SOFTWARE: PRACTICE AND EXPERIENCE
Softw. Pract. Exper. 2017; **47**:1127–1138
Published online 24 February 2017 in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/spe.2487

Deep learning and SVM-based emotion recognition from Chinese speech for smart affective services

Weishan Zhang[1,*,†], Dehai Zhao[1], Zhi Chai[2], Laurence T. Yang[3], Xin Liu[1], Faming Gong[1] and Su Yang[4]

[1]*School of Computer and Communication Engineering, China University of Petroleum, No. 66 Changjiang West Road, Qingdao 266580, China*
[2]*Science and Technology on Optical Radiation Laboratory, Beijing 100854, China*
[3]*Department of Computer Science, St. Francis Xavier University, Antigonish, NS, Canada*
[4]*College of Computer Science and Technology, Fudan University, Shanghai 200433, China*

SUMMARY

Emotion recognition is challenging for understanding people and enhances human–computer interaction experiences, which contributes to the harmonious running of smart health care and other smart services. In this paper, several kinds of speech features such as Mel frequency cepstrum coefficient, pitch, and formant were extracted and combined in different ways to reflect the relationship between feature fusions and emotion recognition performance. In addition, we explored two methods, namely, support vector machine (SVM) and deep belief networks (DBNs), to classify six emotion status: anger, fear, joy, neutral status, sadness, and surprise. In the SVM-based method, we used SVM multi-classification algorithm to optimize the parameters of penalty factor and kernel function. With DBN, we adjusted different parameters to achieve the best performance when solving different emotions. Both gender-dependent and gender-independent experiments were conducted on the Chinese Academy of Sciences emotional speech database. The mean accuracy of SVM is 84.54%, and the mean accuracy of DBN is 94.6%. The experiments show that the DBN-based approach has good potential for practical usage, and suitable feature fusions will further improve the performance of speech emotion recognition. Copyright © 2017 John Wiley & Sons, Ltd.

Received 2 March 2016; Revised 26 October 2016; Accepted 19 January 2017

KEY WORDS: speech emotion recognition; feature fusion; support vector machine; deep belief network

1. INTRODUCTION

Emotion status is an outward manifestation of people's inner thoughts, which plays an important role in rational actions of human being. Emotion recognition is defined as detecting the emotion status of a person. Emotion status is useful for intelligent human–machine interfaces that are helpful for better human–machine communication and decision making [1]. Knowing emotional status is also very helpful for precise services of health care, entertainment, and other services. For example, emotion status can provide an in-car board system with information about the driver to initiate safety strategies [2].

Four decades ago, Williams and Stevens [3] presented an early attempt to analyze vocal emotion, which takes the first step on emotion recognition. Since 2 decades ago when Picard put forward affective computing [4], emotion recognition became a hot topic, and great efforts have been made in this field. Researchers have tried to predict high-level affective content from low-level

*Correspondence to: Weishan Zhang, School of Computer and Communication Engineering, China University of Petroleum, No. 66 Changjiang West Road, Qingdao 266580, China.
†E-mail: zhangws@upc.edu.cn

human-centered signal cues using every possible features extracted from the whole body such as speech, facial expression, and heart rate.

However, emotion recognition is a complex task that is furthermore complicated because there is no unambiguous answer to what the correct emotion is for a given sample [5]. Emotion recognition is still a challenging task because of the following reasons. First, it is difficult to decide which feature should be chosen for the recognition system. For speech emotion recognition, the acoustic variability introduced by the existence of different sentences, speakers, speaking styles, and speaking rates adds another obstacle because these properties directly affect most of the common extracted speech features such as pitch and energy contours [6]. Moreover, there may be more than one perceived emotion in one sample, and it is difficult to determine the boundaries between different portions. Additionally, there are various noises in practical applications, and how to handle influences of the noises in sophisticated applications is difficult [7].

In this paper, we extracted five kinds of speech emotion features, including pitch, short-term energy, short-term zero-crossing rate, formant, and Mel frequency cepstrum coefficient (MFCC), and fused them in different ways. In addition, we applied both support vector machine (SVM) and deep belief network (DBN) to conduct speech emotion recognition and compared these two methods. In fact, our work is more challenging than those who use Berlin emotion database or USC-IEMOCAP database because we use Chinese emotion database. The Chinese language's prosodic features are complex, which make Chinese speech emotion recognition more difficult.

