International Journal of Recent Engineering Science (IJRES)
ISSN: 2349 – 7157 / doi: 10.1445/23497157/IJRES-V5I3P101

Volume 5 Issue 3, 1-7, May-Jun 2018
© 2018 Seventh Sense Research Group®

Improving Speech Emotion Recognition Method of Convolutional Neural Network

ZENG Runhua[1], ZHANG Shuqun[1]
[1]*(College of Information Science and Technology, Jinan University, Guangzhou, China*

Abstract

In this paper, we studied speech emotion recognition and proposed an improved speech emotion recognition method of the convolutional neural network. Improved methods are improving the algorithm of updating convolution kernel weight and transforming the data matrix of the Mel-Frequency Cepstral Coefficients (MFCC) obtained by preprocessing the speech signal. This makes that the algorithm of updating the convolution kernel weight during the training process of traditional convolutional neural networks was related to the number of iterations and increase the difference of emotional phonetic features. Therefore this improved the expressive ability of convolutional neural networks. Experiments showed that the error recognition rate of the improved speech emotion recognition method of the convolutional neural network was about 7% lower than that of the traditional method.

Keywords - *speech emotion recognition, Mel-frequency cepstral coefficients (MFCC), convolutional neural networks, recognition rate*

I. INTRODUCTION

With the development of science and technology, artificial intelligence is becoming more and more mature. However, the development of artificial intelligence in self-awareness and self-awareness is negligent, hindering the natural interaction between humans and machines. An essential approach to solving the problem of self-awareness and self-awareness of artificial intelligence is to make the machine understand the emotional state of humans and make the intelligent machine possess the emotional ability. At present, speech emotion recognition by emotional signals has drawn more and more attention [1]

Speech emotion recognition refers to recognizing advanced and effective emotion state from the low-level features of Speech, which can be regarded as the classification problem based on speech sequence. The main processes of speech emotion recognition include the establishment of an emotion library, the extraction of speech emotion feature, dimension reduction and selection, and emotion classification recognition. At present, there are many methods for speech emotion recognition, such as Hidden Markov Model (HMM) [2], Artificial Neural Network (ANN), Gaussian Mixture Model (GMM), Support Vector Machine (SVM)[3-4], K-Nearest Neighbor (KNN) and Maximum Likelihood Bayesian Classification [5] and that had achieved some results. However, due to the different subjects (languages), and there is no uniform standard in the corpus database, there is a significant difference between the recognition results [6-8].

SVM and KNN are often used in deterministic models, while human emotions have complex and uncertain information. Neural Networks are typical non-deterministic models with nonlinear I/O mapping, powerful generalized ability, self-learning, self-organizing and self-adaptive ability. It has unique advantages in dealing with the problem of uncertainty and nonlinear mapping. It can detect and extract the law and trend that human or other classification techniques cannot detect. In various neural network models, The Convolutional Neural Networks (CNN) is the most used and most successful multi-layer feedforward network in pattern recognition. Because human emotion has intense complexity and uncertainty, the recognition rate of speech emotion based on convolutional neural networks is still not high [8]. Therefore, to increase the feature difference between Speech personal information, this paper proposes an improved speech feature extraction algorithm and an improved convolutional neural network algorithm.

II. EXTRACTING OF SPEECH EMOTIONAL FEATURES

The traditional method of extracting Speech emotional features is to globally analyze emotional speech signals and extract parameters, such as pitch frequency, amplitude energy, speech rate, and formant. Analyzing the timing and distribution features of these feature parameters and finding the rhythm law of different emotional Speech regard based on emotion recognition. In this paper, we extract the 40-dimensional emotion feature vectors A using the method of extract the relevant emotional characteristics of speech signals (speech rate, spectral features, fundamental frequency characteristics, and energy mean features) in the reference [9]. The form of feature vectors A is as follows.

$$A=\begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,40} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,40} \\ \vdots & \vdots & \ddots & \vdots \\ a_{40,1} & a_{40,2} & \cdots & a_{40,40} \end{bmatrix} \qquad (1)$$

Simultaneously, from the signal analysis point of view, the speech signal is composed of many different frequency signals that overlap together. Analysis of the spectral properties of signals also contributes to the study of affective recognition. However, MFCC is based on the acoustic properties of the human ear [10]. It uses a nonlinear frequency unit (Mel frequency) to simulate the human auditory system. Scholars at home and abroad have done much-related research, and MFCC is applied to speech recognition [11]. Therefore, the MFCC feature extraction method is used in this paper.

