Deep Learning Based Emotion Recognition from Chinese Speech

Weishan Zhang[1(✉)], Dehai Zhao[1], Xiufeng Chen[2], and Yuanjie Zhang[1]

[1] Department of Software Engineering, China University of Petroleum,
No.66 Changjiang West Road, Qingdao 266580, China
zhangws@upc.edu.cn, {525044691,307304660}@qq.com

[2] Hisense TransTech Co., Ltd., No.16 Shandong Road, Qingdao, China
chenxiufeng@hisense.com

Abstract. Emotion Recognition is challenging for understanding people and enhance human computer interaction experiences. In this paper, we explore deep belief networks (DBN) to classify six emotion status: anger, fear, joy, neutral status, sadness and surprise using different features fusion. Several kinds of speech features such as Mel frequency cepstrum coefficient (MFCC), pitch, formant, et al., were extracted and combined in different ways to reflect the relationship between feature combinations and emotion recognition performance. We adjusted different parameters in DBN 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 highest accuracy was 94.6 %, which was achieved using multi-feature fusion. The experiment results show that DBN based approach has good potential for practical usage of emotion recognition, and suitable multi-feature fusion will improve the performance of speech emotion recognition.

1 Introduction

Emotion status is an outward manifestation of people's inner thoughts, which plays an important role in rational actions of human beings. There is a desirable requirement for intelligent human-machine interfaces for better human-machine communication and decision making [4]. However, there are still many problems existing in the process, such as limited range of application, low recognition rate, etc.

It is generally believed that emotion is quite a complex reaction that even human cannot distinguish perfectly. It is also a useful data for human-computer interaction. Therefore, finding an effective way to recognize hidden emotion efficiently and accurately is really meaningful work. This has introduced a relatively new research field named emotion recognition, which is defined as detecting the emotion status of a person. Four decades ago, Williams and Stevens [16] presented an early attempt to analyze vocal emotion, which takes the first step on emotion recognition. Since two decades ago when Picard put forward affective

C.K. Chang et al. (Eds.): ICOST 2016, LNCS 9677, pp. 49–58, 2016.
DOI: 10.1007/978-3-319-39601-9_5

computing [12], emotion recognition became a hot topic and great efforts has been made in this field. Researchers have tried to predict high-level affective content from low-level human-centred signal cues by using every possible features extracted from the whole body such as speech, facial expression, heart rate, etc.

Emotion recognition is particularly useful for applications which require a natural man-machine interaction where the response to a user depends on the detected emotion. For example, if computers are able to give real-time response depend on users' affect, it will be more life-like than conventional systems which operate according to rigid rules. In addition, if some important posts such as car, aircraft and workshop where the mental status of the workers may affect the working status seriously can be monitored and get the information of tiredness or stressfulness out in advance, it's possible to guarantee workers' safety and avoid the accident [5]. What's more, it will be helpful for doctors to make the right diagnosis and it can also be employed as a diagnostic tool for therapists [13].

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 [15]. 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 [2]. Moreover, there may be more than one perceived emotion in one sample and it is difficult to determine the boundaries between different portions.

In this paper, we apply DBN to conduct speech emotion recognition. In addition, we compare and analyze different feature combinations. 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 makes Chinese vocal emotion recognition more difficult. Thus, we try to extract more feature from speech samples and combine them in a suitable way to solve this problem.

The rest of the paper is organized as follows. Section two describes some features that used widely in speech emotion analyzation. Section three introduces the classifier we use. Section four 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 which 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 Mel frequency cepstrum coefficient (MFCC), linear prediction cepstrum coefficient (LPCC) and prosodic features such as pitch frequency, formant, short-term energy, short-term zero-crossing rate, and so on [14]. Each speech signal is divided into small intervals of 20 ms to 30 ms, which are known as frames [14], and features are extracted from every frame respectively. All the extracted features have its own unique significances to the speech emotion recognition and they are described as follows:

