Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2014, Article ID 749604, 7 pages
http://dx.doi.org/10.1155/2014/749604

Research Article

A Research of Speech Emotion Recognition Based on Deep Belief Network and SVM

Chenchen Huang, Wei Gong, Wenlong Fu, and Dongyu Feng

Department of Computer, Communication University of China, Beijing 100024, China

Correspondence should be addressed to Chenchen Huang; hcc.1990@163.com

Received 27 May 2014; Revised 21 July 2014; Accepted 21 July 2014; Published 12 August 2014

Academic Editor: Stefan Balint

Feature extraction is a very important part in speech emotion recognition, and in allusion to feature extraction in speech emotion recognition problems, this paper proposed a new method of feature extraction, using DBNs in DNN to extract emotional features in speech signal automatically. By training a 5 layers depth DBNs, to extract speech emotion feature and incorporate multiple consecutive frames to form a high dimensional feature. The features after training in DBNs were the input of nonlinear SVM classifier, and finally speech emotion recognition multiple classifier system was achieved. The speech emotion recognition rate of the system reached 86.5%, which was 7% higher than the original method.

1. Introduction

Speech emotion recognition was a technology that extract emotional feature from speech signals by computer and contrasts and analyses the characteristic parameters and the emotional change acquired. Finally, the law of speech and emotion was concluded and speech emotional states were judged according to the law. At present, speech emotion recognition was an emerging crossing field of artificial intelligence and artificial psychology; besides, it was a hot research topic of signal processing and pattern recognition [1]. The research was widely applied in human-computer interaction, interactive teaching, entertainment, security fields, and so on.

Speech emotion processing and recognition system was generally composed of three parts, which were speech signal acquisition, feature extraction, and emotion recognition. System framework is shown in Figure 1.

In this system, the quality of feature extraction directly affected the accuracy of speech emotion recognition. In the process of feature extraction, it usually took the whole emotion sentence as units for feature extracting, and extraction contents were four aspects of emotion speech, which were several acoustic characteristics of time construction, amplitude construction, fundamental frequency construction, and formant construction. Then contrast emotion speech with no emotion sentence from these four aspects, acquiring the law of emotional signal distribution, then classify emotion speech according to the law [2].

Deep neural network (DNN) has unprecedented success in the field of speech recognition and image recognition [3]; however, so far no research on deep neural network has been applied to speech emotion processing. We found that the deep belief network (DBN) of DNN in speech emotion processing has a huge advantage [4]. Therefore, this paper proposed a method to realize the emotional features automatically extracted from the sentence. It used DBNs to train a 5-layer-deep network to extract speech emotion features. It incorporates the speech emotion features of more consecutive frames, to build a high latitude characteristic, and uses SVM classifier to classify the emotional speech. We compared other traditional feature extraction methods with this method and concluded that the speech emotion recognition rate reached 86.5%, which was 7% higher than the original method.

2. Feature Extraction of Speech Emotion

Emotion can be expressed by speech because speech contains the characteristic parameters that can reflect emotion information [5]. We can extract and observe the change

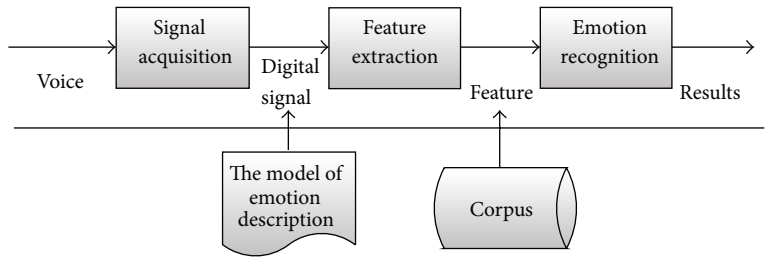

Figure 1: Speech emotion recognition system block diagram.

of characteristic parameters to measure the corresponding speech emotional changes. The key above is extracting characteristic parameters of speech emotion from speech signals. The quality of feature extraction directly affects the accuracy of speech emotion recognition. Meanwhile, speech signals contain not only emotional feature information but also the speaker's own important information, therefore research on how to extract and which speech emotion characteristic parameters to extract are of great importance [6].

