Hindawi Publishing Corporation
Journal of Electrical and Computer Engineering
Volume 2016, Article ID 7437860, 11 pages
http://dx.doi.org/10.1155/2016/7437860

Research Article

A Novel DBN Feature Fusion Model for Cross-Corpus Speech Emotion Recognition

Zou Cairong,[1,2] Zhang Xinran,[2] Zha Cheng,[2] and Zhao Li[2]

[1] *Department of Information and Communication Engineering, Guangzhou Maritime University, Guangzhou 510006, China*
[2] *Key Laboratory of Underwater Acoustic signal Processing of Ministry of Education, Southeast University, Nanjing 210096, China*

Correspondence should be addressed to Zhang Xinran; zxrzxr87324@126.com

Received 24 June 2016; Revised 19 October 2016; Accepted 15 November 2016

Academic Editor: Alexey Karpov

The feature fusion from separate source is the current technical difficulties of cross-corpus speech emotion recognition. The purpose of this paper is to, based on Deep Belief Nets (DBN) in Deep Learning, use the emotional information hiding in speech spectrum diagram (spectrogram) as image features and then implement feature fusion with the traditional emotion features. First, based on the spectrogram analysis by STB/Itti model, the new spectrogram features are extracted from the color, the brightness, and the orientation, respectively; then using two alternative DBN models they fuse the traditional and the spectrogram features, which increase the scale of the feature subset and the characterization ability of emotion. Through the experiment on ABC database and Chinese corpora, the new feature subset compared with traditional speech emotion features, the recognition result on cross-corpus, distinctly advances by 8.8%. The method proposed provides a new idea for feature fusion of emotion recognition.

1. Introduction

In recent years, more attention is paid to the study of emotion recognition. Speech, as one of the most important ways of communication in human daily life, contains rich emotional information. Speech emotion recognition (SER), because of its wide application significance and research value in intelligence and naturalness of human-computer interaction aspects [1], gets more and more attention from the researchers in recent years. Emotion recognition system performance determines the quality of information feedback and the efficiency of human-computer interaction, while overall performance of SER depends on the matching degree between features and classifiers [2]. Although the earlier temporal features may not be suitable for the current corpus structures [3], the emotional information contained on the time domain still has good representation ability to be reserved. In order to research SER on the broader technology level, extending the database source and searching suitable fusion model for big emotional information data have become new focuses [3, 4].

Feature layer fusion is the integration of data after preprocessing and feature extraction, so many related researches [5, 6] are applied to this area. Through specific means such as fusion, the scale of feature sources is enhanced and the data sets are expanded. Further, some effective data analysis techniques are introduced and applied, such as Neural Network and Deep Learning. Common feature fusions are often used for single source data samples. Because the emotional properties of different features are various, the cross-corpus recognition effects of current fusion methods are not satisfactory. The development of Deep Learning technology brings a new orientation to SER. Using appropriate algorithm to train the deep neural network model, more valuable features can be derived from the vast amounts of original databases which are multiple sources [7]. Accordingly, Deep Belief Nets (DBN) model, which is a commonly used model in Deep Learning area [8], is introduced in our work. Through Restricted Boltzmann Machine (RBM) [9], DBN could constantly adjust the connection weight, which can realize effective fusion of features. Previous cross-corpus studies are dependent on traditional suprasegmental acoustic global features, which are often used in emotion recognition technology [10]. Since the emotional features have great significance to SER, exploring new features to promote the development of SER has an irreplaceable role in the cross-corpus research. Thus, this article introduces a new emotional feature category based on visual attention mechanism. The new feature space includes three kinds of image vectors: color, brightness, and orientation. Features extracted from spectrogram connect the time domain and frequency domain, so they have important significance for cross-corpus SER research. The new direction of the research which uses the spectrogram emotional features [4, 5] has advantages for its overall information. The integrated features with traditional acoustic traits could combine the global and temporal features, which supplement the original feature space.