1. Pitch frequency. Generally, pitch frequency is related with the length, thickness, toughness of one's vocal cords and reflect personal characteristics to a large degree. As one of the most important parameter of describing excitation source in speech signal processing filed, 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.
2. 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 is 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's hard to get a good performance for voiceless sound perception.
3. 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 since 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.
4. 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.
5. MFCC. MFCC is extracted based on the human ears' hearing characteristics, which is the most common kind of characteristic parameter of speech emotion recognition. Work on human auditory systems show that the response of human ears to different frequency is nonlinear but logarithmic. Human auditory system is good enough that it can not only extract semantic information but also emotion information. MFCC is extracted by imitating human auditory system, which is helpful for improving accuracy. The process of extracting MFCC is shown in Fig. 1.

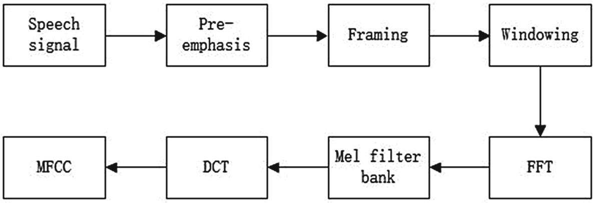

Fig. 1. The process of extracting MFCC

3 Deep Belief Networks

Deep learning has been used widely in multi-classification problems and it performs well. The characteristic of self-learning reduces the heavy work of feature selection, which makes classification more effective. Supposing that there is a bipartite graph, in which one layer is visible layer and the other layer 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 (DBN) is constituted by many RBM. Comparing with conventional neural networks with discrimination model, DBN is a probability generation model, which is a joint distribution of observation and labels [6]. That is to say, the probability generation model estimates P(Observation|Label) and P(Label|Observation) since 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 Fig. 2. The networks contain only one visible layer and one hidden layer. All the layer connects with each other but nodes in the same layer is 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 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 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 [11]. 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 learned in the hidden layer. The learned features from the hidden layer are used as the input to the output layer. Finally, a binary list which represents the emotion category of the speech sample is obtained.

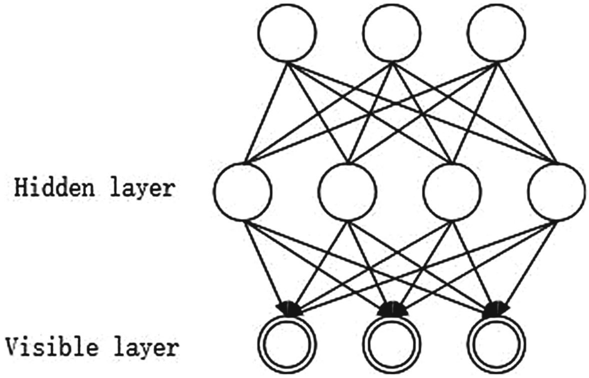

Fig. 2. A simple structure of DBN

4 Evaluation

4.1 Database

The database we use in the experiments is Chinese Academy of Sciences emotional speech database[1], which is provided by the Chinese Academy of Sciences Institute of Automation State Key Laboratory of Pattern Recognition Research Group. The corpus is recorded by four processional announcers, including two males and two females. The total 1200 sentences contains 6 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 are extracted from speech segments and the corresponding statistic characteristics such as minimum, maximum, mean, variance and standard deviation are calculated at the same time. Finally, a feature dataset with 74 dimension is obtained.

4.2 Experiments with DBN

Allowing for the difference between male and female, whose vocal characteristics are so disparate that will influence speech analyzation greatly, we divide the dataset into two groups, the male group and the female group. Each group contains 500 training samples and 100 testing samples. And MFCC is the most widely used feature for speech emotion recognition because it can simulate human auditory system in a large degree. So the first group of experiment used MFCC to train with DBN. After loading one group of dataset with labels added by ourselves, We adjusted the parameters of DBN to get a good model. Then we evaluated the trained model with the rest of speech samples. During the training process, we found that this four parameters of *numpochs*, *bachsize*, *momentum*

[1] http://www.datatang.com/data/39277.

and *alpha* mainly affect the results. The challenge is finding suitable parameters to achieve the best results. One group of parameters may be suitable for detecting male emotion but unsuitable for detecting female emotion because of the difference between male and female voice. So it's important to find the rule of parameter adjustment by a lot of experiments. The final true positive of using MFCC are shown in Table 1.