At present, the traditional MFCC feature extraction method uses 256 sampling points as a frame length, 160 sampling points as a frameshift, and an MFCC coefficient order is 12. Moreover, calculating the energy level, the first-order difference, and the second-order difference for each frame length, then getting the average value of each frame. So each frame will be a 40-dimensional Mel band filter coefficients. However, each sample of speech samples is not uniform; the number of frames is also inconsistent. To get a unified number of frames, the studies following two methods to extract the characteristics of the database data.

Method 1: The commonly used feature extraction scheme is to extract the feature data directly from each speech sample. However, the number of extracted feature data is not uniform, and most of them are between 140 and 170 frames. The features of the Speech preserved mostly by the data of each sample feature is uniformly cropped to 160 frames (less than 160 frames to complement 0 expansion, more than 160 frames of a direct crop). The feature data is further transformed into an 80 × 80 matrix form as input to the convolutional neural network.

Method 2: Since the data extracted by method 1 is too large, the training time is longer. The feature data dimensionality reduced by the speech sampling points of each sample are uniformly structured as 7136 sampling points and then get unified 40 frames of characteristic data through the extraction of the MFCC coefficients. Further, the feature data is transformed into a 40 × 40 matrix A_1 as the input of the convolutional neural network. The form of A_1 is the same as A.

After many experiments comparing the two recognition results, the results show that the second method of training needs less-time, and the recognition rate is high. At the same time, this paper considers whether to normalize the data A_1, comparing the experimental results shows that the recognition rate is not improved after the normalization of the feature data A_1, and the identification is not stable. So we use the unnormalized data A_1 as the input of the neural network. At the same time, the feature vectors A and A_1 are respectively taken as the input of the convolutional neural network model, and the results of the two experiments are shown in Table 4. Experiments show that speech emotion error the recognition rate is lower when MFCC feature data vectors A_1 is used as the input of the neural network.

Human emotions have very complex and uncertain information. To reflect the feature difference between speech emotions and increase the influence of convolution neural network input feature data on speech emotion recognition. This paper proposes an improved emotion-speech feature extraction algorithm: The A_1 matrix element is multiplied by n times as the input characteristic data of the convolutional neural network; the specific method is as follows:

$$A_n = n\times\begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,40} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,40} \\ \vdots & \vdots & \ddots & \vdots \\ a_{40,1} & a_{40,2} & \cdots & a_{40,40} \end{bmatrix} \qquad (2)$$

Among them, A_n is used as the input characteristic data for the enlarging convolutional neural network. From formula (2), we can see that the input characteristic data of the convolutional neural network is *n* times larger without increasing the workload of extracting feature data. To study the value of *n*, in this paper, we make much experimental comparison to *n*=1,2,3,4,5 (convolutional neural network input characteristic data is the A_1, A_2, A_3, A_4, and A_5). The results are shown in Table 1.

Table 1 Comparison of convolutional neural network models of different values of n，such as the error recognition rate.

n	1	2	3	4	5
Number of iterations	60	60	60	60	60
Convergence iteration times	21	20	23	25	30
Training speed / second	665	718	735	756	788
Error recognition rate /%	51.21	49.68	50.17	50.69	52.30

As shown in Table 1, when *n* = 2, under the premise of not affecting the training speed, the error recognition rate is the lowest, and the convergence iteration number is the lowest.

III. INTRODUCTION AND IMPROVEMENT OF THE CNN

CNN is composed of one or more sets of convolutional layers and aggregation layers [12]. A convolutional layer contains several different convolvers, which extract the local features of the Speech. The aggregation layer reduces the number of input nodes of the next layer through the fixed window length of the polymerization on the convolution layer output node; therefore, it controls the complexity of the model. General aggregation layer using the maximum aggregation algorithm that output the maximum fixed length the node. Finally, through the whole network layer, the output value of the aggregated layer is combined to get the final classification decision result. It obtains better performance in image processing [13]. The CNN is implemented the extraction of the local information of speech features by convolution, and it enhances the original signal characteristics and reduces the noise.

