Speech Emotion Recognition Based on Deep Belief Networks and Wavelet Packet Cepstral Coefficients

Yongming Huang [1,2], Ao Wu [1,2], Guobao Zhang [1,2] and Yue Li[1,2]

[1.] *School of Automation*, Southeast University, Nanjing 210096, China
[2.] Key Laboratory of Measurement and Control of Complex Systems of Engineering,
Ministry of Education

***Abstract* — A wavelet packet based adaptive filter-bank construction combined with Deep Belief Network(DBN) feature learning method is proposed for speech signal processing in this paper. On this basis, a set of acoustic features are extracted for speech emotion recognition, namely Coiflet Wavelet Packet Cepstral Coefficients (CWPCC). CWPCC extends the conventional Mel-Frequency Cepstral Coefficients (MFCC) by adapting the filter-bank structure according to the decision task. And Deep Belief Networks (DBNs) are artificial neural networks having more than one hidden layer, which are first pre-trained layer by layer and then fine-tuned using back propagation algorithm. The well-trained deep neural networks are capable of modeling complex and non-linear features of input training data and can better predict the probability distribution over classification labels. Speech emotion recognition system is constructed with the feature set, DBNs feature learning structure and Support Vector Machine as classifier. Experimental results on Berlin emotional speech database show that the Coiflet Wavelet Packet is more suitable in speech emotion recognition than other acoustics features and proposed DBNs feature learning structure combined with CWPCC improve emotion recognition performance over the conventional emotion recognition method.**

Keywords - Speech emotion recognition, Coiflet Wavelet packets Cepstral Coefficients (CWPCC), Acoustic features, Deep Belief Networks (DBNs), Deep learning

I. INTRODUCTION

Speech emotion recognition is well used in all kinds of applications where natural human-computer interaction is needed. For example, in the design of interactive films and computer games [1], in call-centers [2, 3], in intelligent automobile systems[4] and in health-care service to help diagnosing depression and suicide risk [5].

As we all know, wavelet packet (WP) is efficient in providing flexible and adaptive frequency band division methods[6], and is a prominent technique for quasi-periodic and non-stationary signal processing, such as speech processing. In this paper, the Coiflet WP-based acoustic features are proposed to combine with deep learning for speech emotion classification.

Recently, the applications of DBN or Deep Learning (DL) make great breakthroughs in lots of difference fields [7]. DBN represents a series of multi-layer architecture that training with the greedy layer-wise unsupervised pre-training algorithms[8], [9]. DBN can reconstruct the raw data set by the greedy layer-wise unsupervised pre-training mechanism. And the intelligent models, like classifiers usually can achieve better recognition performance with the learned features.

In this paper, we tried to find out the best DBNs feature learning model parameters based on the Support Vector Machine. And then we tested three different kinds of features, CWPCC, MFCC, PLP and there combinations to find out whether the performance could benefit from employing deep learning.

II. WAVELET PACKET TRANSFORM

In this paper, instead of the short-time Fourier transform which is widely used, WPT is adopted for analysing the speech signal. The reason is introduced in Section 1. The wavelet packet transform method is introduced in detail in our previous work[10].

III. DEEP BELIEF NETWORK

The Restricted Boltzmann Machine(RBM) is a two-layer structure, which is constructed by a visible layer and a hidden layer. An illustration of RBM architecture is shown in Fig.1. In Fig.1, we can see the standard type of RBM has binary-valued m hidden and n visible neurons, and consists of a matrix of weights $W = (w_{i,j})$(size $m \times n$) associated with the connection between hidden neurons h_j and visible neuron v_i, as well as bias weights a_i for the visible units and b_j for the hidden units.

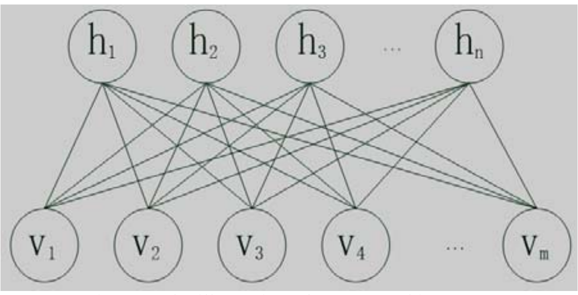

Figure.1 The illustration of RBM architecture

Given these, the energy function of a configuration (v,h) is defined as:

$$E(v,h) = -\sum_i a_i v_i - \sum_j a_j v_j - \sum_i \sum_j v_i w_{ij} h_j \quad (1)$$