2.1. Emotion Speech Database. Before the speech emotion feature extraction, at first, we need to input emotional speech signal. Emotion speech library is the foundation of speech emotion recognition, which provides standard speech for speech emotion recognition.

At present, there is much literature on this research aspect [7], and throughout the world English, German, Spanish, and Chinese single language emotion speech databases have been built. A few speech libraries also contain a variety of languages. This paper is aimed at using the language of Chinese to recognize speech emotion. In order to establish a perfect speech data sampling library, when selecting experimental sentence we need to follow the following principles.

(i) The sentence selected must not contain a particular aspect of emotional tendency, ensuring that the recorded statements will not affect the experimenter's judgment [8].

(ii) The sentence selected has relatively emotional freedom. That is, the sentence can express different emotions, not just a single emotion; otherwise we will be unable to compare the emotional speech parameters in the same emotional sentence under different emotional states [9].

According to the above principles and considering all aspects, to ensure the accuracy in this paper, we selected Buaa emotional speech database, which passed the effectiveness measurement, instead of recording speech database by ourselves. This database was recorded by 7 males and 8 females and it consisted of 7 kinds of basic emotions, which are sadness, anger, surprise, fear, joy, hate, and calm. Each emotion had 20 sentences referential scripts. That is to say, the database consisted of 2100 emotion sentences, and all the emotion sentences above were recorded into WAV format files. The sampling rate of speech above was 16000 Hz all, and the quantitative accuracy was 16 bits. This paper selected 1200 sentences which contain sadness, anger, surprise, and happiness, four basic emotions for training and recognition.

2.2. Traditional Emotional Feature Extraction. Traditional emotional feature extraction was based on the analysis and comparison of all kinds of emotion characteristic parameters, selecting emotional characteristics with high emotional resolution for feature extraction. In general, traditional emotional feature extraction concentrates on the analysis of the emotional features in the speech from time construction, amplitude construction, and fundamental frequency construction and signal feature [10].

2.2.1. Time Construction. Speech time construction refers to the emotion speech pronunciation differences in time. When people express different feelings, the time construction of the speech is different. Mainly in two aspects, one is the length of continuous pronunciation time, the other is the average rate of pronunciation. One is the length of continuous pronunciation time and the other is the average rate of pronunciation.

Zhao Li's research showed that different emotional pronunciations are different in pronunciation length and pronunciation speed. Compared with the length of calm pronunciation time, the pronunciation time of joy, anger, and surprise significantly shortened. But comparing with calm pronunciation time, sad pronunciation length is longer. Compared with quiet pronunciation rate, sad pronunciation rate is slow, while joy, anger, and surprise pronunciation rates are quick relatively.

In conclusion, if we extract the time construction characteristics parameters of the speech, it is easy to distinguish sad emotion from other emotional states. Of course, we also can set a certain time threshold to distinguish joy, anger, surprise, and speech. However it is obvious that using only speech time construction is not enough to recognize speech emotional state.

2.2.2. Amplitude Construction. Speech signal amplitude construction and speech emotional state also have a direct link. When the speaker is angry or happy, the volume of speech is generally high. When speaker is sad or depressed, the volume of speech is generally low. Therefore, the analysis of speech emotion features of amplitude construction is more meaningful.

Figure 2 is the comparison of emotional speech and calm speech, which is shown with the average amplitude difference.

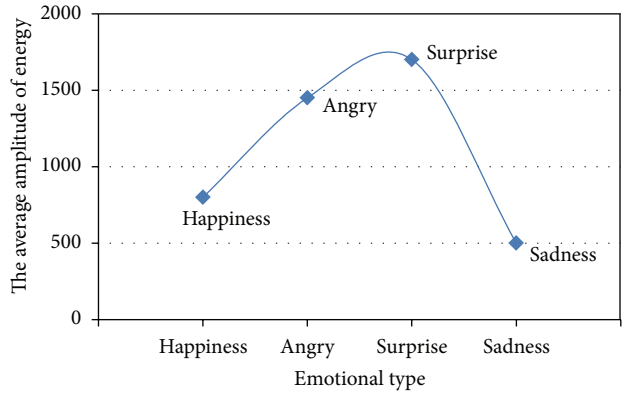

Figure 2: The distribution of emotional speech amplitude energy.