The contributions of this paper are as follows:

- We explored two popular classification methods, SVM and DBN, to conduct speech emotion recognition on Chinese emotion database and compared their performance.

  The most widely used speech emotion feature, MFCC, was extracted from speech samples and used to evaluate three kinds of SVM methods: one versus one, one versus rest, and minimum output coding (MOC). At the same time, the extracted features were used to evaluate DBN. We will consequently find the advantages of each classification method by analyzing the results of them. During the optimization process, we found some rule of adjusting parameters.
- We tried to extract more speech features and evaluated various feature fusions for speech emotion recognition.

  Five kinds of common speech features, including pitch, short-term energy, short-term zero-crossing rate, formant, and MFCC, were extracted and used to evaluate the performance of speech emotion recognition with DBN. Feature fusion is not a simple feature vector connection but a weight adjustment process. That is to say, we should explore a suitable weight for each feature and combine them into a new feature vector to achieve the best result.

The rest of the paper is organized as follows. Section 2 describes some features that used widely in speech emotion analyzation. Section 3 introduces the classifier we use. Section 4 is the evaluation of our methods. Conclusion and future work end the paper.

2. BACKGROUND KNOWLEDGE OF SPEECH EMOTION FEATURE EXTRACTION

It is known that any emotion from the speaker is represented by a large number of parameters that are contained in the speech signals and the changes in these parameters will result in corresponding change in emotions. Thus, an important step in the design of speech emotion recognition system is extracting suitable features that efficiently characterize different emotions. Many researches have shown that effective parameters to distinguish a particular emotion status with high efficiency are spectral features such as MFCC, linear prediction cepstrum coefficient, and prosodic features such as pitch frequency, formant, short-term energy, and short-term zero-crossing rate [8]. Each speech signal is divided into small intervals of 20 to 30 ms, which are known as frames [8], and features are extracted from every frame. All the extracted features have its own unique significances to the speech emotion recognition, and they are described as follows:

- Pitch frequency

DOI: 10.1002/spe

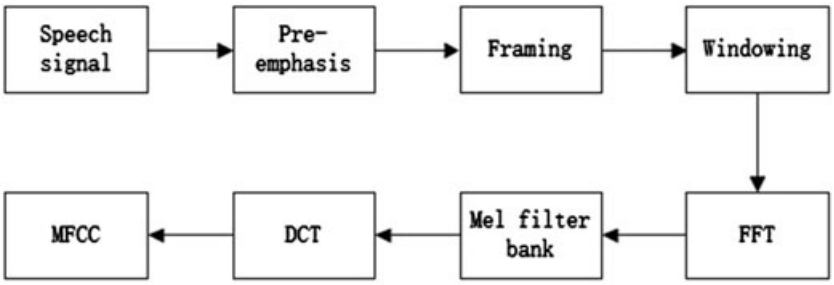

Figure 1. The process of extracting Mel frequency cepstrum coefficient (MFCC).

Generally, pitch frequency is related with the length, thickness, and toughness of one's vocal cords and reflects personal characteristics to a large degree. As one of the most important parameter of describing excitation source in speech signal processing field, pitch frequency is used widely to solve speaker identification problem, especially for Chinese. Because Chinese has intonation, which is very helpful to Chinese semantic comprehension, accurate pitch detection plays outstanding role in Chinese speech emotion recognition.

- Short-term energy

  Short-term energy can be used to distinguish voice and noise because voice has more energy than noise. If the environment noise and the input noise are low enough, voice and noise can be separated easily by computing short-term energy of the input speech signal. Besides, short-term energy-based algorithm performs well when detecting voiced sound because the energy of voiced sound is much more than voiceless sound. But it is hard to have a good performance for voiceless sound perception.
- Short-term zero-crossing rate

  Short-term zero-crossing rate indicates the times of speech signal wave crossing horizontal axis in each frame. It can be used to distinguish voiced sound and voiceless sound because the high band of speech signal has a high zero-crossing rate because low band has a low zero-crossing rate. In addition, short-term zero-crossing rate and short-term energy are complementary approximately because high short-term energy corresponds low short-term zero-crossing rate while low short-term energy corresponds high short-term zero-crossing rate.
- Formant

  Formant is an important parameter for reflecting vocal track information, which carries speech identity attribute like an ID card. The position of formant varies with different emotion because the pronunciation with different emotion can make vocal track change accordingly.
- MFCC

  The MFCC is extracted on the basis of the human ears' hearing characteristics, which is the most common kind of characteristic parameter of speech emotion recognition. Work on human auditory systems shows that the response of human ears to different frequencies is nonlinear but logarithmic. Human auditory system is good enough that it cannot only extract semantic information but also emotion information. MFCC is extracted by imitating human auditory system, which is helpful for speech emotion recognition. The process of extracting MFCC is shown in Figure 1.

