, January **2017**

ICIC International ©2017 ISSN 2185-2766
pp. **201–206**

AUTOMATIC EMOTIONAL SPEECH RECOGNITION BASED ON MULTI-SCALE FEATURE FUSION AND DEEP NEURAL NETWORK

Haoyu Wang[1] and Chengxian Yu[2]

[1]The Institute of Education
Anqing Normal University
No. 128, Linghu South Rd., Anqing 246011, P. R. China

[2]School of Economics and Management
Anqing Normal University
No. 1318, Jixian North Road, Anqing 246133, P. R. China
ychx1974@126.com

Received February 2016; accepted May 2016

Abstract. *In this paper we study the automatic emotional speech recognition problem based on deep learning. First, a novel multi-scale feature fusion algorithm is proposed. Word level and sentence level features are constructed and optimized for acoustic recognition. Second, deep neural network is used to model the speech emotional features in various languages. Four types of languages are studied for emotion recognition. Third, we verify the proposed algorithm in experiments using the generalized model in German, Chinese, Vietnamese and English. Experimental results show that the proposed multi-scale feature constantly improves the recognition rate and deep neural network based model outperforms the traditional Gaussian model in cross-language emotion recognition.*
Keywords: Speech recognition, Feature optimization, Information fusion, Deep learning

1. **Introduction.** Speech interaction is one of the most convenient ways in man-machine communication [1, 2, 3]. The traditional speech recognition focuses on the processing of linguistic information. The affective information has been treated as noise. Speech emotion recognition (SER), on the other hand, focuses on the recognition of speaker emotional states using various prosodic and voice quality features [4, 5].

In the past researches, many speech emotion databases were restricted to one or two types of languages [6]. The Berlin Emotional Speech Database [7] was designed for German emotional speech synthesis. The AIBO database included both German and English [8]. Although promising results have been reported on these databases, a generalized emotion recognition model is still an unsolved challenge.

There are many acoustic parameters that can be applied to speech analysis. MFCC based features are proposed to apply with hidden Markov model, and several successful applications are reported on German database [9]. Pitch and other prosodic features are the most widely used emotional features, and they are successfully used in many types of languages. Further studies have shown that the cross-language emotion recognition problem is largely dependent on the generalization of emotional features [10]. Finding the optimized features for more than one type of language has become a key step in current speech emotion recognition research.

In this paper we propose a novel multi-scale feature fusion algorithm combined with deep neural network modelling. The proposed algorithm has a strong ability of representing various acoustic features in different languages. The rest of the paper is organized as follows: Section 2 gives an introduction of the database used; Section 3 describes the proposed feature fusion algorithm; Section 4 gives details on deep neural network modelling;

experimental results are given in Section 5; and finally, the conclusions are provided in Section 6.

2. **The Databases.** In order to study the generalization ability of speech emotion model, we select several available databases. They are German database (Berlin Emotional Speech Database, EMODB), Chinese database [11], Vietnamese database [12], and English database (eNTERFACE) [13]. These four databases have covered four different languages from Germanic Language, Sino-Tibetan language and South-Asian Language.

The German database includes 10 speakers, 5 males and 5 females. The speech material comprises about 800 sentences (seven emotions, ten actors, and ten sentences). The Chinese database comprises six emotions, including happiness, sadness, fear, anger, surprise and neutrality. Six professional actors and actresses participated in the recording. The emotional corpus includes 2,268 words, 2,916 short sentences, and 210 paragraphs. The Vietnamese database contains the same six emotions as the Chinese database. The speech utterances are induced by emotional scenes. The speech text consists of 30 sentences without specific emotional meanings. The eNTERFACE database contains both audio and video data. The emotional utterance is in English, and 42 people coming from 14 different nationalities participated in the data collection. The collected emotions contains happiness, sadness, surprise, disgust, anger, fear and neutrality. We select six common emotions from eNTERFACE database for experiment. Compared to the other databases eNTERFACE provides the largest amount of subjects.