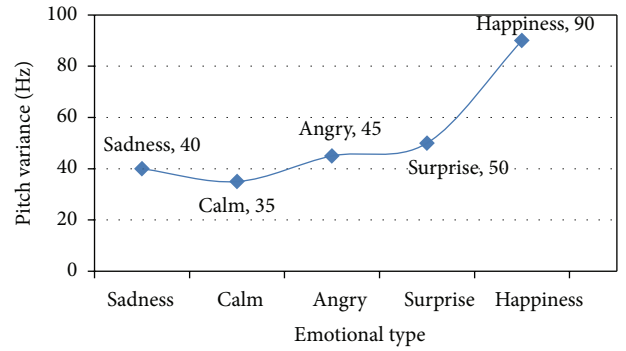

Figure 3: Curves of different emotional variance.

Figure 2 shows that the amplitude of joy, anger, and surprise, three kinds of emotional speech, compared with calm voice signal is larger, while the sad speech amplitude is smaller.

2.2.3. Fundamental Frequency Construction. Bänziger and Scherer [11] proposed that, for the same sentence, if the emotions expressed were different, fundamental frequency curves were also different; besides the mean and variance of fundamental frequency were also different. When the speaker is in a state of happiness, the fundamental frequency curve of speech generally is bent upwards. And when the speaker is in a state of sadness, the fundamental frequency curve of speech generally is bent downward. Figure 3 shows the curves of different emotional variance.

Compared to calm emotional state, happiness, surprise, and angry speech signal characteristics variation is larger. Thus, analyzing fundamental frequency curve of the same sentence under different emotional states, we can contrast and acquire fundamental frequency construction of different emotion speech.

3. The Depth of the Belief Network

Deep neural network stems from artificial neural network [12]. Deep neural network is literally a deep neural network. In 2006, Professor Hinton in the University of Toronto presented a deep belief network (DBN) structure in [13]. Since then, deep neural network and deep learning have become the most popular hot spot research in artificial intelligence. He clarified the effectiveness of unsupervised learning and training learning at each layer and pointed out that each layer can conduct unsupervised train again on the basis of the results output from previous layer training. Compared with the traditional neural network, deep neural network has deep structure with multiple nonlinear mapping which can complete complex function approximation [14].

Hinton first put forward DBNs in 2006. Since then DBNs have got unprecedented success in areas such as speech recognition and image recognition. Microsoft researcher Dr. Deng cooperated with Hilton found deep neural network can significantly improve the accuracy of speech recognition; however; so far no research on deep neural network has been applied to speech emotion recognition. In this paper, research found that deep belief networks (DBNs) have a great advantage in speech emotion recognition; therefore, we choose deep belief networks to extract emotional feature automatically in the speech.

A typical deep belief network is a highly complex directed acyclic graph, which is formed by a series of restricted Boltzmann machine (RBM) stacks. Training DBNs is realized by training RBMs layer by layer from bottom to up. Because RBM can be trained rapidly via layered contrast divergence algorithm, training RBM can avoid a high degree of complexity of training DBNs, simplifying the process to training each RBM. Numerous studies have demonstrated that deep belief network can solve low convergence speed and local optimum problems in traditional backpropagation algorithm training multilayer neural network.

3.1. Restricted Boltzmann Machine. DBNs are stacked regularly by restricted Boltzmann machine (RBM); RBM is a kind of typical neural network. RBM is composed via the connection of visible layer and hidden layer, but there is no connection between visible layer, visible layer and hidden layer, and hidden layer. In Figure 4, training RBM used unsupervised greedy method step by step. That is, in training, the characteristic value of the visible layer maps to the hidden layer, then visible layer can be reconstructed through the hidden layer; this new layer visible characteristic value maps to the hidden layer again, then acquiring a new hidden layer. Its main purpose is to obtain the generative power value. Thus, the main characteristics of RBM are the activation features of the layer inputted to next layer as training data, and as a consequence the learning speed is fast [15]. This is a layer by layer and efficient learning strategy theory, and proof of procedure is in [16].

In Figure 4, DBNs are stacked from bottom to up by RBM. Using Gauss-Bernoulli RBM and Bernoulli- Bernoulli RBM to connect, the lower layer output is the input for the upper layer.