Because of the binary-valued neurons, the probabilities of the states of the visible and hidden neurons can be calculated by the following function

$$p_{vi} = p(vi = 1) = \frac{1}{1 + \exp(-\sum_i w_{ij} h_j)} \quad (2)$$

$$p_{hj} = p(hj = 1) = \frac{1}{1 + \exp(-\sum_i w_{ij} h_j)} \quad (3)$$

In RBM, the probability distributions of hidden and visible vectors can be defined as

$$P(v) = \frac{1}{Z} \sum_h e^{E(v,h)} \quad (4)$$

The RBMs are trained to maximize the product of probabilities of training dataset

$$\arg\max_W \prod_{v \in V} P(v) \quad (5)$$

In each iteration, we update the w_{ij} according MCD rule[12, 13] by

$$\Delta w_{ij} = \varepsilon \left(v \cdot h^T - \hat{v} \cdot \hat{h}^T \right) \quad (6)$$

where $\hat{v}, \hat{h}$ is the reconstructed states of the node in the last iteration.

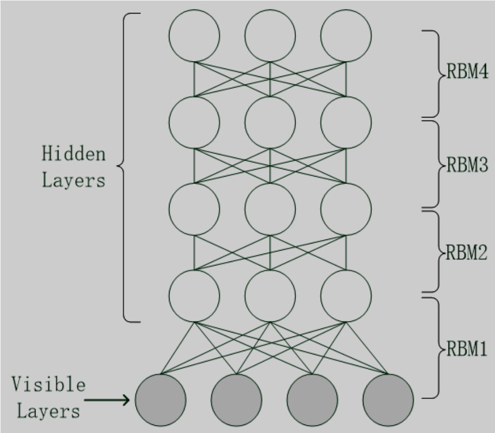

Figure.2 Schematic diagram of DBNs

As shown in Figure.2, deep Belief Networks (DBNs) are stacked up by many layers of restricted Boltzmann machines, in which latent units are typically allocated with binary values randomly. The DBNs combine the acoustic features to a high-dimensional feature, which describe the relationships between speech emotion features. Besides, DBNs is good at learning relationships between features in high-dimensional space.

IV. PROPOSED SYSTEM

In this section, we describe details of the proposed speech emotion recognition system with emphasis on the WP filter-bank based acoustic feature extraction. The following subsections give detailed description of each part of the proposed system. Framework of the proposed speech emotion recognition system is shown in Fig.3

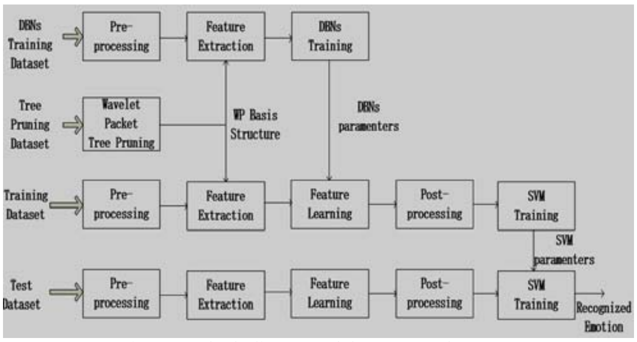

Figure.1 Block diagram of the proposed system

A. *Dataset Division*

The whole emotional speech database is divided into four parts. The DBNs training dataset tree-pruning dataset, training dataset and test dataset are used for DBNs training, WP tree pruning, classifier training and emotion recognition respectively, and are not overlapped with each other.

B. *Wavelet Packet Filter-Bank Construction*

An optimal filter-bank structure is obtained using the tree-pruning algorithm. The optimal WP filter-bank structure is then applied on the training and test samples to calculate WP-based acoustic features. With the fast tree-pruning algorithm, a sequence of WP admissible trees with different number of leaf nodes is obtained, and correspondingly the set of filter-bank structures with different number of sub-bands. The obtained WP filter-bank structures are then used to calculate emotion-discriminative acoustic features from original speech signal.

C. *Pre-processing*

Before feature extraction, conventional speech signal processing operations including pre-emphasis, frame blocking and windowing are performed first. The speech signal is first pre-emphasized by a high-pass FIR filter

$1-0.9375z^{-1}$ to spectrally flatten the signal and make it less susceptible to finite precision effects later in the signal processing [10]. The pre-emphasized speech signal is then blocked into frames of K samples with an overlap of $K^{'}$ between adjacent frames. Here we use $K=256$ and $K^{'}=K/2$. And each individual frame is multiplied by a Hamming window to reduce ripples in the spectrum. The pre-processing is the same as we did in our previous work[10].

D. Feature Extraction

A type of wavelet packet features called the Coiflet Wavelet Packet Cepstral Coefficients (CWPCC), is extracted in a similar way as the conventional MFCC features, for the speech emotion recognition task.

Firstly, the windowed speech frame is passed through a filter-bank obtained from WP tree pruning step. Secondly, we calculated the sub-band log-energy coefficients and frequency ordered them. Thirdly we de-correlated the log filter-bank energy by the Discrete Cosine Transform (DCT). Finally, we adopted the first DCT coefficients together with the log-energy of the frame and their first and second order derivatives to constitute the feature vector.