Moreover, it also makes the convolution neural network model based on convolution operation have better anti-noise performance. Through the aggregation to enhance the robustness of the model. At the same time, weight sharing dramatically reduces the storage size of the model.

Because the structure of a convolutional neural network can be used in a multi-layer convolutional layer, this paper compares the traditional convolutional neural network model with different levels. The results are shown in Table 2. The size of each layer in the model is the same. As shown in Table 2, the structure model of one convolution layer has the least training time, but the false recognition rate is highest. The lowest error recognition rate of the structure model of two convolution layers and the training speed (the time of completing a model training) far less than model of three or four convolutional layer and the convergence number of iterations (has been iterated times when converging) is least. Therefore, this paper uses two convolutional network model of two convolution layer, as shown in Fig. 1. The input data is 40 × 40 eigenvector matrix, the output is 6-dimensional data, two convolutional layer, two polymerization layer, and fully connected layer.

Table 2 Comparison of performance parameters of traditional convolutional neural networks models of different layers，such as the error recognition rate

Traditional Convolutional Neural Network Model	Input Data	Convolution kernel size	Polymer layer size	Number of iterations	convergence number of iterations	The training speed	the error recognition rate /%
One layer convolution layer	A	33×33	2×2	60	43	312	68.32
two layer convolution layer	A	9×9 9×9	2×2	60	21	665	51.21
three layer convolution layer	A	9×9 5×5 3×3	2×2	60	49	1263	51.34
four layer convolution layer	A	5×5 3×3 3×3 3×3	2×2	60	55	3216	51.52

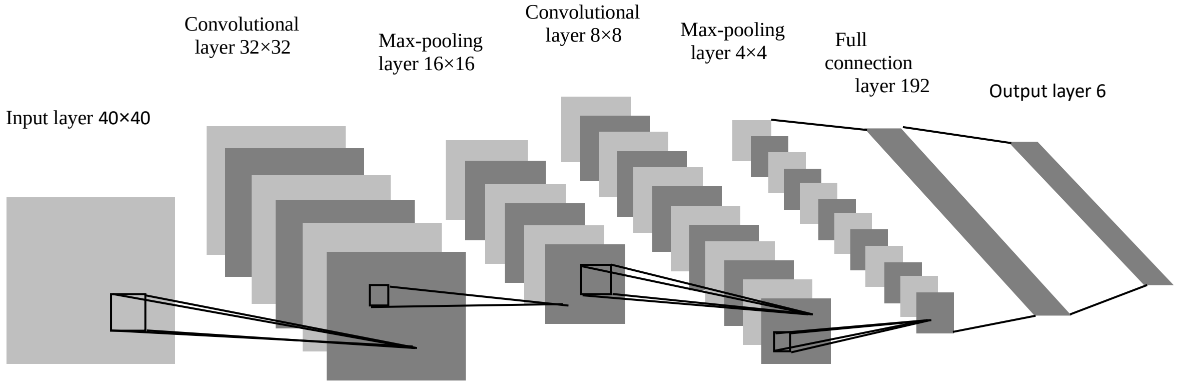

Figure 1 Network Structure of Convolutional Neural Networks

A. The Convolution Layer

In the convolution layer, the feature data of the previous layer is convolved with a learnable convolution kernel, and then the layer's feature data is outputted by an activation function. Every output feature data may contain a convolution of a plurality of input data. The general form of a convolutional layer is as follows:

$$x_j^l = f(\sum_{i\in M_j} x_i^{l-1} * k_{ij}^l + b_j^l) \quad (3)$$

Wherein x_j^l represents the *j*th feature data in the *l*th layer of the convolutional layer; $f(\cdot)$ represents the activation function; k_{ij}^l represents the convolution kernel weight of the *j*th output characteristic data corresponding to the *i*th input characteristic data in the *l*th layer of the convolutional layer; M_j represents set of input data; $*$ represents a convolution operation; b_j^l is offset.

B. The aggregation layer

The aggregation layer performs a sampling operation on the input [14]. If there are n input feature data, the output feature data is also n, but the output feature data becomes smaller relative to the input data. The down-sampling general form is as follows:

$$x_j^i = f(\beta_j^l down(x_j^{l-1}) + b_j^l) \quad (4)$$

Wherein $down(\cdot)$ represents the downsampling function. The down-sampling function is generally a weighted summation of the input data to a k × k-sized area of the layer. So the size of the output data is 1/k times of the input data, β_j^l is the weighting coefficient, and b_j^l is offset.

