International Journal of Multimedia and Ubiquitous Engineering
Vol.12, No.2 (2017), pp.107-116
http://dx.doi.org/10.14257/ijmue.2017.12.2.08

Spoken Emotion Recognition by Combining Deep Belief Networks and Multi-layer Perceptron

Shiqing Zhang, Yueli Cui, Yuelong Chuang, Wenping Guo, Ying Chen and Xiaoming Zhao*

Institute of Intelligent Information Processing, Taizhou University,China
E-mail: tzxyzxm @163.com
Corresponding author: Xiaoming Zhao

Abstract

The recently-emerged deep learning theory has attracted extensive attentions in machine learning, signal processing, artificial intelligence and pattern recognition. Deep belief networks (DBN) is the well-known method of deep learning due to its strong ability of unsupervised feature learning. In this paper, a new method of spoken emotion recognition by combining DBN and multi-layer Perceptron (MLP), is proposed. DBN is used to perform unsupervised feature learning on the extracted emotional acoustic features. Then the learning results of the topside hidden layer of DBN is employed to initialize the MLP model for emotion classification. Experimental results on the popular Berlin database of German emotional speech, demonstrate the promising performance of the proposed method on spoken emotion recognition tasks, outperforming the other used methods.

***Keywords**: Deep learning, Deep belief networks, multi-layer Perceptron, Spoken emotion recognition, Unsupervised feature learning*

1. Introduction

In the last two decades, enormous efforts have been devoted to developing methods for automatically identifying human emotions from affective speech signals, *i.e*., called spoken emotion recognition. At present, spoken emotion recognition has become a very attractive research topic in signal processing, pattern recognition, artificial intelligence, *etc.*, due to its vital applications to human-machine interactions [1-2].

In recent years, deep learning is originally developed by Geoffrey Hinton *et al*. [3], according to the hierarchical architecture of information processing in the primate visual perception system. In this primate visual perception system, it has been found that the mammal brain is referred to be structured in a deep architecture [4], in which a given input percept can be represented with multiple levels of abstraction, and each level corresponds to a various area of cortex. The mammal brain is capable of implementing information processing by means of multiple stages of transformation and representation. Compared with the shallow learning methods, such as artificial neural network (ANN) and support vector machine (SVM), deep learning has two important properties. For one thing, deep learning is multi-layered with a deep architecture. For another, deep learning emphasizes the importance of unsupervised feature learning.

So far, deep learning, as a recently-emerged machine learning theory, has attracted extensive attentions in machine learning [5], signal processing [6], artificial intelligence [7] and pattern recognition [8]. Deep belief networks (DBN) [9], as a representative method of deep learning, exhibits a strong ability of unsupervised feature learning. In recent years, DBN has been successfully applied for acoustic modeling [10], natural language understanding [11], speech recognition [12] and so on.

ISSN: 1975-0080 IJMUE

Although DBN has a strong ability of unsupervised feature learning, it could not directly used for classification. To address this issue, in this work we present a new method of spoken emotion recognition by combining DBN with a traditional multi-layer perceptron (MLP) model, endowing DBN with a classification ability. To testify the performance of the proposed method, spoken emotion recognition experiments are conducted on the Berlin database of German emotional speech [13]. Experiment results demonstrate the effectiveness of the proposed method.

2. Review of Deep Belief Networks

Deep belief networks (DBN) [9] is developed by means of stacking a number of restricted Boltzmann machine (RBM).

2.1. RBM

Restricted Boltzmann machine (RBM) has one layer of (typically Bernoulli) stochastic hidden units and one layer of (typically Bernoulli or Gaussian) stochastic visible units. RBM could be represented as a bipartite graph, in which the whole visible units could be connected to the whole hidden units. In an RBM, the visible-visible or hidden-hidden connections are non-existent.

Given the model parameters q and the energy function $E(v,h;q)$, in a RBM the joint distribution $p(v,h;q)$ over the visible units v and the hidden units h is represented as

$$P(v,h;q) = \frac{\exp(- E(v,h;q))}{Z} \tag{1}$$

where $Z = \text{å}_{v}\text{å}_{h}\exp(- E(v,h;q))$ is a normalization factor. The energy function $E(v,h;q)$ in a Bernoulli (visible)-Bernoulli (hidden) RBM is defined as

$$E(v,h;q) = - \overset{I}{\underset{i=1}{\text{å}}}\ \overset{J}{\underset{j}{\text{å}}}\ w_{ij}v_ih_j - \overset{I}{\underset{i=1}{\text{å}}}\ b_iv_i - \overset{J}{\underset{j=1}{\text{å}}}\ a_jh_j \tag{2}$$