Figure 5 is the structure diagram of DBNs. The number of layer and unit is an example, and the number of hidden layer is not necessarily the same in actual experiment.

In Bernoulli RBM, visible and hidden layer units are binary: $V \in \{0,1\}^D$ and $h \in \{0,1\}^K$. D and K present unit numbers of visible and hidden layers. In Gaussian RBM,

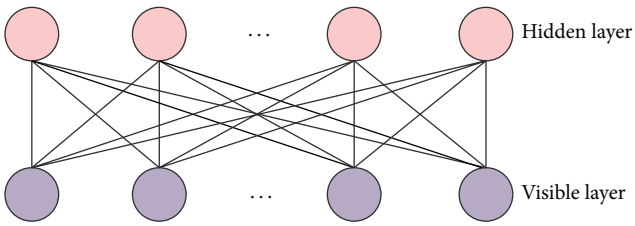

Figure 4: RBM module.

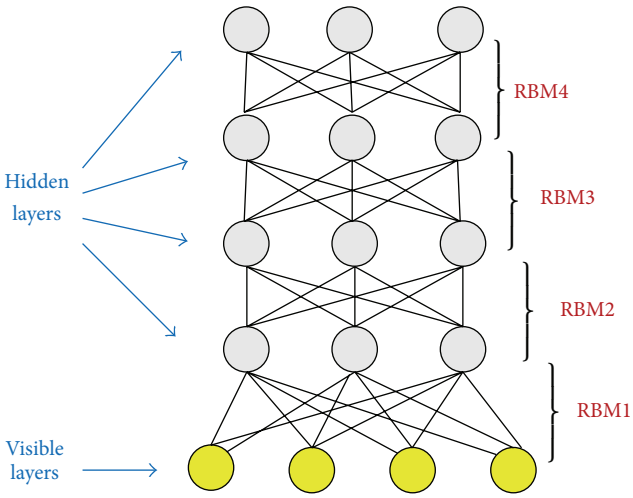

Figure 5: DBNs structure sketch map.

visible layer unit is a real number: $V \in R^D$. V and h joint probability is expressed as

$$P(v,h) = \frac{1}{Z}\exp(-E(v,h)). \quad (1)$$

Z is a normalized constant and $E(v,h)$ is an energy equation. For Bernoulli RBM, energy equation is

$$E(v,h) = -\sum_{i=1}^{D}\sum_{j=1}^{K} w_{ij}v_i h_j - \sum_{i=1}^{D} b_i v_i - \sum_{j=1}^{K} a_j h_j. \quad (2)$$

w_{ij} presents the quality of unconnected visible layer nodes v_i and hidden layer nodes h_j and a and b are implicit bias of visible and hidden units. For Gaussian RBM, energy equation is

$$E(v,h) = \sum_{i=1}^{D}\frac{(v_i - b_i)^2}{2} - \sum_{i=1}^{D}\sum_{j=1}^{K} w_{ij}v_i h_j - \sum_{j=1}^{K} a_j h_j. \quad (3)$$

DBNs combined emotional speech signal features of continuous frames, forming a high dimensional feature vector, fully describing the correlation between the emotional speech features. Meanwhile, DBNs use these high dimensional features to simulate [17]. Besides, DBNs extracting speech information process is similar to brain processing speech, using RBM to extract emotional information layer by layer, eventually acquiring the most suitable high dimensional characteristics for pattern recognition. DBNs can combine well with traditional speech emotion recognition technology in practical application (e.g., SVM), improving the accuracy of speech emotion recognition.

Table 1: Hyperparameters and training statistics of the chosen DBN.

Number of hidden layers	5
Units per layer	50
Unsupervised learning rate	0.001
Supervised learning rate	0.01
Number of unsupervised epochs	50
Number of supervised epochs	475
Total training time (hours)	136
Classification accuracy	0.865

3.2. DBNs Model Training. We have used Theano to train the deep belief networks (DBNs) in this paper. Theano is a mathematical symbols compilation kit in Python, which is extremely a powerful tool in machine learning, because it combines the power of Python and C, making it easier to establish deep learning model.

The DBNs were first pretrained with the training set in an unsupervised manner [18]. We split the Buaa dataset and used 40% of the voice data for training and 60% of the voice data for testing. Then we proceeded to the supervised fine tuning using the same training set and used the validation set to do early stopping [19].