Table 1. Results only using MFCC features with DBN

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

As other features also play the important role of reflecting speech emotion, three group of experiments were conducted. We choose the most representative feature combinations in the experiments. Table 2 shows the true positive using MFCC and pitch. Table 3 shows the true positive using MFCC and formant. And Table 4 are the true positive using all features we extracted

Table 2. Results using MFCC and pitch features with DBN

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

Table 3. Results using MFCC and formant features with DBN

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

Though the results of each group are very similar, we can find some characteristics of different feature combinations. For example, when using MFCC and pitch, neutral status has the lowest accuracy. But when using MFCC and formant, fear performs badly. What's more, 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's necessary to note that the fusion features are not simple series of different features and they need adjusted weights for every feature. The weights of one feature can be adjusted according to its contribution to reflect emotion or the dimension of it. For example, the dimension of MFCC is higher than pitch, so we give pitch a larger weight than MFCC. All the adjustments should be changed by the practical applications.

Table 4. Results using all features with DBN

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

4.3 Experiments with SVM

Additionally, we conducted the experiment using SVM as a comparison of DBN. Three kinds of SVM multi-classification methods, one versus one, one versus rest and minimum output coding (MOC) are evaluated. Meanwhile, the conventional dichotomous method is implemented. The true positive of male emotion recognition are shown in Table 5.

Table 5. Results of male emotion recognition with SVM

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

As we can see from Table 5, when using one versus one method to deal with male dataset, fear, neutral status and sadness have a higher accuracy of 92 %, 94 % and 80 % respectively. Similarly, these three emotions have a higher accuracy when using one versus one method. Since 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 true positive of female emotion recognition are shown in Table 6.

Table 6. Results of female emotion recognition with SVM

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

From Table 6 we can see that when using one versus rest method, only joy has a low recognition of 66 % since the rest five emotions have a high accuracy.

And one versus one method can get a good result for all the emotions. MOC method performs better than other methods, which is same with the result of male emotion recognition. We can also see that conventional dichotomous method can't get a good result.

As shown in the experiments above, the highest accuracy of DBN is 94.6 %, which is higher than SVM method of 84.54 %. That is to say, when applying DBN, the performance of speech emotion recognition is better than SVM, which satisfies the expected result.

4.4 Related Work

Prashant Aher et al. [1] put forward a noise robust speech emotion recognition system. They find 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 are fed into SVM with RBF kernel function and get an accuracy of 89 % on Berlin emotion database. We also extracted MFCC as the main feature but we fused other speech features to achieve a better performance.

In [7], 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 (BOF) 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 were suitable for speech emotion recognition task. Though they used both acoustic and lexical feature, they didn't use deep learning method.

In [17], 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 a good performance.

Yelin Kim et al. [9] focused on deep learning techniques which can overcome the limitations of complicated feature selection by explicitly capturing complex non-linear feature interactions in multi-modal data. They proposed and evaluated a suite of deep belief networks models and demonstrate 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. We also used deep learning method but we applied some artificial assistance of selecting suitable feature fusion, which leaded to a better performance.

In [3], the authors removed the gaussian white noise with the adaptive filter. Then the Mel Frequency Cepstrum Coefficients (MFCC) based on Empirical Mode Decomposition (EMD) was extracted and with its difference parameter to improve. At last they presented an effective method for speech emotion recognition based on Fuzzy Least Squares Support Vector Machines (FLSSVM) so as to realize the speech recognition of four main emotions. The experiment results showed that this method has the better anti-noise effect when compared with traditional Support Vector Machines. However, deep learning method that we used has been confirmed to have a robust performance when facing noisy speech.