C. The improve the algorithm

In training convolution kernel weights in the traditional convolutional neural network, the convolutional kernel weights are updated by backtracking errors. The general form of which is as follows:

$$k^l_{ij_{new}} = k^l_{ij_{old}} - d^l_{ij} \quad (5)$$

Where $k^l_{ij_{old}}$ - represents the convolution kernel weight before update the *i*th input characteristic data corresponding to the *j*th output characteristic data in the *l*th layer of the convolution layer; $k^l_{ij_{new}}$ - represents the updated convolution kernel weight of the *i*th input characteristic data corresponding to the *j*th output characteristic data in the *l*th layer of the convolution layer; d^l_{ij} is the convolution kernel weight error of the *j*th output characteristic data corresponding to the *i*th input characteristic data in *l*th layer obtained by feedback. d^l_{ij} - The convolution kernel weight error of each layer that the Loss function square error gets by de-sampling inversely propagated. In the process of training the convolution kernel weight according to the traditional algorithm, the learning rate n is 1 and is always unchanged and too broad. The modification of convolution kernel weight at each iteration is realized by directly subtracting the convolution kernel weight error, and convolution kernel weight is easy to cross the extreme point during training. Therefore, the traditional update algorithm may result in instability of convolution kernel weight update with the increase of iteration times. For this reason, in this paper, we propose an improved algorithm that makes every change of convolution kernel weight decrease non-linearly or linearly with the increase of the iteration number, and the learning rate n decreases gradually with the increase of the number of iterations. Nonlinear decreasing formula (5) will be improved as follows:

$$k^l_{ij_{new}} = k^l_{ij_{old}} - d^l_{ij} + \exp(\ln(i/e)) \times d^l_{ij} \quad (6)$$

Wherein e represents the total number of iterations of training, i is the number of current iterations, and its learning rate is $\eta = 1 - \exp(\ln(\mathrm{i/e}))$. According to formula (6), we can see that the size of each modified value of convolution kernel weight decreases non-linearly with the increase of the iteration number.

Linear decreasing formula (5) will be improved as follows:

$$k^l_{ij_{new}} = k^l_{ij_{old}} - d^l_{ij} + a \times (i/e) \times d^l_{ij} \qquad (7)$$

Wherein e represents the total number of iterations of training; *i* is the number of current iterations. To $k^l_{ij_{old}} - k^l_{ij_{new}}$ is greater than 0 and is less than d^l_{ij}, in which the value of the parameter a must be greater than 0 and less than 1. After many experimental comparisons, the training effect is best when a=0.9 and the learning rate is $\eta = 1 - a \times (\mathrm{i/e})$. According to formula (7), we can be seen that the change of each update of convolution kernel weight decreases linearly with the increase of iteration number. In this paper, a large number of experiments were made to compare the nonlinear improvement algorithm of formula (6) and the linear improvement algorithm of formula (7); we can be seen that using the improved linear algorithm of formula (7), the convolutional kernel weights become more stable with the increase of the number of iterations and speech emotion recognition rate is higher. Specific results see the third part of this article.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

A. *Experimental conditions*

In this paper, the database we use is CASIA Chinese Emotional Corpus, recorded by CAS Automation. Under the pure recording environment (signal to noise ratio is about 35 dB), four recording artists (two men and two women) recorded 50 sentences of Speech respectively under six different types of emotions (angry, scared, happy, neutral, sad, astonished), 16 kHz sampling, 16-bit quantization, a total of 1200 samples.

Experiments using characteristic data formula (1) A_1 were described in the second part of the paper. The improved convolutional neural network is also used to compare several multi-layer network structures. The results show that the two-layer convolutional neural network model is the best one, as shown in Figure 1. So the convolution neural network structure model used in this paper is shown in Figure 1: the convolution kernel size is 9 × 9, the aggregation layer size is 2 × 2, and the batch size is B. After a large number of experiments, it is found that the best training result is obtained when B = 10. Therefore, the experimental batch size of B = 10. The experimental hardware platform: Core (TM) i5 processor, 3.20GHz Memory. Software platform: the deep learning framework Caffe, Matlab language.