E. Feature Extraction

In the pre-training process of neural networks, we set the learning rate, the mini-batch size and the weight cost to be 0.01, 256 and 0.0001, respectively. In the training process of one RBM, the range of momentum was 0.5 to 1.0. To ensure the output probabilities sum up to 1.0 after using back propagation algorithm to train the network, we added a softmax regression to the output layer.

We selected 15 per cent of the training data as cross validation set for the parameter tuning process. At first, the mini-batch size was set to 128 and learning rate was set to 0.8. We evaluated the performance of our system on the cross validation set after every iteration. The learning rate was halved for the next iteration when the performance did not achieve great improvements.

F. Classifier

We adopted Support vector machine (SVM) as the speech emotion classification in this paper. The implementation of the SVM classifier is provided by a publicly available Matlab toolbox named LIBSVM Matlab Toolbox [13].

V. EXPERIMENT

A. Emotional Speech Database and Experimental Setup

The proposed speech emotion recognition system is evaluated on the Berlin emotional speech database [14], which contains 7 simulated emotions (anger, boredom, disgust, fear, joy, neutral and sadness). In this paper, six emotions (no disgust) with a sum of 489 utterances are used for the classification task. The disgust emotion is discarded because the number of disgust utterances is fairly limited. 20% of the database is randomly selected to form the tree-pruning dataset, 20% of the database is randomly selected to train the DBNs structure and we apply 5-fold cross validation on the remaining 60% utterances to assess the classification performance.

B. Experimental Results

To evaluate performance of the extracted feature, a set of experiments are conducted and the results are presented in our previous work[10]. WP filter-bank structures generated by coif 3 achieved the highest accuracy rate in speech emotion recognition compared with other wavelet packets.

It is believed that when the size of the hidden layers is changed, the recognition accuracy of the system varies. So we changed the size of the hidden layers to 512 and 2048 in experiments. Figure 2 summarizes the effects of different layer size on the performance of system.

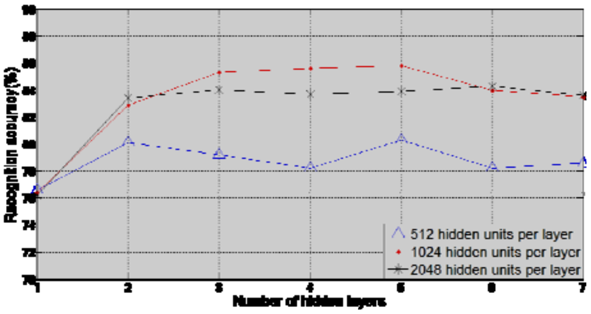

Figure.4 Emotion recognition rate using CWPCCs, with different sizes of hidden layers and hidden units per layer.

According to the above figure, the system with a layer size of 1024 and 2048 achieved better performance. In the recognition system architecture with hidden units of 2048, pre-training the first two layers of RBMs was enough for this emotion recognition task and the results stayed stable when increasing the number of hidden layers. The classifier with 512 units per layer achieved better performance than the other two classifiers when there was one hidden layer in the networks, but its performance was not as good as the other two when it formed a deep network. We found the best performance occurred when the size of the networks was 1024 and the number of hidden layers was 5, so we fixed them and then investigated the performance of varying the acoustic features that input to the neural networks.

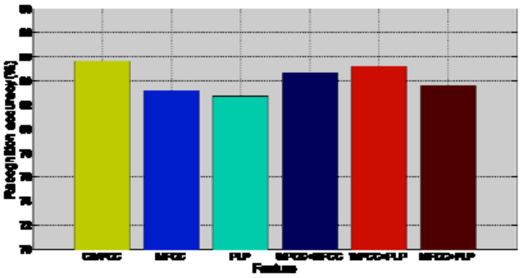

Figure 5 Emotion recognition rate using CWPCC, MFCC, PLP and their combinations

We can see from the above figure that the emotion recognition accuracy with CWPCC achieved the best performance. A highest classification accuracy of 85.60% is achieved for the proposed CWPCC feature set combined with DBNs feature learning. So we fixed the the size of the networks, the number of hidden layers and the acoustic feature CWPCC to compare with the conventional emotion recognition method that extracting WPCC features and the recognition as we did in our previous work[10].