where w_{ij} denotes the symmetric interaction term between the visible unit v_i and the hidden unit h_j, b_i and a_j the bias terms, I and J are separately the amounts of the visible units and the hidden units. In this case, the conditional probabilities are denoted by

$$\begin{gathered} p(h_j = 1 \mid v;q) = s(\overset{I}{\underset{i=1}{\text{å}}}\ w_{ij}v_i + a_j) \\ p(v_i = 1 \mid h;q) = s(\overset{J}{\underset{j=1}{\text{å}}}\ w_{ij}h_j + b_i) \end{gathered} \tag{3}$$

where

$$s(x) = 1/(1+ \exp(x)) \tag{4}$$

In a RBM, the RBM weights are updated by using the following rule

$$\text{D}w_{ij} = E_{data}(v_ih_j) - E_{\text{mod}el}(v_ih_j) \tag{5}$$

where $E_{data}(v_ih_j)$ is the expectation value over in the training data and $E_{\text{mod}el}(v_ih_j)$ is that same expectation value over the distribution defined by the given model.

2.2. DBN

DBN can be developed by means of stacking a number of RBM learned layer-by-layer from bottom-up, as shown in Figure 1. In DBN the output of the lowest layer of DBN is taken as the input of the next layer, and then the output of the next layer is subsequently taken as the input of the higher level's layer. In practice, the training procedure of DBN includes two steps: pre-training and fine-tuning. Hinton *et al.* [13] has proved that a greedy learning method for unsupervised training is effective. This method is also called contrastive divergence (CD). And this CD learning method has effectively improved the training data's lower bound of likelihood probability, which is based on a hybrid model. And this learning procedure of DBN is unsupervised.

When using DBN for classification, the pre-training procedure of DBN could be employed to initialize the weights of a standard neural network such as a multi-layer perceptron (MLP) model, which could then be discriminatively fine-tuning by using the back propagating (BP) algorithm. The obtained weights of the trained DBN are taken as the weights of a standard neural network.

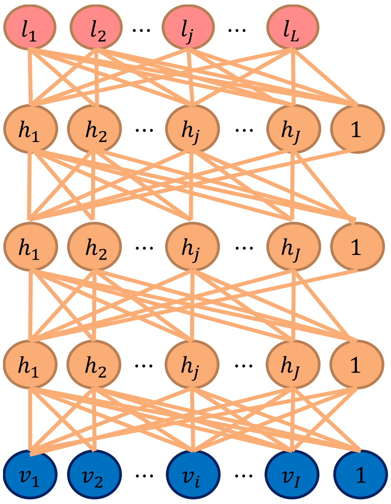
l_1 l_2 ⋯ l_j ⋯ l_L
h_1 h_2 ⋯ h_j ⋯ h_J 1
h_1 h_2 ⋯ h_j ⋯ h_J 1
h_1 h_2 ⋯ h_j ⋯ h_J 1
v_1 v_2 ⋯ v_i ⋯ v_I 1

Figure 1. An Illustration of DBN

3. The Proposed Method

Although DBN has a strong ability of unsupervised feature learning, it could not directly adopted for emotion recognition. To tackle this problem, we propose a new method of spoken emotion recognition by means of combining DBN with MLP. As a result, we can endow DBN with a capability of identifying spoken emotion.

Figure 2 illustrates the proposed method of spoken emotion recognition by combining DBN and MLP. The proposed method is composed of four key steps: acoustic feature extraction, DBN feature learning, MLP initialization and emotion classification, as described as follows.

First, acoustic feature extraction is to derive the relevant acoustic features related to human emotion expression from original speech signals. Second, DBN feature learning, including DBN's pre-training and fine-tuning, aims to perform feature learning on the low-level acoustic features extracted from speech signals, resulting in producing high-level feature representation. The extracted high-level feature representation is reflected in each hidden layer of DBN. In this work, the learning high-level feature representation of the topside hidden layer of DBN is employed to implement the latter used MLP initialization tasks. Third, the MLP initialization is to use the high-level feature representation learned by DBN to initialize the MLP classification model. The initialized MLP model owns the same parameters as DBN, including the number of hidden layers, the nodes of each hidden layer, and the weights of each hidden layer. Four, we employ the initialized MLP model as an emotion classifier to identify emotions, such as anger, joy, sadness, *etc.*