This paper mainly studies the feature fusion method based on DBN which fuse spectrogram features and acoustic global features for SER. In Section 2, by selective attention mechanism, the features with time-frequency domain correlation traits are extracted while the emotion recognition abilities are analyzed. Then in Section 3, based on the DBN fusion method on feature level, an alternative DBN (so-called DBN21) feature fusion layer model is proposed for the extraction of spectrogram feature fusion. After that, the approximate optimal feature subset is obtained for overcoming the shortage of recognition ability differences between adjacent frames, which often appears in the traditional feature fusion method. Furthermore, as for the cross-corpus cases, a modified DBN network model (so-called DBN22) is designed for spectrogram features and acoustic features fusion. In Section 4, proven by simulation experiments on four databases, the features of proposed fusion method effectively improve the performance of SER system on cross-corpus.

2. Spectrogram Feature Extraction Based on Selective Attention

Spectrogram, namely, speech spectrum diagram, is based on the time domain signal processing, which has the horizontal axis representing time, the vertical axis representing the frequency, and the depth of the midpoint chart color representing the strength of the corresponding signal. Spectrogram is a communication between the time domain and frequency domain, which reflects the correlation of the two domains. Because spectrogram is the visual expression of the time-frequency distribution of the speech energy [11], it contains characteristic information, such as energy and formant. In our research, based on STB/Itti model [12], selective attention features on the orientation, color, and brightness of spectrogram are extracted as new characteristics for SER. Meanwhile, the dimension reduction and optimization for the features are conducted by proposed DBN model. And then an improved kernel learning K-nearest neighbor algorithm based on feature line centroid (kernel-KNNFLC) [13] classifier is carried on the experiment. The results show that the extracted features possess more powerful emotion recognition ability than their contrast. The spectrogram feature extraction process in SER with selective attention mechanism and DBN is shown in Figure 1.

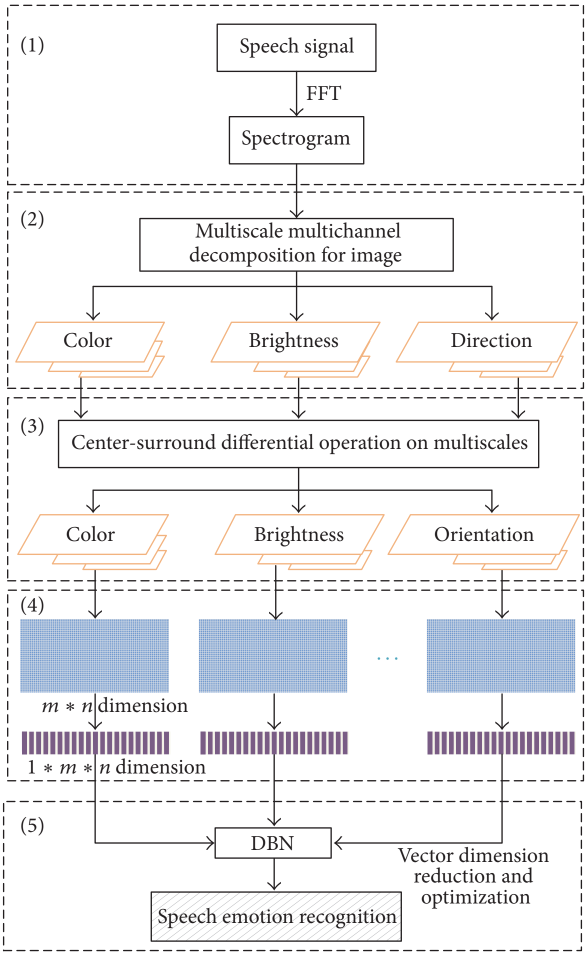

(1)
Speech signal
FFT
Spectrogram
(2)
Multiscale multichannel decomposition for image
Color
Brightness
Direction
(3)
Center-surround differential operation on multiscales
Color
Brightness
Orientation
(4)
$m * n$ dimension
$1 * m * n$ dimension
(5)
DBN
Vector dimension reduction and optimization
Speech emotion recognition

FIGURE 1: Spectrogram feature extraction process in SER.