3. CLASSIFICATION METHODS IN OUR WORK

On the one hand, the SVM-based method has unique advantages because of its simplicity and capabilities for classification. However, the choice of SVM kernel function and the parameter optimization is still difficult. Additionally, the system can recognize more than two emotion categories, which need a multi-classification SVM classifier. On the other hand, deep learning has been used widely in multi-classification problems, and it performs well. The characteristic of self-

DOI: 10.1002/spe

learning reduces the heavy work of feature extraction and selection, which makes classification more effective.

3.1. Principles of support vector machine multi-classification algorithm

There are mainly two kinds of implementation of SVM multi-classification algorithm [9] :

- The original optimization in proposed SVM multi-classification problem can be changed properly so that it can calculate all the multi-classification decision function simultaneously.
- The SVM multi-classification problem is divided into a series of dichotomous SVM problems, which can be solved directly. The final discrimination will be obtained on the basis of the results of this series of dichotomous SVM problems.

Although the first method seems simpler, it is actually difficult to achieve because the process of its parameter optimization is too complicated and the calculation of this algorithm is very large. So we choose the second method, which mainly includes five forms:

- One versus rest

  One versus rest is one of the earliest and most widely used method. When solving a k-class problem, k dichotomous machines are constructed, among which the i-th dichotomous machine is used to distinguish the i-th emotion status and the rest of emotion status. That is to say, during the training process, class i is taken as the positive category, and the remaining classes are taken as the negative category. The input signal passes through k classifier respectively and a total of k output values (1 or 0) are obtained. If only one 1 is presented, obviously, it is the ideal result; we will have the emotion status directly. But in practice, errors occur in the construction of decision function, and there may be more than one 1 or no 1. So we choose the maximum value of the corresponding category as the result. The advantages of one versus rest are that only k dichotomous machines need to be trained for a k-class problem. Thus, the calculation is few, and the algorithm is easy to implement. The speed of the classification is fast relatively.
- One versus one

  This method trains a classifier between every two categories. Thus, there will be $k(k-1)/2$ decision functions for a k-class problem. The unknown samples will input into every classifier, and the classifier will vote for the corresponding category. The final results will be obtained by finding the category with the most votes. But sometimes, the voting method may lead to a problem that one unknown sample belongs to multiple categories because one class can be voted several times.
- Directed acyclic graph

  Directed acyclic graph (DAG) is a directed graph with no directed cycles. It is formed by a collection of vertices and directed edges; each edge connects one vertex to another [10]. DAG is simple because only $(k-1)$ decision functions are needed to figure out the outcomes for a k-class problem. Therefore, the classification speed is faster than one versus one method. And there is no misclassification and refused subregion. In addition, it has a high fault tolerance and high accuracy than dichotomous method because of its special structure. However, the top–down error accumulation phenomenon is an inherent drawback to hierarchy structure; DAG is no exception, that is, if a classification error occurs on one node, it will spread to the following nodes.
- Decision tree

  Decision tree starts from the root that contains the full set of samples, and the root is divided into two subcategories using decision methods. Then each subcategory is divided into two smaller subcategories continually until the child node contains only one class. There are many SVM multi-classification methods for decision tree, and the main difference lies in the different designs of tree's structure. The average number of classifier for a complete binary tree is $\log_2 k$, which is fewer than that of DAG. Thus, the classification speed is faster.
- Error-correcting output codes

DOI: 10.1002/spe

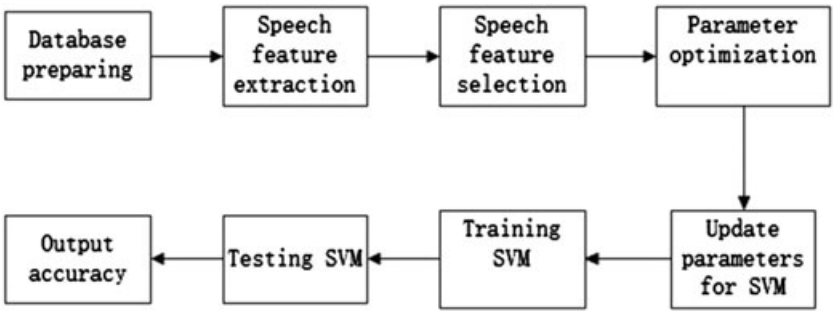

Figure 2. Parameter optimization of support vector machine (SVM).