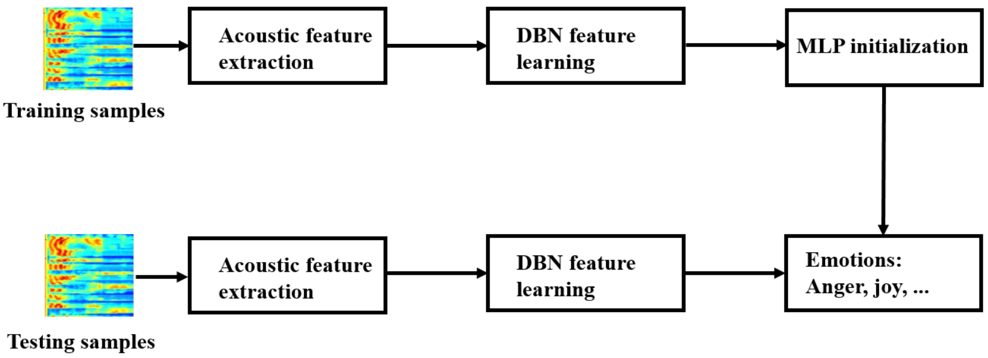

Figure 2. The Proposed Method of Spoken Emotion Recognition by Combining DBN and MLP

4. Acoustic Feature Extraction

So far, there is no general accordance regarding the most effective acoustic features used for spoken emotion recognition. Nevertheless, the most well-known acoustic features regarding human emotion expression contains three categories: prosody features, voice quality features as well as the spectral features. In our experiment, the extracted prosody features are pitch, intensity and duration. For the voice quality features extraction, we used are the first three formants (F1, F2, F3), spectral energy distribution, harmonics-to-noise-ratio (HNR), pitch irregularity (jitter) and amplitude irregularity (shimmer). For the spectral features extraction, the well-known Mel-Frequency Cepstrum Coefficients (MFCC) features are employed. When extracting these acoustic features, some typical statistical parameters such as mean, standard derivations (std), *etc.*, are computed. In summary, for each emotional utterance from the Berlin speech corpus, we finally extracted 25 prosody features, 23 voice quality features as well as 156 MFCC features. These extracted 204 features in total are statistical in Table 1. More details about each feature type can be found in our previous work [2].

5. Experiments

To verify the effectiveness of the presented DBN method on spoken emotion recognition tasks, we performed spoken emotion recognition experiments on the popular German Berlin emotional speech corpus.

5.1. Dataset

In our experiments, we employed the popular Berlin database of German emotional speech [13] for spoken emotion recognition. The Berlin speech corpus is an acted database, and contains about 530 emotional utterances with seven different acted emotions: anger, joy, sadness, neutral, boredom, disgust and fear. Ten professional native German-speaking actors (five female and five male) were asked to simulate these emotions, giving 10 German utterances (five short and five long sentences) which were able to be used in everyday communication. These actors were demanded to read these predefined sentences in the targeted seven emotions. The length of each speech utterance changed from three seconds to eight seconds. The recordings in this Berlin database were taken in an anechoic chamber with high-quality recording equipment and produced at a sampling rate of 16 kHz with a 16-bit resolution and mono channel.

Table 1. Acoustic Feature Extraction and Statistics

Feature types	Feature groups	Statistics
Prosody features	Pitch	maximum, minimum, range, mean, std, first quartile, median, third quartile, inter-quartile range, the mean-absolute-slope
	Intensity	maximum, minimum, range, mean, std, first quartile, median, third quartile, inter-quartile range
	Duration	total-frames, voiced-frames, unvoiced-frames, ratio of voiced vs. unvoiced frames, ratio of voiced-frames vs. total-frames, ratio of unvoiced-frames vs. total-frames
Voice quality features	Formants	mean of F1, std of F1, median of F1, bandwidth of median of F1, mean of F2, std of F2, median of F2, bandwidth of median of F2, mean of F3, std of F3, median of F3, bandwidth of median of F3
	Spectral energy distribution	band energy from 0 Hz to 500 Hz, band energy from 500 Hz to 1000 Hz, band energy from 2500 Hz to 4000 Hz, band energy from 4000 Hz to 5000 Hz.
	HNR	maximum, minimum, range, mean, std
	Jitter, Shimmer	Jitter, Shimmer
Spectral features	MFCC	mean, std of the first 13 MFCC, and their first-deltas and second-deltas

5.2. Experimental Setup

At first, all the extracted acoustic features were normalized with one variance and zero mean. When using DBN, the nodes of its visible layer correspond to the number of the input acoustic features. Two hidden layers are adopted for DBN, each of which has the nodes of 50, 100, 200, 300, 400. The DBN's recognition results were reported for the used nodes of hidden layer, *i.e.*, 50, 100, 200, 300, 400. In order to give better convergence results, the number of cycles was set to be 200 for DBN's pre-training and fine-tuning. All the algorithms are performed in the platform of MATLAB2014a.