TABLE 1. CONFUSION MATRIX WITH CONVENTIONAL EMOTION RECOGNITION METHOD

Emotion	Anger	Boredom	Fear	Joy	Neutral	Sadness
Anger	**89.22%**	0.00%	2.94%	7.84%	0.00%	0.00%
Boredom	0.00%	**84.62%**	1.54%	0.00%	6.15%	7.69%
Fear	1.82%	1.82%	**76.36%**	7.27%	5.45%	7.27%
Joy	28.07%	0.00%	14.04%	**57.89%**	0.00%	0.00%
Neutral	0.00%	4.76%	3.17%	0.00%	**88.89%**	3.17%
Sadness	0.00%	16.00%	0.00%	0.00%	2.00%	**82.00%**

TABLE II. CONFUSION MATRIX WITH PROPOSED EMOTION RECOGNITION METHOD

Emotion	Anger	Boredom	Fear	Joy	Neutral	Sadness
Anger	**93.46%**	0.00%	1.96%	4.58%	0.00%	0.00%
Boredom	0.00%	**88.53%**	1.54%	0.00%	3.17%	6.76%
Fear	1.82%	1.82%	**80.54%**	4.76%	4.76%	6.29%
Joy	23.75%	0.00%	11.73%	**64.52%**	0.00%	0.00%
Neutral	0.00%	3.72%	1.66%	0.00%	**92.47%**	2.14%
Sadness	0.00%	12.73%	0.00%	0.00%	1.07%	**86.34%**
Average recognition rate: 85.60%						

From Table I and Table II, we can know that the recognition accuracy of our proposed method has been improved compared with conventional emotion recognition method. In a conclusion, the acoustic feature CWPCC combined with the DBNs, whose size of the networks was 1024 and number of hidden layers was 5, improved 4.48% recognition accuracy.

VI. CONCLUSION

In this paper we explored the wavelet packet based acoustic feature extraction approach combined with DBNs feature learning for speech emotion recognition. The CWPCC feature is calculated following the conventional Mel-frequency Cepstral analysis paradigm. We tried different sizes of hidden layers (512, 1024, 2048), different number of DBNs layers and different types of acoustic features (CWPCC, MFCC, PLP) and their different combinations to model the emotion recognition system. The best result of recognition rate was 85.60%, achieved by combining the CWPCC features and DBNs feature learning model. The emotion recognition system using deep learning performed better than the conventional systems just using SVM as the classifiers. Future work also includes investigating more robust deep learning models. Apart from this, seeking for robust feature representation is also considered as part of the ongoing research, as well as efficient classification techniques for automatic speech emotion recognition.

CONFLICT OF INTEREST

The authors confirm that this article content has no conflicts of interest.

ACKNOWLEDGMENT

This work was supported by open Fund of Jiangsu Province Natural Science Foundation (no.BK20140649) and National Natural Science Foundation (no.61503081, no.61473079).

REFERENCES

[1] Caponetti L, Buscicchio C A, Castellano G.: Biologically inspired emotion recognition from speech. Eurasip Journal on Advances in Signal Processing, 2011(1), 1-10(2011)

[2] Morrison D, Wang R L, De Silva L C.: Ensemble methods for spoken emotion recognition in call-centres. Speech Communication, 49(2), 98-112(2007)

[3] Petrushin V.: Emotion recognition in speech signal: experimental study, development, and application. In: ICSLP 2000, Beijing, pp.222-225(2000)

[4] Malta L, Miyajima C, Kitaoka N et al.: Multimodal estimation of a driver's spontaneous irritation. In: Intelligent Vehicles Symposium, 2009 IEEE, pp.573-577(2009)

[5] France D J, Shiavi R G, Silverman S et al.: Acoustical properties of speech as indicators of depression and suicidal risk. IEEE Transactions on Biomedical Engineering, 47(7), 829-837(2000)

[6] Stephane, M.: A Wavelet Tour of Signal Processing, 3rd edn. Academic Press, Burlington (2009)

[7] Y. Bengio.: Deep learning of representations for unsupervised and transfer learning. Journal of Machine Learning Research-Proceedings Track, 27(2), 17-36(2012)

[8] G. E. Hinton, R. R. Salakhutdinov, Reducing the dimensionality of data with neural networks. Science, 313(5786), 504-507(2006)

[9] Y. Bengio.: Learning deep architectures for AI. Now Publishers Inc. 2(1), 67-76(2009)

[10] Yongming Huang, Ao Wu, Guobao Zhang, Yue Li.: Speech Emotion Recognition Based on Coiflet Wavelet Packet Cepstral Coefficients. In: Chinese Conference on Pattern Recognition 2014, pp.436-443(2014)

[11] G. E. Hinton and T. J. Sejnowski: Learning and relearning in boltzmann machines. MIT Press, Cambridge, Mass, 1, 282-317(1986)

[12] G. Hinton.: Training products of experts by minimizing contrastive divergence. Neural Computation,14(8), 1771-1800(2002)

[13] Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines. ACM Transac-tions on Intelligent Systems and Technology (TIST) 2(3), 1-27 (2011)

[14] Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W., Weiss, B.: A database of german emotional speech. In: Proceeding INTERSPEECH 2005, ISCA, pp. 1517-1520 (2005)