Error-correcting output codes train a binary classifier for every class, which is similar with one versus rest method. The positive class (1) and negative class (0) constitute the codebook, whose row represents class and column represents binary classifier. When inputting an unknown sample, the output from the binary classifier will be combined as a code. Then the code will be compared with that in codebook, and the class with minimum distance is taken as the final result. The most common method of distance computing is Hamming distance.

3.2. Parameter optimization of support vector machine multi-classification algorithm

In general, the penalty factor sig2 and kernel parameter gam are the most important parameters in SVM multi-classification algorithm, which are influenced by the size of samples and training iteration. Fixed initial values are given by experience and optimized by experiments. The optimization process are shown in Figure 2.

Step 1. Parameter initialization. According to experience, sig2 and gam are initialized to 2.
Step 2. Input the dataset, and train the SVM with an iteration number of 20, which should be large enough for the dataset.
Step 3. Analyze the statistical outcomes after completing 20 iterations, and record the corresponding sig2 and gam value with the highest accuracy.
Step 4. Update the SVM parameters with the recoded sig2 and gam value. And the highest accuracy of each emotion status will be obtained.

We can find the most optimal parameter by comparing several different experiments and make statistics of the range of accuracy. Then get a better performance for the parameter optimization.

3.3. Deep belief networks

Supposing that there is a bipartite graph, in which one is visible layer and another is hidden layer. The nodes in the same layer are unconnected, and all the nodes in the bipartite structure are random binary variables. Meanwhile, the probability distribution is Boltzmann distribution. This structure is known as restricted Boltzmann machine (RBM).

Deep belief networks are constituted by many RBMs. Comparing with conventional neural networks with discrimination model, DBN is a probability generation model, which is a joint distribution of observation and labels [11]. That is to say, the probability generation model estimates P(Observation|Label) and P(Label|Observation) because the discrimination model only estimates P(Label|Observation). There may exist some problems when using DBN. For example, a labeled dataset is needed for training, the speed of learning process is slow, and the learning results may converge to local optimal solution because of unsuitable parameter.

A typical DBN model is shown in Figure 3. The networks contain only one visible layer and one hidden layer. All the layers connect with each other, but notes in the same layer are unconnected. The hidden layer is trained to catch the correlation between high-level data expressed by the visible layer. The connection of DBN is determined by top–down weights generation, and RBMs constitute

DOI: 10.1002/spe

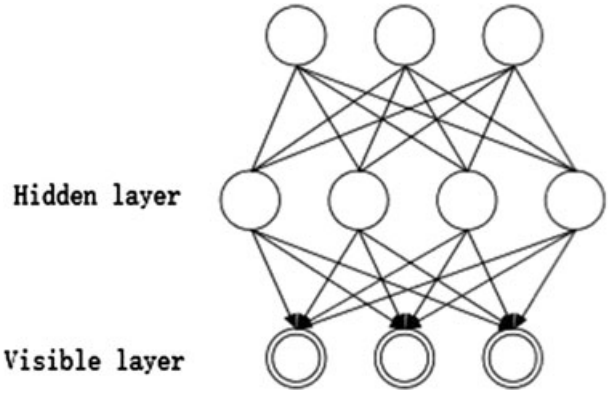

Figure 3. A simple structure of deep belief network.

the whole structure like building blocks. Comparing with sigmoid belief networks, DBN performs well in weights connection learning.

Ignoring the complex work on feature representation, extraction, and selection, DBN can effectively generate discriminative features that approximate the complex nonlinear dependencies between features in the original set. So it has been widely applied to speech processing as well as emotion recognition tasks [12]. In our work, we present a DBN model to investigate audio feature learning in emotion domain, which is a three-layer model. The audio features from the input layer are learnt in the hidden layer. The learnt features from the hidden layer are used as the input to the output layer. Finally, a binary list that represents the emotion category of the speech sample is obtained.