Then, the best obtained performance of the proposed method combining DBN and MLP (denoted by DBN+MLP) was compared with three typical classification methods, such as K-nearest neighbor (KNN), artificial neural network (ANN), support vector machines (SVM). For KNN, the best values of K were found by using an exhaust search within the range [1, 20] with a step of 1. As one of the representative ANN, the typical radial basis function neural networks (RBFNN) method was employed due to its computation simplicity and promising performance. For SVM, The LIBSVM package, available at http://www.csie.ntu.edu.tw/cjlin/libsvm, was employed to perform the SVM algorithm with the simple linear kernel function, one-versus-one strategy for multi-class classification problem. Given the reliability of recognition results obtained by each classification method, a five-fold cross validation scheme was adopted in all spoken emotion recognition experiments.

Table 2. Recognition Results of DBN+MLP when Using DBN with Different Hidden Nodes

Hidden node of DBN	50	100	200	300	400
Accuracy (%)	76.07	80.75	83.55	79.81	74.39

Table 3. Comparison (%) of Different Classification Methods

Methods	**DBN+MLP**	RBFNN	KNN	SVM
Accuracy (%)	**83.55**	71.22	70.84	78.75

5.3. Results and Analysis

Table 2 presents the recognition results of DBN+MLP for the used nodes of hidden layer, *i.e.*, 50, 100, 200, 300, 400. Table III presents the comparison of recognition performance obtained with different classification methods, including DBN+MLP, RBFNN, KNN as well as SVM. It can be seen in Table 2 that the presented DBN method gives the highest accuracy of 83.55% when the node of hidden layer is set to be 200. The results in Table 3 show that the proposed DBN+MLP method obtains better performance than the other used methods such as RBFNN, KNN and SVM. In detail, the obtained recognition accuracy is 83.55% for DBN+MLP, 71.22% for RBFNN, 70.84% for KNN, and 78.75% for SVM. This demonstrates that the proposed DBN+MLP is an effective emotion classification method.

To further investigate the recognition accuracy per emotion when DBN+MLP performs best, the confusion matrix of seven emotion recognition results achieved by DBN+MLP is given in Table 4, in which the bold numbers denote the recognition accuracy for each emotion. The confusion matrix in Table 4 shows that anger and sadness, were classified well with an accuracy of 91.34% and 88.71%, respectively. In comparison, joy was identified with the lowest accuracy of 71.83% because joy was highly confused with anger.

Table 4. Confusion Matrix of Recognition Result when DBN+MLP Gives an Accuracy of 83.55% (*Ang-Anger, Joy-Joy, Sad-Sadness, Neu-Neutral, Fea-Fear, Bor-Boredom, Dis-Disgust)

	Ang	Joy	Sad	Neu	Fea	Bor	Dis
Ang	**91.34**	7.08	0.00	0.00	0.79	0.00	0.79
Joy	18.30	**71.83**	0.00	2.82	4.23	0.00	2.82
Sad	0.00	1.61	**88.71**	4.84	0.00	3.23	1.61
Neu	0.00	2.53	2.53	**81.01**	1.27	8.86	3.80
Fea	4.35	5.80	4.35	4.35	**79.71**	0.00	1.44
Bor	0.00	0.00	2.47	11.11	0.00	**83.95**	2.47
Dis	4.36	0.00	6.52	2.17	2.17	2.17	**82.61**

6. Conclusions

Automatic spoken emotion recognition has increasingly attracted attention duo to its important applications to human computer interaction. This paper proposes a new method of spoken emotion recognition by combining DBN and MLP. Experimental results on the Berlin database of German emotional speech demonstrate the effectiveness of the proposed method on spoken emotion recognition tasks. This can be attributed to DBN having a strong ability of unsupervised feature learning. In future, we will focus on more advanced deep learning methods for spoken emotion recognition.

Acknowledgments

This work is supported by Zhejiang Provincial Natural Science Foundation of China under Grant No. LY16F020011, No. LY14F020036 and No. LQ15F020001, National Natural Science Foundation of China under Grant No. 61203257 and No. 61272261, Scientific Research Fund of Zhejiang Provincial Education Department under Grant No.Y201636701, and Taizhou Science Plan No.1501KY64.

References

[1] J. Zhengbiao, Z. Feng and Z. Ming, “An Algorithm Study for Speech Emotion Recognition Based Speech Feature Analysis”, International Journal of Multimedia and Ubiquitous Engineering, vol. 10, no. 11, **(2015)**, pp. 33-42.