B. *Experimental results*

In this experiment, 80% of each emotion was selected randomly as the training set and 20% as the test set. The training set data were used to train the network structure model, and the test set data were used to test the network structure model. Under using the trained network structure model to test the Speech Emotion recognition, the model of the output vector is not a binary integer vector, and its actual component is close to 0 or 1. At this time, the output vector of each component rounded up to the nearest integer to form a binary vector, and each component can be identified corresponding to what kind of emotion.

To further verify the affection of the emotion-speech feature extraction algorithm in the first part of this paper on the improved CNN model, we also use A_1, A_2、A_3、A_4, and A_5 as the improved neural network inputs. As shown in Table 1, the experimental results show the same conclusion: when n = 2, under the premise of not affecting the training speed, the error recognition rate is the lowest, and the convergence iteration number is the lowest. Therefore, we use the characteristic data A_2 in formula (2) as an improved CNN model input.

The improved Convolution Neural Networks model CNN (A_2) (the characteristic data A_2 of formula (2) as input, using improved algorithm of formula (7)). The test results are shown in Table 3.

In order to test the performance of the model, we compare three different models of convolutional neural networks (the network structure is shown in Fig. 1): the traditional Convolution Neural Network model CNN (A_1) (the characteristic data A_1 as input), the improved Convolution Neural Network model CNN (A_1) (the characteristic data A_1 as input, using an improved algorithm of formula (7)) and the improved Convolution Neural Network model CNN (A2). The results, as shown in Fig. 2.

Table 3 Test results of improved Convolutional Neural Networks model (CNN(A_2))

Emotion category	number of Samples	Number of training	Number of tests	Number of error recognition	the error recognition rate
angry	200	160	40	9	22.50%
fear	200	160	40	25	62.50%
happy	200	160	40	16	40.00%
neutral	200	160	40	29	72.50%
sad	200	160	40	6	15.00%
surprise	200	160	40	21	52.50%

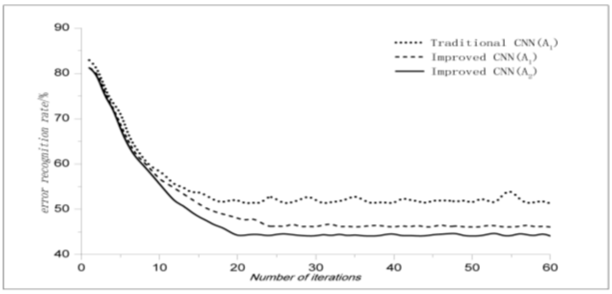

Figure 2 Comparison curve of the error recognition rates

As can be seen from the comparison of the error recognition rates in Fig. 2: 1), Although the traditional CNN (A_1) model starts to converge on the 21st iteration gradually, convergence is instable; The improved CNN (A_1) model starts to converge at the 24th iteration and is relatively stable. This is because the amount of change of each renewal of the convolution kernel weight decreases gradually with the number of iterations. Therefore, with the number of iterations increasing, the convolution kernel weights gradually become stable. The convergence rate of identifying the model is stable, and the error recognition rate is significantly reduced; 2) Under the same training conditions, the improved CNN (A_2) model begins to converge and stabilize at the 20th iteration model converge speed is faster. This is because doubling characteristic data indeed better reflects the feature differences of speech emotion.

Table 4 Comparison of performance parameters of different convolutional neural networks models，such as the error recognition rate

Models	Input Data	Convolution kernel size	Polymer layer size	The number of iterations	The convergence number of iterations	The training speed	the error recognition rate /%
Traditional CNN (A)	A	9×9 9×9	2×2	60	23	721	55.46
Traditional CNN (A_1)	A_1	9×9 9×9	2×2	60	21	665	51.21
Traditional CNN (A_2)	A_2	9×9 9×9	2×2	60	20	718	49.68
Improved CNN (A)	A	9×9 9×9	2×2	60	21	752	50.35
Improved CNN (A_1)	A_1	9×9 9×9	2×2	60	24	711	46.12
Improved CNN (A_2)	A_2	9×9 9×9	2×2	60	20	723	44.17
nonlinear Improved CNN (A_2)	A_2	9×9 9×9	2×2	60	26	826	46.02