4. EVALUATION

4.1. Database

The database we used in the experiments was Chinese Academy of Sciences emotional speech database,[‡] which was provided by the Chinese Academy of Sciences Institute of Automation State Key Laboratory of Pattern Recognition Research Group. The corpus were recorded by four processional announcers, including two men and two women. The total 1200 sentences contains six kinds of emotion status: anger, fear, joy, sadness, surprise, and neural status; and each emotion status has 50 recording scripts. Pitch frequency, short-term energy, short-term zero-crossing rate, formant, and MFCC were extracted from speech samples, and the corresponding statistic characteristics such as minimum, maximum, mean, variance, and standard deviation were calculated at the same time. Finally, a feature dataset with total 74 dimensions, which contained 60 dimensions of MFCC and 14 dimensions of statistical features, was obtained.

4.2. Experiments with support vector machine and deep belief network

Firstly, we compared the two classification methods introduced in Section 3 and analyze the performance of them. Generally, MFCC is the most widely used feature for speech emotion recognition because it can simulate human auditory system in a large degree. So this experiment used MFCC that was organized as a feature vector to train SVM and DBN.

As is known to all that speech has many different inherent characteristics between man and woman. For example, the pitch has large discrepancies between different genders, and women have higher pitch than men. Considering differences between man and woman, whose vocal characteristics are so disparate even if they are in the same emotion status, which affects emotion analysis greatly, we divided the dataset into two groups, the male group and the female group, and each group

[‡]http://www.datatang.com/data/39277

DOI: 10.1002/spe

Table I. True positive of male emotion recognition with SVM.

Emotion	OneVsAll (%)	MOC (%)	OneVsOne (%)	Dichotomy (%)
Anger	36	86	74	84
Fear	92	100	98	96
Joy	18	84	10	74
Neutral	94	100	100	98
Sadness	80	98	94	74
Surprise	26	98	18	96

MOC, minimum output coding; SVM, support vector machine.

Table II. True positive of female emotion recognition with SVM.

Emotion	OneVsAll (%)	MOC (%)	OneVsOne (%)	Dichotomy(%)
Anger	94	96	98	90
Fear	96	100	98	72
Joy	66	100	86	98
Neutral	86	96	96	86
Sadness	100	98	100	100
Surprise	84	98	94	98

MOC, minimum output coding; SVM, support vector machine.

contained 500 samples as the training set and 100 samples as the test set. Three kinds of SVM multi-classification methods, one versus one, one versus rest, and MOC, were evaluated. Meanwhile, the conventional dichotomous method was implemented to make comparison with these three methods. The results of male emotion recognition are shown in Table I.

As we can see from Table I, when using one versus one method to deal with male dataset, fear, neutral status, and sadness achieve a high accuracy of 92%, 94%, and 80%, respectively. Similarly, these three emotions achieve a high accuracy when using one versus one method, because all the emotions have the highest accuracy when using MOC method. We can see obviously that the three methods perform better than conventional dichotomous method.

The results of female emotion recognition are shown in Table II.

We can see from Table II that when using one versus rest method, only joy has a low accuracy of 66% while the rest five emotions have a high accuracy. And one versus one method can have a good result for all the emotions. MOC method performs better than other methods, which is the same with the results of male emotion recognition. We can also see that traditional dichotomous method cannot have a good result.

The next experiment used DBN to implement emotion recognition. We also divided the dataset into two groups, the male group and the female group, and each group contained 500 samples as the training set and 100 samples as the test set. After loading training set with labels added by ourselves, we trained DBN with the adjusted parameters for six kinds of emotions. Then we evaluated the trained model with test set. During the training process, we found that these four parameters, *numpochs*, *bachsize*, *momentum*, and *alpha*, mainly affect the results. Therefore, the challenge is finding different parameters to have the best results, and it is important to find the rule of parameter adjustment by more experiments. The final results are shown in Table III.

The mean accuracy of DBN is 90.33%, which is higher than SVM of 84.54%. That is to say, when we apply MFCC to analyze speech emotion, the performance of DBN is better than SVM, which satisfies the expected result. So we only use DBN to compare various feature fusions in the following experiments.

4.3. Experiments with feature fusion

As other features are also very important for reflecting speech emotion, the recognition process should take all the useful features into consideration. But the feature fusion schemes are so many that we cannot test them one by one, and not every feature plays the decisive role in analyzing

DOI: 10.1002/spe

Table III. True positive using Mel frequency cepstrum coefficient with deep belief network.