2.1. Spectrogram Feature Extraction. The computation formula of spectrogram is as follows:

$$L = |Y| = \left| \sum_{n=0}^{N-1} s(n)\,\omega(n)\,e^{-j(2\pi/N)kn} \right|, \quad k \in [0, N]. \tag{1}$$

$s(n)$ represents the input signal, $\omega(n)$ represents the hamming window function, and N is the window length. Figure 2 is the spectrogram extracted from a piece of the speech labeled "aggressive" emotion in ABC corpus.

2.2. Gaussian Pyramid Decomposition. Based on the mechanism of selective attention, the area is easy to get the attention of people in a picture, which usually has strong difference compared to the surrounding area [14]. Multiscale multichannel filtering can be resolved by convolution operation with the linear Gaussian kernel. A 6×6 Gaussian kernel is used in this paper. The resulting image formula after Gaussian Pyramid Decomposition (GPD) is as follows:

$$I(\sigma + 1) = \frac{I(\sigma)}{2}, \quad \sigma = [0, 1, 2, 3, 4, 5, 6, 7, 8], \tag{2}$$

where σ represents the layer number and $I(\sigma)$ is the σ layer of the image after decomposition, in which $I(0)$ is the original

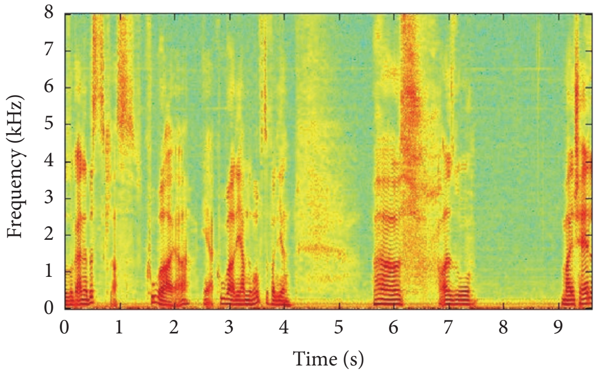


Figure 2: Spectrogram sample graph.

image. After the multiscale multichannel filtering, feature extractions are conducted of each scale image in orientation, color, and brightness, and then sequence images are formed, respectively.

In the retinal cone photoreceptors response level, the model is trichromatic mechanism. However, in the process of choosing messages for center selection in the brain, it changes into 4 primary mechanisms. As a result, 4 primary channels are defined in Itti model. Therefore, the antagonism of the R-G and B-Y colors could be used to simulate the saliency contribution which is made by the colors to images. And then the computation formula is

$$\begin{aligned} P_{R\text{-}G}(\sigma) &= \frac{(r-g)}{\max(r,g,b)}, \\ P_{B\text{-}Y}(\sigma) &= \frac{(b-\min(r,g))}{\max(r,g,b)}. \end{aligned} \tag{3}$$

In formula (3) r, g, and b, respectively, represent the three primary colors: red, green, and blue. Here are 16 GPD images based on the extracted color features of the different scale images.

The GPD images of brightness features are obtained after calculating the average of the normalized r, g, and b:

$$P_I(\sigma) = \frac{(r+g+b)}{3}. \tag{4}$$

Here are 8 GPD images of brightness.

The 2D Gabor directional filter could be used to simulate the directional selection mechanism of the retina [15]; therefore, we can use its convolution with the GPD of brightness feature to get the GPD images of local orientation feature. It has been proven that the angle $\theta \in (0^\circ, 45^\circ, 90^\circ, 135^\circ)$ can be used to represent the orientation feature:

$$P_\theta(\sigma) = \left\| P_I(\sigma) * G_0(\theta) \right\| + \left\| P_I(\sigma) * G_{\pi/2}(\theta) \right\|. \tag{5}$$

The corresponding formula is as follows:

$$G_\psi(\theta) = \exp\left(-\frac{x'^2 + \gamma^2 y'^2}{2\delta^2}\right)\cos\left(2\pi\frac{x'}{\lambda} + \psi\right). \tag{6}$$

γ is the orientation rate which has a value of 1; δ and λ, respectively, represent the standard deviation and the wavelength which have the value of 7/3 pixels and 7 pixels; ψ is the phase and $\psi \in \{0, \pi/2\}$. 32 GPD images of orientation feature could be obtained by using a total of 8 scales and 4 directions, 2D Gabor.