3. **Multi-Scale Feature Analysis.** Speech emotion analysis is based on the assumption that emotions are expressed over a certain period of time in the signals. According to the related psychology study, the felt emotion can last about one or two minutes. In practice, the recognition of the perceived emotion is carried out in a much shorter duration. Past studies have shown that the frame-wise and turn-wise recognition methods are both successful. However, the optimal duration for recognition speech is still not clear.

In this paper, we proposed to analyze the emotional feature in both word level and sentence level, since these two components are the most common and natural segments in speech. The optimal duration may be different from emotion to emotion, as shown in Table 1.

TABLE 1. Emotion recognition rate under various duration (%)

Emotions	**1 word**	**2 words**	**3 words**	**5 words**	**8 words**
Happiness	61.3	66.9	70.8	73.1	72.1
Sadness	57.4	56.7	55.4	59.2	61.0
Fear	59.1	59.1	61.0	60.4	62.2
Surprise	66.7	70.0	65.9	71.5	70.1
Anger	71.3	72.9	77.9	78.2	77.3
Neutrality	70.1	70.1	66.7	68.1	67.3

A fusion algorithm is proposed to optimize the emotional features. Speech signal $s(t)$ is defined on the time domain, and the basic acoustic parameters $\phi_h(l)$ are constructed on the frame-wise sequence $s'(n)$. ϕ denotes the acoustic parameter, h denotes the index of different parameters, and l is the order of the parameter. For formant frequencies $l = 1, 2, 3, 4$, the first four formants are adopted. For mel frequency cepstrum coefficients, $l = 0, 1, 2, \ldots, 12$, 12-order coefficients are adopted. s' denotes one frame of speech signal, and n is the index of speech frame. The emotional features are then constructed from the basic acoustic parameters:

$$\mathbf{f}_d(h,l) = \left[\max_n(\phi_h(l,n)), \min_n(\phi_h(l,n)), \operatorname*{mean}_n(\phi_h(l,n)), \operatorname*{var}_n(\phi_h(l,n))\right] \tag{1}$$

where d is the duration of the speech signal, in frame-wise feature extraction this duration equals to the length of a frame, in turn-wise feature extraction this duration equals the length of a speech segment, and in multi-scale feature extraction this duration is not fixed. "max()" stands for the maximization function over all frames, "min()" stands for the minimization function, "mean()" stands for the average function, and "var()" stands for the variance of the basic acoustic features. The total dimension of features is determined by $h \times l$.

In traditional feature construction framework, the static features extracted from low level parameters are sent to the learning algorithm directly with proper feature reduction. In this paper we adopt probabilistic model to further extract the high level features that are less dependent on the duration of the feature analysis. The word level feature and the sentence level feature are modeled for the posterior probability, as shown in Figure 1.

$$\mathbf{x} = \mathbf{a} \cdot \mathbf{p}\left(\boldsymbol{f}^{w}|\lambda_j\right) + \mathbf{b} \cdot \mathbf{p}\left(\boldsymbol{f}^{g}|\lambda_j\right) \tag{2}$$

The model parameter λ_j differs from emotion to emotion, and j denotes the index of emotions. $\boldsymbol{f}^{w}$ stands for the feature vector of word level analysis, and $\boldsymbol{f}^{g}$ stands for the sentence (utterance) level feature. The fusion weights $\mathbf{a}$ and $\mathbf{b}$ are optimized using a standard three layer perceptron.

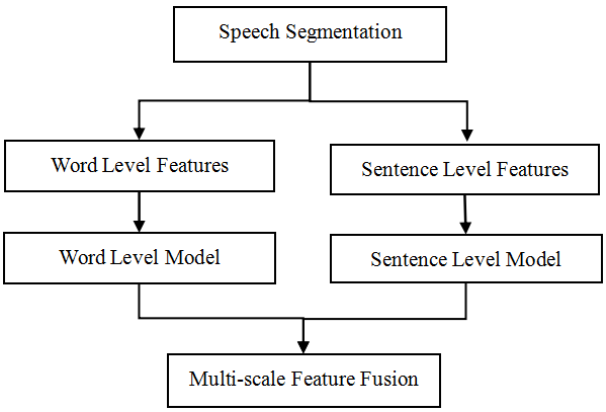

FIGURE 1. Multi-scale feature fusion for word and sentence level features

4. **Recognition Methodology.** Deep Neural Network (DNN) is used for feature modeling and recognition. In the feature analysis section we propose to fuse two feature models in order to get a super vector for recognition. The establishment of word level model and sentence level model can be implemented by DNN. Deep learning is different from traditional machine learning algorithms in the way that it can abstract higher level features that are independent from low level data. The outputs of the DNN are then used to train the multi-scale feature model by an ordinary three-layer neural network.