Emotion	Male (%)	Female (%)
Anger	86	98
Fear	88	82
Joy	94	86
Neutral	88	98
Sadness	90	88
Surprise	94	82
Average	90	90.3

speech emotion, or some features are too weak to reflect emotion clearly, such as short-term energy and short-term zero-crossing rate. So we selected three most widely used speech features including MFCC, pitch, and formant and fused them in different ways.

Tables IV and V are the results using MFCC fused with pitch and formant, respectively. Although the results of two group are very similar, we can find some characteristics of different feature fusions. For example, when using MFCC and pitch, neutral status has the lowest accuracy. But when using MFCC and formant, fear performs badly. This result confirms that different feature fusions will contribute to a certain kind of emotion recognition.

What is more, we conducted another experiment using all the five kinds of features we extracted, including pitch, short-term energy, short-term zero-crossing rate, formant, and MFCC, in which short-term energy and short-term zero-crossing rate are represented by their minimum, maximum, mean, variance, and standard deviation values. We can see from Table VI that the fusion of all features interaction decreases the contingency of single feature, leads to similar results of male group and female group, and at the same time, achieves a higher accuracy. It is necessary to note that the fusion features were not simple series of different features and they needed adjusted weights. The weights of one feature can be adjusted according to its contribution to reflect emotion or the

Table IV. True positive using Mel frequency cepstrum coefficient and pitch features with deep belief network.

Emotion	Male (%)	Female (%)
Anger	92	94
Fear	90	90
Joy	94	90
Neutral	88	90
Sadness	94	93
Surprise	92	90
Average	91.6	91

Table V. True positive using Mel frequency cepstrum coefficient and formant features with deep belief network.

Emotion	Male (%)	Female (%)
Anger	92	94
Fear	88	92
Joy	94	94
Neutral	94	90
Sadness	94	94
Surprise	94	90
Average	92.6	92.3

DOI: 10.1002/spe

Table VI. True positive using all features with deep belief network.

Emotion	Male (%)	Female (%)
Anger	92	94
Fear	94	96
Joy	90	96
Neutral	98	96
Sadness	98	90
Surprise	96	96
Average	94.6	94.6

dimension of it. For example, the dimension of MFCC was higher than pitch, so we gave pitch a larger weight than MFCC. For a speech sample, we modified suitable weights for various features and linked them as a new vector. All the adjustments should be changed by the practical applications.

4.4. Discussion

Comparing with the existing work, our work has some characteristics:

- Researchers have used various SVM methods to detect speech emotion [5] [13], with the accuracy of 74.37% and 89%, but deep learning method was confirmed as a better choice [14], whose method was similar to ours. But comparing with their accuracy of 62.42%, we achieved a higher accuracy of 94.6% with DBN.
- Different feature fusions will contribute to some kinds of emotion recognition, but more features can improve the performance, provided that we have optimized weights for every feature and connected them into a new vector.
- Usually, training deep neural network needs large amount of data. Berlin speech emotion database is one of the most widely used database that contains about 800 sentences, but it is even smaller than the Chinese speech emotion database used in this paper. Such little data lead to some problems like overfitting. Therefore, we designed a suitable DBN architecture for the small database that contains three hidden layers. And in the experiment, 3000 iterations with the learning rate of 0.01 performed well for this database.
- The features extracted by different algorithms have diverse scale, which makes the dataset unbalanced. It has been demonstrated that the unbalanced dataset would make the machine learning algorithms can not be convergent. Therefore, normalization has been conducted on all the features so that they could have equal contribution to training SVM and DBN and make training process easy to be convergent.
- In terms of process of parameter optimization, generally, a big basic learning rate such as 0.1 and a learning rate decay is set up during the training process. But actually, if the training process is convergent, the gradient would get smaller and smaller, which is equivalent to learning rate decay, and the double learning rate decay would slow the convergence speed. So a smaller basic learning rate without decay has been set up in our work.
- It has been found in the experiments that changing parameters would achieve diverse accuracy for different emotions. For example, one set of parameters could make happy get the highest accuracy, but another set of parameters could make the algorithm be more sensitive to sad. Actually, it is always a hard task to optimize parameters for deep learning because deep learning is a black box with low interpretability of intermediate result. So it would be more practical to find a set of suitable parameters by doing a lot of experiments to make all kinds of emotions achieve the highest mean accuracy.