We tried approximately 100 different hyperparameters combinations and selected the model that has the smallest error rate. The selected DBN model has been shown in Table 1.

We fixed the DBN architecture for all experiments and used 5 hidden layers (the first layer is a Gaussian RBM; all other layers are Bernoulli RBMs) and 50 units per layer, and the only variable is size of input vector, which depends on the context window length. The hyperparameters for generative pretraining are shown in Table 1. The unsupervised layers ran for 50 epochs with 0.001 learning rate. Supervised layers ran for 475 epochs with 0.01 learning rate.

4. Support Vector Machine Classifier

SVM is a kind of machine learning methods based on statistical theory and structural risk minimization principle. Its principle is to map the low dimensional feature vector to high-dimensional feature vector space, so as to solve nonlinear separable problem. SVM has been widely used in pattern classification field [20].

The key to solve nonlinear separable problem is to construct the optimal separating hyperplane; the construction of the optimal hyperplane eventually translates into the calculation of optimal weights and bias. We set the training sample set for $\{(X^1,d^1)(X^2,d^2)\cdots(X^P,d^P)\cdots(X^P,d^P)\}$; the cost function of minimization weight W and slack variable ε_p is

$$\Phi(W,\varepsilon) = \frac{1}{2}W^T W + C\sum_{P=1}^{P}\varepsilon_P. \quad (4)$$

Conditions for its limitation. $d^p(W^T W^P + b) \geq 1 - \varepsilon_P$, $p = 1, 2, \ldots, p$, using to measure the deviation degree of a sample points relative to linear separable ideal conditions. Due to the sample set, we can calculate the dual to know the optimal weight and bias. To solve quadratic optimization problem, when optimal hyperplane in the feature space is constructed, the optimal hyperplane can be defined as the following function:

$$g(x) = \sum_{i=1}^{Q} u_i a_i^* K\left(X, X_i^*\right) + b^*, \quad i = 1, 2, 3, \ldots, Q. \tag{5}$$

In the formula above, $K(X, X_i^*)$ is kernel function, X_i^* is support vector of nonzero Lagrange multiplier a_i^*, Q is the number of support vectors, and b^* is bias parameter. For nonlinear SVM, nonlinear mapping kernel function mapping data to higher-order feature space, optimal hyperplane exists in this space. Several kinds of kernel function are applied to nonlinear SVM, such as Gaussian kernel function and polynomial kernel function. This paper takes the following Gaussian kernel function to do the research:

$$K(x, y) = \exp\left(-\gamma |x - y|^2\right). \tag{6}$$

In this function, γ is a Gaussian transmission coefficient.

When we use support vector machine (SVM) to solve the problem of classification, there are two solutions: one-to-all/one-to-one. In previous studies we found that the accuracy of one-to-one way of classification is higher. Therefore, this paper chose the one-to-one way of classification to deal with four kinds of emotion (surprise, joy, angry, and sadness).

"One-to-one" mode is to construct hyper-plane for every two emotions, so the number of child classifiers we need to be trained is $k * (k-1)/2$ [21]. In this experiment and during the whole process of training, the number of SVM classifier we need is C_4^2, namely 6. Each child classifier was trained by surprise, joy, anger, sadness, or any of these two kinds of emotion, namely, joy-anger, joy-sadness, joy-surprise, anger-sadness, anger-surprise, and sadness-amazement.

Training a classifier for any two categories, when classifying an unknown speech emotion, each classifier judges its category and votes for the corresponding category. Finally take the category with most votes as the category of unknown emotional. Decision stage uses voting method, it is possible that votes tied for multiple categories, leading to the unknown samples belonging to different categories at the same time, therefore affecting the classification accuracy.

SVM classifier defines a label for each speech emotion signal before training and recognition, to indicate speech emotional signal's emotional categories. Label type must be set to double. In the process of emotion recognition, input feature vector into all SVMs; then, the output of each SVM passes through logic judgment to choose the most probable emotion. Finally, the emotion with the highest weight (most votes) is the emotional state of speech signal. The setting of penalty factor C in classifier training and γ in kernel function can be determined via cross-validation of training set. One thing to note here is that C and γ are fit for recognition effect of training set but do not necessarily also fit for test set. After repeated testing, in the experiments of speech emotion recognition in this paper, the magnitude of Gaussian transmission coefficient γ SVM was set to 10^{-3} and the magnitude of parameter C is 10^6. These parameters will change according to the experiment and the error rate, to improve the accuracy rate of training data classification.