Deep Neural Network has been applied to speech emotion recognition [14]. In our paper, we study the age estimation problem, and emotion is treated as noise in our case. DNN contains many hidden layers, which is different from traditional neural network. DNN belongs to the family of directed graphical models and it provides the posterior probability $\mathrm{p}_{y|x}(y = q|x)$ where q stands for a class, x stands for an input feature and y stands for the output of the deep neural network. If we denote the input of the first layer L as vectors $\boldsymbol{v}^{l}$ and l is the index of nodes. The hidden binary vectors are then represented as $\boldsymbol{h}^{l}$. Let $\boldsymbol{h}_{m}^{l}$ be the hidden unites and m is the index of the unit. The total number of the hidden units is N. The posterior probability then can be modelled according to Equation (3)

$$\mathrm{p}^{l}\left(\mathbf{h}^{l}|\mathbf{v}^{l}\right) = \prod_{m=1}^{N} \frac{e^{z_m^l(\mathbf{v}^l)\mathbf{h}_m^l}}{e^{z_m^l(\mathbf{v}^l)} + 1} \tag{3}$$

where $\mathrm{z}^l\left(\mathbf{v}^l\right) = \left(\mathbf{W}^l\right)^T \mathbf{v}^l + \mathbf{a}^l$ $\mathbf{W}$ is the weight and $\mathbf{a}$ is the bias vector. The output layer provides the posterior probabilities:

$$\mathrm{p}_{y|x}(y = q|x) = \frac{e^{z_q^L(\mathbf{v}^L)}}{\sum e^{z_{q'}^L(\mathbf{v}^L)}} \tag{4}$$

The Gaussian mixture model (GMM) is adopted for comparison with DNN. GMM is defined as the weighted sum of M members as shown in Equation (5).

$$p(\boldsymbol{X}_t|\lambda) = \sum_{i=1}^{M} a_i b_i(\boldsymbol{X}_t) \tag{5}$$

where $\boldsymbol{X}$ is a D-dimension vector, $b_i(\boldsymbol{X})$, $i = 1, 2, \ldots, M$ is the Gaussian distribution of each member; a_i, $i = 1, 2, \ldots, M$ is the mixture weight.

$$b_i\left(\boldsymbol{X}_t\right) = \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}_i|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{X}_t - \boldsymbol{U}_i)' \sum\nolimits_i^{-1} (\boldsymbol{X}_t - \boldsymbol{U}_i)\right\} \tag{6}$$

where the mixture weight satisfies: $\sum\limits_{i=1}^{M} a_i = 1$.

The complete GMM parameters may be represented as: $\lambda_i = \{a_i, \boldsymbol{U}_i, \boldsymbol{\Sigma}_i\}$, $i = 1, 2, \ldots, M$.

According to the Bayes theory, the classification can be made by maximizing the posterior probability: $E = \arg\max\limits_k \{p(\boldsymbol{X}_t|\lambda_k)\}$.

5. **Experimental Results.** In order to verify the effectiveness of the proposed multi-scale emotional feature, we compare the recognition results between the multi-scale feature with two traditional features, namely the turn-wise feature and the frame-wise feature. The turn-wise feature takes a segment of speech to analyze and the frame-wise feature takes a frame of the speech to analyze. According to the comparison results shown in Figure 2, the proposed multi-scale feature outperforms the traditional features constantly over all six types of emotions.