In this paper, we also made a comparative experiment on the nonlinear modified CNN (A_2) model (the characteristic data A_2 in formula (2) as input and using the improved algorithm of formula (6)). The results are shown in Table 4. From Table 4, it can be seen that the improved CNN (A_2) model has a faster convergence rate, less training time, and lower error recognition rate than the nonlinear modified CNN (A_2) model. Its error recognition rate of speech emotion is 44.17%. The literature [7] published in foreign publications used the traditional MFCC feature extraction method (see Part 1) in the emotional Speech feature extraction, and the error recognition rate of speech emotion based on the traditional convolution neural network in literature [7] is 63.70%. The error recognition rate of speech emotion based on deep convolutional neural networks in literature [8] is 59.98%. In contrast, Compared with previously published convolutional neural network speech emotion recognition, we can find that the error recognition rate of speech emotion in the improved CNN (A_2) model in this paper is significantly lower.

V. CONCLUSION

In this paper, the algorithm principle of speech emotion recognition based on a convolutional neural network is discussed. MFCC feature data matrix is obtained by preprocessing the Speech according to the characteristics of Speech. To increase the feature difference between speech emotions characteristics, we transform the feature data matrix and propose CNN(A_2) model. The convolution kernel weight update algorithm is improved to improve the recognition ability of speech emotion. Experimental results show that convolutional neural networks can not only in image recognition and language recognition but also be applied in speech emotion recognition.

REFERENCES

[1] Anagnostopoulos C N, Iliou T, Giannoukos I. “*Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011*“[J]. Artificial Intelligence Review, 2015, 43(2): 155-177.

[2] Juang B H, Rabiner L. “*Mixture autoregressive hidden Markov models for speech signals*”[J]. Procedia Computer Science, 2015, 61(6): 328-333.

[3] Kirandeep Singh "*Speech Recognition: A Review of Literature*", International Journal of Engineering Trends and Technology (IJETT), V37(6),302-310 July 2016.

[4] Hu H, Xu M X, Wu W. “*GMM super vector-based SVM with spectral features for speech emotion recognition*”[C], IEEE International Conference on Acoustics, 2007: IV-413-IV-416.

[5] Lee C M, Yildirim S, Bulut M, et al. “*Emotion recognition based on phoneme classes*”[J]. Proc Icslp', 2004: 889-892.

[6] Mao Q, Dong M, Huang Z, et al. “*Learning salient features for speech emotion recognition using convolutional neural networks*” [J]. IEEE Transactions on Multimedia, 2014, 16(8): 2203-2213.

[7] Zhang B, Quan C, Ren F. “*Performance of convolution neural network on the recognition of speech emotion and images*”[C]. AIA International Advanced Information Institute. 2016: 12-21.

[8] Zheng W Q, Yu J S, Zou Y X. “*An experimental study of speech emotion recognition based on deep convolutional neural networks”*[C]. International Conference on Affective Computing and Intelligent Interaction. 2015: 827-831.

[9] Guo P. “*Research on emotion recognition from speech-features and models*”[D]. Northwestern Polytechnical University, 2007.

[10] Davis S, Mermelstein P. “*Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences”*[J]. Readings in Speech Recognition, 1990, 28(4): 65-74.

[11] Hinton G, Deng L, Yu D, et al. “*Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups*” [J]. IEEE Signal Processing Magazine, 2012, 29(6): 82-97.

[12] Bengio Y, Lecun Y. “*Convolutional networks for images, Speech, and Time-Series”*[J]. 1995.

[13] Krizhevsky A, Sutskever I, Hinton G E. “*Image net classification with deep convolutional neural networks”*[J]. Advances in Neural Information Processing Systems, 2012, 25(2): 2012.

[14] Lecun Y, Bottou L, Bengio Y, et al. “*Gradient-based learning applied to document recognition”*[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.

[15] Shaveta Sharma , Parminder Singh "*Speech Emotion Recognition using GFCC and BPNN*", International Journal of Engineering Trends and Technology (IJETT), V18(6),321-322 Dec 2014.

[16] Vlassis N, Likas A. “*A greedy EM algorithm for Gaussian mixture learning*”[J]. Neural Processing Letters, 2002, 15(1): 77-87.