Chi-Chun Lee et al. [10] 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, MFCC, etc. using the software SPSS and obtained a reduced feature set which was in a range of 40-60. The results show 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. The features they extracted was more than ours but not all the features are helpful.

Akshay S. Utane et al. [15] handled five emotion status (joy, sadness, surprise, anger and neutral status) using two different classifier, 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. The accuracy they achieved had a large range between various emotions, which represented that it was an unstable method.

In [8], the authors proposed a speech emotion recognition system using multi-algorithm, i.e., 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, which is also a large range.

5 Conclusions and Future Work

In this work, we investigate the utility of single feature and fusion features for speech emotion recognition. The results show that fusion features can improve the performance of DBN. However, the combination is 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. But the evaluation of DBN shows 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 multi-dimensional 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's 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 is the biggest challenge.

Acknowledgement. This research was supported by the International S&T Cooperation Program of China (ISTCP, 2013DFA10980).

References

1. Aher, P., Cheeran, A.: Auditory processing of speech signals for speech emotion recognition (2014)
2. Banse, R., Scherer, K.R.: Acoustic profiles in vocal emotion expression. J. Pers. Soc. Psychol. **70**(3), 614 (1996)
3. Chu, Y.Y., Xiong, W.H., Chen, W.: Speech emotion recognition based on EMD in noisy environments. In: Advanced Materials Research, vol. 831, pp. 460–464. Trans Tech Publication (2014)
4. El Ayadi, M., Kamel, M.S., Karray, F.: Survey on speech emotion recognition: features, classification schemes, and databases. Pattern Recogn. **44**(3), 572–587 (2011)
5. France, D.J., Shiavi, R.G., Silverman, S., Silverman, M., Wilkes, D.M.: Acoustical properties of speech as indicators of depression and suicidal risk. IEEE Trans. Biomed. Eng. **47**(7), 829–837 (2000)
6. Hinton, G.E.: Deep belief networks. Scholarpedia **4**(5), 5947 (2009)
7. Jin, Q., Li, C., Chen, S., Wu, H.: Speech emotion recognition with acoustic and lexical features. In: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4749–4753. IEEE (2015)
8. Joshi, D.D., Zalte, M.: Recognition of emotion from marathi speech using MFCC and DWT algorithms (2013)
9. Kim, Y., Lee, H., Provost, E.M.: Deep learning for robust feature generation in audiovisual emotion recognition. In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 3687–3691. IEEE (2013)
10. Lee, C.C., Mower, E., Busso, C., Lee, S., Narayanan, S.: Emotion recognition using a hierarchical binary decision tree approach. Speech Commun. **53**(9), 1162–1171 (2011)
11. Mohamed, A.R., Dahl, G.E., Hinton, G.: Acoustic modeling using deep belief networks. IEEE Trans. Audio Speech Lang. Process. **20**(1), 14–22 (2012)
12. Picard, R.W., Picard, R.: Affective Computing, vol. 252. MIT Press, Cambridge (1997)
13. Schuller, B., Rigoll, G., Lang, M.: Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2004), vol. 1, pp. I-577. IEEE (2004)
14. Shen, P., Changjun, Z., Chen, X.: Automatic speech emotion recognition using support vector machine. In: 2011 International Conference on Electronic and Mechanical Engineering and Information Technology (EMEIT), vol. 2, pp. 621–625. IEEE (2011)
15. Utane, A.S., Nalbalwar, S.: Emotion recognition through speech using gaussian mixture model and support vector machine. Emotion **2**, 8 (2013)
16. Williams, C.E., Stevens, K.N.: Emotions and speech: some acoustical correlates. J. Acoust. Soc. Am. **52**(4B), 1238–1250 (1972)
17. Wu, S., Falk, T.H., Chan, W.Y.: Automatic speech emotion recognition using modulation spectral features. Speech Commun. **53**(5), 768–785 (2011)