5. RELATED WORK

Prashant Aher *et al*. [13] put forward a noise robust speech emotion recognition system. They found that MFCC is good in clean environment but degrades when there exists data mismatch between training and testing phase. Features extracted from input speech samples were fed into SVM with

DOI: 10.1002/spe

RBF kernel function and obtained an accuracy of 89% on Berlin emotion database. Our work used more features to conduct a more comprehensive speech emotion analysis, and the DBN we applied is robust to noise.

In [15], the authors generated feature representations from both acoustic and lexical levels. They extracted both low-level acoustic features and lexical features. And they also used the traditional Bag-of-Words features. Their work was evaluated on USC-IEMOCAP database and achieved an accuracy of 69.2%. The experimental results showed that late fusion of both acoustic and lexical features is suitable for speech emotion recognition task. Although we did not take lexical features into consideration, the deep learning method we used could achieve a better performance than the traditional method they used.

In [16], the authors proposed modulation spectral features for the automatic recognition of human affective information from speech. The features were obtained using an auditory filter bank and a modulation filter bank for speech analysis. They evaluated their system on Berlin emotion database and Vera am Mittag database with SVM and achieved an overall recognition rate of 91.6%, which was really a good result. But the features they extracted were not the popular features such as MFCC and Pitch which are demonstrated to be effective for speech emotion recognition.

Yelin Kim *et al*. [14] focused on deep learning techniques that can overcome the limitations of complicated feature selection by explicitly capturing complex nonlinear feature interactions in multimodal data. They proposed and evaluated a suite of DBNs models and demonstrated that these models show improvement in emotion classification performance. An accuracy of 62.42% has been achieved on the Interactive Emotional Dyadic Motion Capture Database.Their work has shown that using deep learning feature can achieve a better performance than using artificial methods, which inspires our work. So we fine tuned the DBN and got a better result.

In [17], the authors removed the Gauss white noise with the adaptive filter. Then the MFCC based on empirical mode decomposition was extracted. At last, they presented an effective method for speech emotion recognition on the basis of fuzzy least squares support vector machines so as to realize the speech recognition of four main emotions. The experiment results showed that this method has the better antinoise effect when compared with traditional SVMs. But comparing with deep learning method that we used, SVM is still a kind of shallow model, which can not express so many speech emotion features.

Chi-Chun Lee *et al*. [18] proposed a hierarchical binary decision tree approach to realize emotion recognition. They performed feature selection on the 384 features including pitch frequency, short-term energy, and MFCC using the software SPSS and obtained a reduced feature set that was in a range of 40–60. The results showed that comparing with SVM baseline model, this approach has an improved accuracy of 3.37% and 7.44% on the AIBO database and the USC-IEMOCAP database, respectively. Their work used third party software to refine features, which complicated the whole process. Actually, deep learning can refine features during the training process.

Akshay S. Utane *et al*. [5] handled five emotion status (joy, sadness, surprise, anger, and neutral status) using two different classifiers, SVM and GMM. Prosodic features such as pitch frequency and spectral features such as MFCC were extracted. And the system was evaluated on Berlin emotion database and achieved an accuracy of 12.18–74.37% among different emotion status. They found that both SVM and HMM provide relatively similar accuracy for classification. Similar to their work, we also used prosodic features and spectral features, but we demonstrated that using deep learning features can achieve a better result.

In [19], the authors proposed a speech emotion recognition system using multi-algorithm, that is, the MFCC and discrete wavelet transform-based algorithm have been successfully used to extract emotional information from speech signal. Similarly, they used SVM as the classifier, but the speech database they use is created by themselves. An accuracy of 11–89% is achieved among different emotion status.

DOI: 10.1002/spe

6. CONCLUSIONS AND FUTURE WORK

In this work, we investigated the utility of single feature and multi-features for speech emotion recognition. The results showed that multi-features fusion can improve the performance of speech emotion recognition. However, the combination was not a simple series process, and it needs suitable weights for different features according to the practical applications. Moreover, both SVM and DBN can be used for multi-classification well. The comparison of the classification performance between SVM and conventional dichotomous method demonstrated that SVM is a better choice for multi-classification problem. But the evaluation of DBN showed that deep learning method can make the best of low-level features to complete high-level emotion detection. That is to say, deep learning method has advantage over the SVM method when facing multidimensional features, because it avoids artificial selection of complex features.