2.3. Features Obtaining and Matrix Reconstruction. Relying on the color and brightness features of GPD extracted previously, they cannot attract the selective attention insufficiently, which also needs the difference contrast of image characteristics. These features compared with the traditional acoustic global features, properties, have better characterization of different speech sample sources (language, speakers, including noise, etc.) in which emotional information is implied. In our research the center-surround is applied to the computing method of calculation [16]. Experimental results show that this center-surround method brings the model more reliable robustness in cross-corpus SER. After the calculation of contrastive feature vector, the gist feature images could be obtained based on the merger strategy (local iterative normalized).

$$FM_l(\sigma_c, \sigma_s) = N\left(\left|P_l(\sigma_c) - P_l(\sigma_s)\right|\right), \tag{7}$$

where $l \in \{R\text{-}G, B\text{-}Y, I, 0^\circ, 45^\circ, 90^\circ, 135^\circ\}$ represents the kinds of gist feature images which are a total of 7, including the R-G and B-Y 2 kinds of color features, one kind of brightness features, and four kinds of orientation features; $\sigma_c \in \{2,3,4\}$ is the central scale of Gaussian pyramid; and $\sigma_s = \sigma_c + d$ is the surrounding scale, among which $d = \{2,3\}$. $N(\cdot)$ represents the merger strategy with local iterative normalized [17]. Finally we received 12 color-contrast, 6 brightness-contrast, and 24 orientation-contrast feature images. The extracted gist feature images based on the speech samples are shown in Figure 3.

A feature image is lot into m lines and n columns, forming total $m * n$ subregions. Then each subregion is replaced by its mean. Furthermore, the images are normalized to a $m * n$ feature matrix, so that a low resolution feature matrix of image is used to describe the whole spectrogram. The mathematical representation of the feature matrix is as follows:

$$\begin{gathered} FD_i(p,q) = \frac{mn}{vh} \sum_{g=pv/n}^{(p+1)v/n-1} \sum_{f=qh/m}^{(q+1)h/m-1} FM_i(g,f), \\ p \in [0, n-1], \quad q \in [0, m-1], \end{gathered} \tag{8}$$

among which FM_i is feature image and FD_i is corresponding feature matrix, $i \in [1, 42]$. Here, m is 4 and n is 5. h and v represent the height and width of the feature image, respectively. And then, the characteristic matrix obtained is reconstructed to a $1 \times mn$ vector, in which the feature performance on the cross-corpus SER will be validated in the subsequent experiments.