DBNs proposed multidimensional feature vector as the input of SVM. Since the SVM does not scale well with large datasets, we subsampled the training set by randomly picking 10,000 frames. For nonlinear separable problem in speech emotion recognition, take kernel function to map input characteristics sample points to high-dimensional feature space, making the corresponding sample space linear separable. In simple terms it creates a classification hyperplane as decision surface, making the edge isolation of positive and negative case maximum. Evaluating decision function calculation is still in the original space, making less computational complexity of the high dimensional feature space after mapping.

Figure 6 is the block diagram for the emotion recognition based on SVM system.

5. The Experiment and Analysis

This paper selected 1200 sentences which contain sadness, anger, surprise, and happiness, four basic emotions for training and recognition.

Before inputting the emotional speech signal to the model, we need to preprocess this signal. In this paper, the speech signal to be inputted first goes through pre-emphasis and window treatments. We selected median filter that has the window of the length of 5 to do smoothing processing for denoised emotional speech signal.

This paper used 40% of the voice data for training and 60% of the voice data for testing. The experimental group is speech emotion recognition model established by deep belief network via extracting phonetic characteristics. The control group is established by extracting traditional speech feature parameters to the speech emotion recognition model under the condition of the same emotional speech input. At last, we contrasted and analyzed the experimental data for the conclusion. The process of the experiment is shown in Figure 7.

5.1. The Experimental Group: Research of Speech Emotion Recognition Based on Multiple Classifier Models of the Deep Belief Network and SVM. This paper proposed a method to realize the emotional features extracted automatically from the sentence. It used DBNs to train a 5-layer-deep network to extract speech emotion features. In this experiment, we put the output of the DBNs hidden layer as a characteristic training regression model and the finished training characteristics as the input of a nonlinear support vector machine (SVM). Eventually we set up a multiple classifier model of speech emotion recognition. It would also be possible to train our DBN directly to do classification. However our goal is to compare the DBN learned representation with other representations. By using a single classifier we were able to carry out direct comparisons. The experimental results are shown in Table 2.

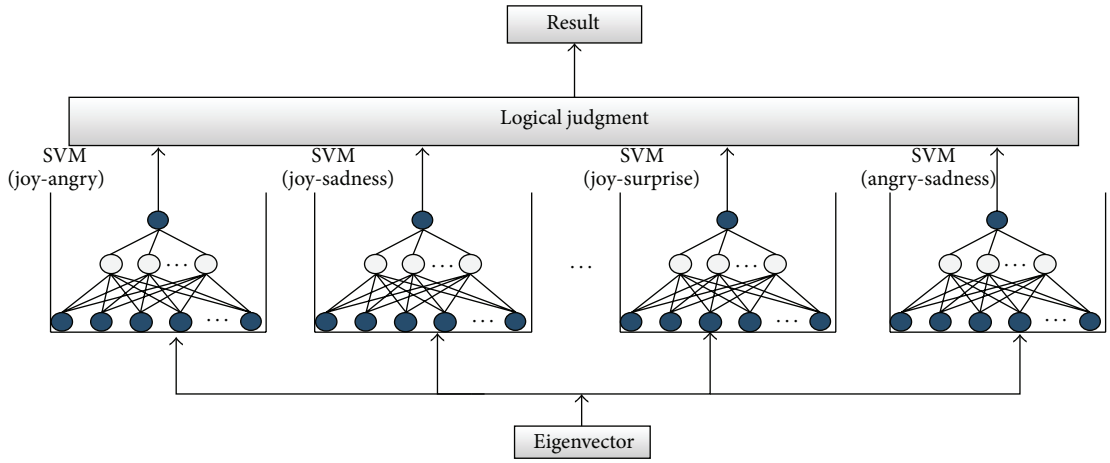

Figure 6: Diagram of the emotion recognition system based on SVM.