In the future, we will collect more speech data to train DBN and generalize the network model. What is more, lexical feature is a very important part of speech. If we combine lexical feature with audio feature, the system will work better. Besides, the single speech feature is too few to reflect one's real emotion. The combination of various features from the whole body such as heart rate and facial expression is a bigger challenge because all these features will form a high-dimension and non-linear dataset, which requires massive computing resources and we should guarantee the efficiency, scalability and reliability of the system [20].

ACKNOWLEDGEMENT

This research was supported by the Program on Innovative Methods of Work from Ministry of Science and Technology, China (Grant No. 2015010300), National Natural Science Foundation of China (Grant No. 61309024), Natural Science Foundation of Shandong Province (Grant No. ZR2014FM038), and also supported by Fundamental Research Funds for the Central Universities.

REFERENCES

1. El Ayadi M, Kamel MS, Karray F. Survey on speech emotion recognition: features, classification schemes, and databases. *Pattern Recognition* 2011; **44**(3):572–587.
2. Schuller B, Rigoll G, Lang M. Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, (ICASSP'04)*, Vol. 1, IEEE, 2004; I–577.
3. Williams CE, Stevens KN. Emotions and speech: some acoustical correlates. *The Journal of the Acoustical Society of America* 1972; **52**(4B):1238–1250.
4. Picard RW, Picard R. *Affective Computing*, Vol. 252. MIT press: Cambridge, 1997.
5. Utane AS, Nalbalwar SL. Emotion recognition through speech using gaussian mixture model and support vector machine. *Emotion* 2013; **2**:742–746.
6. Banse R, Scherer KR. Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology* 1996; **70**(3):614–636.
7. Chen D, Hu Y, Wang L, *et al*. H-PARAFAC: Hierarchical Parallel Factor Analysis of Multidimensional Big Data. *IEEE Transactions on Parallel and Distributed Systems* 2016. DOI: 10.1109/TPDS.2016.2613054.
8. Shen P, Changjun Z, Chen X. Automatic speech emotion recognition using support vector machine. *International Conference on Electronic and Mechanical Engineering and Information Technology (EMEIT), 2011*, Vol. 2, IEEE, Harbin, Heilongjiang, China, 2011; 621–625.
9. Hsu C-W, Lin C-J. A comparison of methods for multiclass support vector machines. *IEEE Transactions on Neural Networks* 2002; **13**(2):415–425.
10. Bang-Jensen J, Gutin GZ. *Digraphs: Theory, Algorithms and Applications*. Springer Science & Business Media: Berlin, 2008.
11. Hinton GE. Deep belief networks. *Scholarpedia* 2009; **4**(5):47–59.
12. Mohamed A-r, Dahl GE, Hinton G. Acoustic modeling using deep belief networks. *IEEE Transactions on Audio, Speech, and Language Processing* 2012; **20**(1):14–22.
13. Aher P, Cheeran A. *Auditory processing of speech signals for speech emotion recognition*, 2014.
14. Kim Y, Lee H, Provost EM. Deep learning for robust feature generation in audiovisual emotion recognition. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2013*, IEEE, Vancouver, BC, Canada, 2013; 3687–3691.

DOI: 10.1002/spe

15. Jin Q, Li C, Chen S, Wu H. Speech emotion recognition with acoustic and lexical features. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2015*: IEEE, Brisbane, Australia, 2015; 4749–4753.
16. Wu S, Falk TH, Chan W-Y. Automatic speech emotion recognition using modulation spectral features. *Speech Communication* 2011; **53**(5):768–785.
17. Chu YY, Xiong WH, Chen W. Speech emotion recognition based on EMD in noisy environments. *Advanced materials research*, Vol. 831, Trans Tech Publications, 2014; 460–464.
18. Lee C-C, Mower E, Busso C, Lee S, Narayanan S. Emotion recognition using a hierarchical binary decision tree approach. *Speech Communication* 2011; **53**(9):1162–1171.
19. Joshi DD, Zalte MB. *Recognition of emotion from marathi speech using MFCC and DWT algorithms*, 2013.
20. Chen D, Hu Y, Cai C, Zeng K, Li X. Brain big data processing with massively parallel computing technology: challenges and opportunities. *Software: Practice and Experience*, 2016. DOI: 10.1002/spe.2418.