We further test the recognition rates over different databases, as shown in Table 2. Ten-fold cross validation is adopted for all tests. On the German database, anger and happiness reach at the higher rates of 83.5% and 82.1% respectively, while neutrality only achieves a rate of 71.0%. On the Chinese database, anger and sadness reach at the highest rates of 87.0% and 88.4% respectively, while fear only reaches at 65.6%. On the Vietnamese database, anger and neutrality reach at the highest rates of 87.2% and 90.7% respectively, while sadness only achieves a rate of 76.2%. On the English database, anger reaches at the highest rate of 87.2% and happiness only achieves a rate of 67.4%.

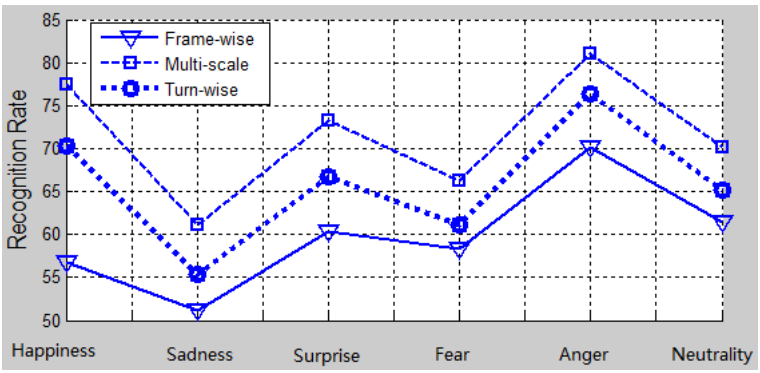

FIGURE 2. Comparison of traditional feature and multi-scale feature fusion

TABLE 2. Recognition rates on different datasets (%)

Languages	Happiness	Sadness	Fear	Surprise	Anger	Neutrality
German	82.1	77.4	78.8	81.2	83.5	71.0
Chinese	73.2	88.4	65.6	76.1	87.0	77.4
Vietnamese	85.1	76.2	83.3	83.2	87.2	90.7
English	67.4	77.3	76.1	69.4	87.2	76.1

TABLE 3. Cross-language emotion recognition rate (%)

Testing Sample Types	Happiness	Sadness	Fear	Surprise	Anger	Neutrality
Happiness	77.5	4.5	3.2	7.1	1.8	5.9
Sadness	3.3	61.2	11.4	8.2	1.9	14.0
Fear	8.4	7.8	66.3	5.7	6.2	5.6
Surprise	8.8	7.3	6.9	73.2	2.1	1.7
Anger	5.4	3.7	5.5	2.3	81.1	2.0
Neutrality	4.3	11.8	8.4	2.7	2.6	70.2

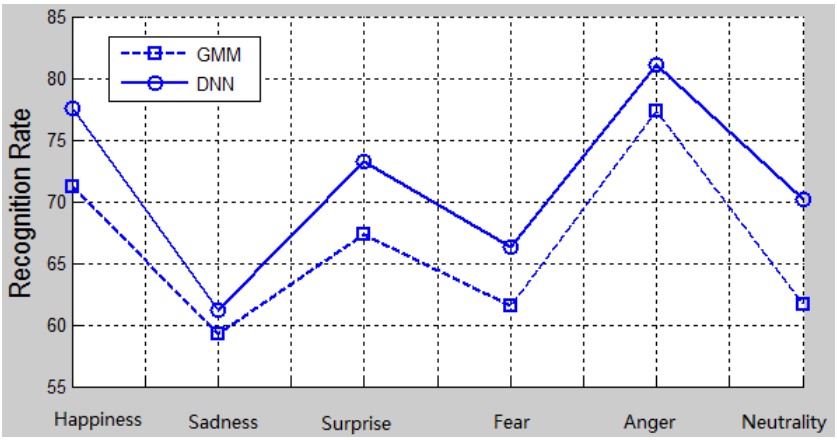

FIGURE 3. Comparison of deep neural network and Gaussian mixture model

We can see from the results that over the different languages, anger has been successfully recognized at relative high rates. Other types of emotions may differ from language to language. We further test the cross-language recognition ability using four languages, and the result is shown in Table 3. Relatively speaking the cross-language recognition result is lower than the result on signal database, and it is more difficult to improve the emotion recognition performance on cross-language dataset than on the single language dataset. According to the confusion matrix in Table 3, the mis-classification between sadness and fear is high, and the mis-classification between sadness and neutrality is also high. The emotion features related to valence level is more difficult to recognize than the features related to arousal level. High arousal level emotion types are well recognized in the cross-language test, such as anger, happiness and surprise. Therefore, the prosodic features that related to arousal level may have a better generalization ability over different languages.