[2] X. Zhao and S. Zhang, “Robust emotion recognition in noisy speech via sparse representation”, Neural Computing and Applications, vol. 24, no. 7-8, **(2014)**, pp. 1539-1553.

[3] G. E. Hinton and R. R. Salakhutdinov, “Reducing the dimensionality of data with neural networks”, Science, vol. 313, no. 5786, **(2006)**, pp. 504-507.

[4] T. Serre, G. Kreiman, M. Kouh, C. Cadieu, U. Knoblich, and T. Poggio, “A quantitative theory of immediate visual recognition”, Progress in brain research, vol. 165, **(2007)**, pp. 33-56.

[5] Y. L. Cun, Y. Bengio, and G. Hinton, “Deep learning”, Nature, vol. 521, **(2015)**, pp. 436-444.

[6] D. Yu and L. Deng, “Deep learning and its applications to signal and information processing [exploratory dsp]”, IEEE Signal Processing Magazine, vol. 28, no. 1, **(2011)**, pp. 145-154.

[7] I. Arel, D. C. Rose, and T. P. Karnowski, “Deep machine learning-a new frontier in artificial intelligence research [research frontier]”, IEEE Computational Intelligence Magazine, vol. 5, no. 4, **(2010)**, pp. 13-18.

[8] J. Schmidhuber, “Deep learning in neural networks: An overview”, Neural Networks, vol. 61, **(2015)**, pp. 85-117.

[9] G. E. Hinton, S. Osindero, and Y.-W. Teh, “A fast learning algorithm for deep belief nets”, Neural computation, vol. 18, no. 7, **(2006)**, pp. 1527-1554.

[10] X. Li, Y. Yang, Z. Pang, and X. Wu, “A Comparative Study on Selecting Acoustic Modeling Units in Deep Neural Networks Based Large Vocabulary Chinese Speech Recognition”, Neurocomputing, vol. 170, **(2015)**, pp. 251-256.

[11] R. Sarikaya, G. E. Hinton, and A. Deoras, “Application of Deep Belief Networks for Natural Language Understanding”, IEEE Transactions on Audio, Speech, and Language Processing, vol. 22, no. 4, **(2014)**, pp. 778-784.

[12] T. Yoshioka and M. J. Gales, “Environmentally robust ASR front-end for deep neural network acoustic models”, Computer Speech & Language, vol. 31, no. 1, **(2015)**, pp. 65-86.

[13] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, “A database of German emotional speech”, in Proceeding of Interspeech-2005, Lisbon, Portugal, **(2005)**, pp. 1-4.

Authors

Shiqing Zhang, received the Ph.D. degree at school of Communication and Information Engineering, University of Electronic Science and Technology of China, in 2012. Currently, he works as an associate professor of the Institute of Intelligent Information Processing, Taizhou University, China. His research interests include affective computing and pattern recognition.

Yueli Cui, received the B.S. degree from the Zhejiang University City College (ZUCC), Hangzhou, in 2006 and the M.S. degree from the Hebei University, Baoding, in 2009, both in electronics and communication engineering. Currently, he works as a lecturer of the Institute of Intelligent Information Processing, Taizhou University, China. His research interests include signal processing and pattern recognition.

Yuelong Chuang, received B. S. degree and M. S. degree in School of computer and communication engineering in Liaoning Shihua University in 2000 and 2008, PhD. degree in College of Computer Science at Zhejiang University in 2013. Currently, he works as a lecturer of the Institute of Intelligent Information Processing, Taizhou University, China. His research interests include computer vision, machine learning and pattern recognition.

Wenping Guo, received his B.S. degree in mechatronic engineering and the M.S. degree in computer science from Southwest Jiaotong University, China, in 1998 and 2005 respectively. He was a visiting scholar at Jacksonville State University, USA, from February 2012 to June 2012. He is currently an associate professor of the Institute of Intelligent Information Processing, Taizhou University, China. His research interests including multimedia annotation and machine learning.

Ying Chen, received his B.S. degree and M.S. degree in computer science from Southwest Jiaotong University, Chengdu, China. Currently he is a Ph.D. candidate at College of Computer and Information in Hohai University, Nanjing, China. He is also working as a Lecturer in the Institute of Intelligent Information Processing at Taizhou University, Taizhou, China. His research interests include machine learning and pattern recognition.

Xiaoming Zhao, received the B.S degree in mathematics from Zhejiang Normal University in 1990 and the M.S degree in software engineering from Beihang University in 2006. He is currently a professor of the Institute of Intelligent Information Processing, Taizhou University, China. His research interests include machine learning and pattern recognition.