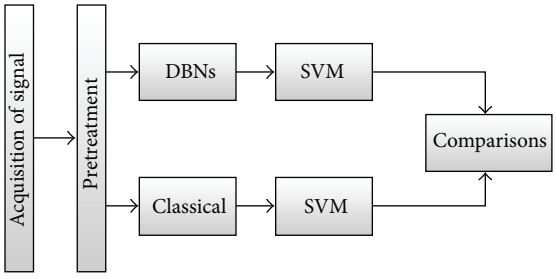

Figure 7: The process of the experiment.

Table 2: The result of recognition rate based on DBNs and SVM.

Test	Train			
	Angry	Joy	Sadness	Surprise
Angry	0.913	0.021	0.021	0.028
Joy	0.031	0.845	0.041	0.122
Sadness	0.029	0.020	0.883	0.061
Surprise	0.047	0.124	0.035	0.819

As it was shown in Table 1, anger and sadness have a higher recognition rate than joy and surprise; they reached 91.3% and 88.3%, respectively, and the overall recognition rate was 86.5%.

5.2. The Control Group: Research of Speech Emotion Recognition Based on Extracting the Traditional Emotional Characteristic Parameters and SVM. In this paper, the control group was established by extraction traditional emotional characteristic parameters: time construction, amplitude construction, and fundamental frequency construction. After extracting them we also need to input the emotional characteristic parameters into the SVM classification of speech emotion recognition. Finally, we compared the difference of experimental group and control group. The experimental results were shown in Table 2.

Table 3: The result of recognition rate based on extraction of traditional emotional characteristic parameters and SVM.

Test	Train			
	Angry	Joy	Sadness	Surprise
Angry	0.861	0.032	0.023	0.088
Joy	0.039	0.742	0.034	0.182
Sadness	0.047	0.034	0.848	0.071
Surprise	0.053	0.174	0.037	0.729

Table 4: Comparison of results of the two methods.

	Angry	Joy	Sadness	Surprise	Average
DBNs	0.913	0.845	0.883	0.819	0.865
Classical	0.861	0.742	0.848	0.729	0.795

As it was shown in Table 3, overall recognition rate based on the traditional emotional characteristic parameters of the SVM system was 79.5%.

As it was shown in Table 4 and Figure 8, compared with the two methods, the method using DBNs to extract speech emotion features improved these four types of (sadness/joy/anger/surprise) emotion recognition rate. The average rate was increasing by 7%, and the recognition rate of joy has increased by 10%.

6. Conclusion

In this paper we proposed a method that used the deep belief networks (DBNs), one of the deep neural networks, to extract the emotional characteristic parameter from emotional speech signal automatically. We combined deep belief network and support vector machine (SVM) and proposed a classifier model which is based on deep belief networks (DBNs) and support vector machine (SVM). In the practical training process, the model has small complexity and 7% higher final recognition rate than traditional artificial extract,

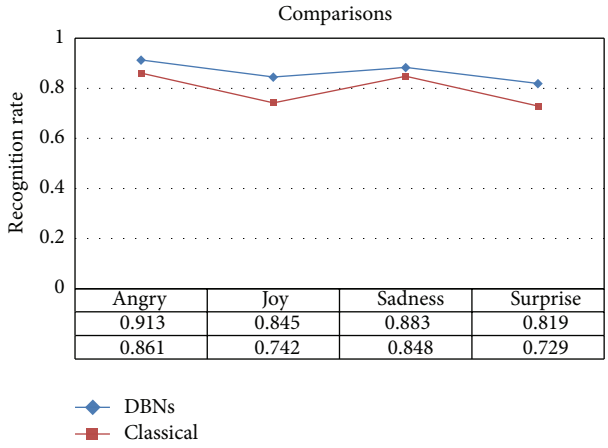

FIGURE 8: Comparison of results of the two methods.

and this method can extract emotion characteristic parameters accurately, improving the recognition rate of emotional speech recognition obviously. But the time cost for training DBNs feature extraction model was 136 hours, and it was longer than other feature extraction methods.

In future work, we will continue to further study speech emotion recognition based on DBNs and further expand the training data set. Our ultimate aim is to study how to improve the recognition rate of speech emotion recognition.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

The authors would like to thank the National Key Science and Technology Pillar Program of China (the key technology research and system of stage design and dress rehearsal, 2012BAH38F05; study on the key technology research of the aggregation, marketing, production and broadcasting of online music resources, 2013BAH66F02; Research of Speech Emotion Recognition Based on Deep Belief Network and SVM, 3132013XNG1442).