Gaussian mixture model is adopted as an alternative module for deep neural network. The comparison of recognition performances is shown in Figure 3. The Gaussian mixture number is set to 16 and the expectation-maximization algorithm is adopted for the estimation of GMM parameters. We can see that DNN outperforms the GMM classifier over all six types of emotions. The deep neural network has a strong ability to abstract high

level features that are less dependent on the low level data. This property may improve the generalization ability over different languages.

6. **Conclusions.** In this paper we apply the deep neural network to the feature extraction under various time durations. The outputs of the DNN are then combined with ordinary neural network for emotion classification task. The proposed method has a strong generalization ability on four different languages, German, Chinese, Vietnamese and English. In the future we may further explore more efficient valence speech features in different languages.

REFERENCES

[1] N. Almeida, S. Silva and A. Teixeira, Design and development of speech interaction: A methodology, human-computer interaction *Advanced Interaction Modalities and Techniques*, Springer International Publishing, pp.370-381, 2014.

[2] J. Freitas, A. Teixeir and M. S. Dias, Multimodal corpora for silent speech interaction, *Proc. of the 9th edition of the Language Resources and Evaluation Conference*, pp.12-16, 2014.

[3] H. Hofmann, A. Silberstein, U. Ehrlich, A. Berton, C. Muller and A. Mahr, Development of speech-based in-car HMI concepts for information exchange Internet apps, *Natural Interaction with Robots, Knowbots and Smartphones*, Springer, New York, pp.15-28, 2014.

[4] M. El Ayadi, M. S. Kamel and F. Karray, Survey on speech emotion recognition: Features, classification schemes, and databases, *Pattern Recognition*, vol.44, no.3, pp.572-587, 2011.

[5] C. Huang, R. Liang, Q. Wang et al., Practical speech emotion recognition based on online learning: From acted data to elicited data, *Mathematical Problems in Engineering*, pp.1-6, 2013.

[6] C. Zou, C. Huang, D. Han and L. Zhao, Detecting practical speech emotion in a cognitive task, *Proc. of the 20th International Conference on Computer Communications and Networks*, pp.1-5, 2011.

[7] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier and B. Weiss, A database of German emotional speech, *Proc. of Annual Conference of the International Speech Communication Association*, vol.5, pp.1517-1520, 2005.

[8] A. Batliner, C. Hacker, S. Steidl, E. Nöth, S. D'Arcy, M. J. Russell and M. Wong, "You stupid tin box" – Children interacting with the AIBO robot: A cross-linguistic emotional speech corpus, *Proc. of Language Resources and Evaluation Conference*, pp.1-4, 2004.

[9] B. Schuller, G. Rigoll and M. Lang, Hidden Markov model-based speech emotion recognition, *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol.2, 2003.

[10] K. R. Scherer, A cross-cultural investigation of emotion inferences from voice and speech: Implications for speech technology, *Proc. of Annual Conference of the International Speech Communication Association*, pp.379-382, 2000.

[11] Z. Wang, L. Zhao and C. Zou, Emotional speech recognition based on modified parameter and distance of statistical model of pitch, *Acta Acustica Sinica*, vol.31, no.1, pp.28-34, 2006.

[12] V. La, C. Huang, C. Zha and L. Zhao, Emotional feature analysis and recognition from Vietnamese speech, *Signal Processing*, vol.29, no.10, pp.1423-1432, 2013.

[13] O. Martin, I. Kotsia, B. Macq and I. Pitas, The eNTERFACE'05 audio-visual emotion database, *Proc. of the 22nd International Conference on Data Engineering Workshops*, p.8, 2006.

[14] K. Han, D. Yu and T. Tashev, Speech emotion recognition using deep neural network and extreme learning machine, *Proc. of Annual Conference of the International Speech Communication Association*, Singapore, pp.223-227, 2014.