References

[1] Z. Yongzhao and C. Peng, "Research and implementation of emotional feature extraction and recognition in speech signal," *Joural of Jiangsu University*, vol. 26, no. 1, pp. 72–75, 2005.

[2] L. Zhao, C. Jiang, C. Zou, and Z. Wu, "Study on emotional feature analysis and recognition in speech," *Acta Electronica Sinica*, vol. 32, no. 4, pp. 606–609, 2004.

[3] H. Lee, C. Ekanadham, and A. Y. Ng, "Sparse deep belief net model for visual area V2," in *Proceedings of the 21st Annual Conference on Neural Information Processing Systems (NIPS '07)*, MIT Press, December 2007.

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proceedings of the 26th Annual Conference on Neural Information Processing Systems (NIPS '12)*, pp. 1097–1105, Lake Tahoe, Nev, USA, December 2012.

[5] Z. Li, "A study on emotional feature analysis and recognition in speech signal," *Journal of China Institute of Communications*, vol. 21, no. 10, pp. 18–24, 2000.

[6] T. L. Nwe, S. W. Foo, and L. C. de Silva, "Speech emotion recognition using hidden Markov models," *Speech Communication*, vol. 41, no. 4, pp. 603–623, 2003.

[7] X. Bo, "Analysis of mandarin emotional speech database and statistical prosodic features," in *Proceedings of the Interntional Conference on Affective Computing and Intelligent Interaction (ACLL '03)*, pp. 221–225, 2003.

[8] L. Zhao, X. Qian, C. Zhou, and Z. Wu, "Study on emotional feature derived from speech signal," *Journal of Data Acquistion & Processing*, vol. 15, no. 1, pp. 120–123, 2000.

[9] G. Pengjuan and J. Dongmei, "Research on emotional speech recognition based on pitch," *Application Research of Computers*, vol. 24, no. 10, pp. 101–103, 2007.

[10] P. Guo, *Research of the Method of Speech Emotion Feature Extraction and the Emotion Recognition*, Northwestern Polytechnical University, 2007.

[11] T. Bänziger and K. R. Scherer, "The role of intonation in emotional expressions," *Speech Communication*, vol. 46, no. 3-4, pp. 252–267, 2005.

[12] R. Rui and M. Zhenjiang, "Emotional speech synthesis based on PSOLA algorithm," *Journal of System Simulation*, vol. 20, pp. 423–426, 2008.

[13] S. Zhijun, X. Lei, X. Yangming, and W. Zheng, "Overview of deep learning," *Application Research of Computers*, vol. 29, no. 8, pp. 2806–2810, 2012.

[14] G. E. Hinton, S. Osindero, and Y. Teh, "A fast learning algorithm for deep belief nets," *Neural Computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

[15] Z. Sun, L. Xue, Y. Xu, and Z. Wang, "Overview of deep learning," *Application Research of Computers*, vol. 29, no. 8, pp. 2806–2810, 2012.

[16] Z. Chunxia, J. Nannan, and W. Guanwei, "Introduction of restricted boltzmann machine," China Science and Technology Papers Online, http://www.paper.edu.cn/releasepaper/content/201301-528.

[17] T. Shimmura, "Analyzing prosodic components of normal speech and emotive speech," *The Preprint of the Acoustical Society of Japan*, pp. 3–18, 1995.

[18] X. Kai, J. Lei, C. Yuqiang, and X. Wei, "Deep learning: yesterday, today, and tomorrow," *Journal of Computer Research and Development*, vol. 50, no. 9, pp. 1799–1804, 2013.

[19] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle, "Greedy layer-wise training of deep networks," in *Advances in Neural Information Processing Systems 19*, pp. 153–160, MIT Press, 2007.

[20] Y. Kim, H. Lee, and E. M. Provost, "Deep learning for robust feature generation in audio-visual emotion recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '13)*, Vancouver, Canada, 2013.

[21] J. Zhu, X. Wu, and Z. Lv, "Speech emotion recognition algorithm based on SVM," *Computer Systems & Applications*, vol. 20, no. 5, pp. 87–91, 2011.