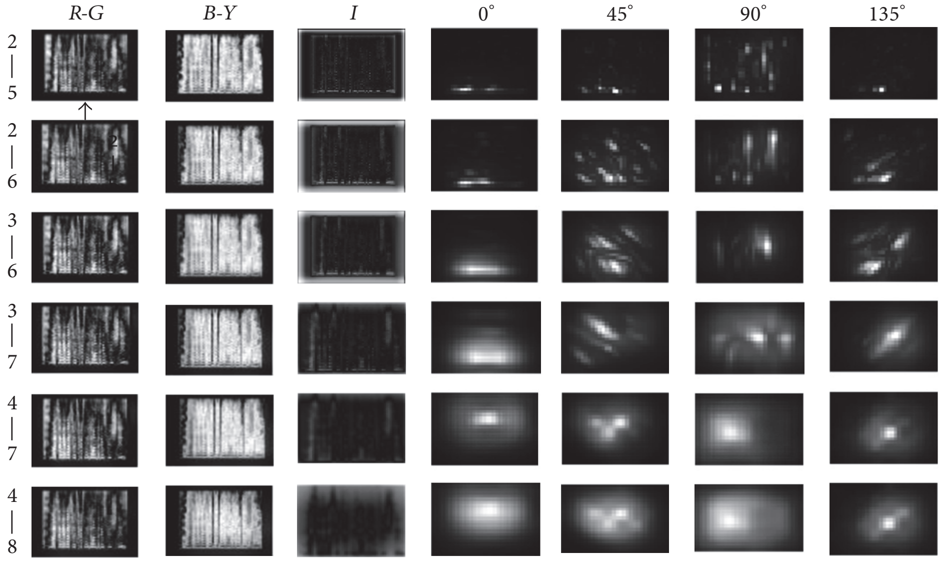


Figure 3: Gist feature images based on spectrogram.

3. DBN Feature Fusion Model for SER

The Deep Belief Nets model rooted in statistical mechanics, which is described through the energy function and probability distribution function. Energy function can reflect the stability of the system. When the system in an orderly state and the probability distribution are intense concentrated, the energy of the system is small. Conversely, if the system is in disorder and the probability distribution is uniform, the energy of system may be larger. DBN model is formed in a multilayer stack RBM, just like constructing a building. The RBM is accumulated in layers and is evaluated one by one from the bottom to the top. The training of each layer is independent while the top RBM has self-associative memory according to the information from the lower. Eventually the Error Back Propagation (BP) algorithm is applied to fine-tune weight. At the top of DBN, the kernel-KNNFLC classifier is connected for classification.

3.1. Restricted Boltzmann Machine in DBN. Boltzmann Machine (BM) is a kind of random neural network model, which is made up of two parts: visible and hidden layer. Although BM has strong unsupervised-learning ability and could learn the complex rules in the data, the training time is tremendous long. To solve this problem, Smolensky proposed the RBM, the structure of which is as shown in Figure 4.

The model figure reveals that it is inexistence of internal connection between the visible layer and hidden layer of RBM, which has the property: if the state of the hidden units is given, activated units in visible layer are conditionally independent, so that if the unit number of hidden and layers visible of RBM is m and n, respectively, which state vectors are h and v, according to a given state (v, h), the energy could be defined as follows:

$$E(v, h \mid \theta) = -\sum_{i=1}^{n} a_i v_i - \sum_{j=1}^{m} b_j h_j - \sum_{i=1}^{n} \sum_{j=1}^{m} v_i W_{ij} h_j, \quad (9)$$

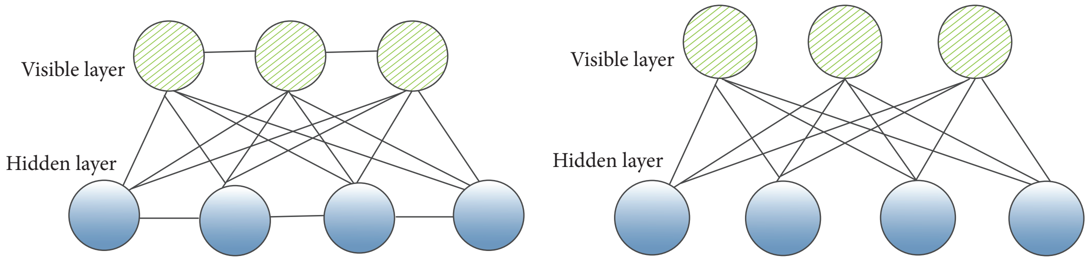
in which a_i and b_j are the values of bias of visible unit i and hidden unit j, respectively, and W_{ij} represents the connection weight of j and i. Here $\theta = \{W_{ij}, a_i, b_j\}$ is used as the whole parameter set in RBM. When the parameter set is determined, the joint probability distribution of (v, h) could be obtained according to formula (10), as shown in the following formula:

$$P(v, h \mid \theta) = \frac{e^{-E(v,h|\theta)}}{Z(\theta)}, \quad Z(\theta) = \sum_{v,h} e^{-E(v,h|\theta)}. \quad (10)$$

Here $Z(\theta)$ is called the partition function. Because, with the unit being given by RBM, the activated states between each hidden unit are independent, if it is in a given unit state, the activation probability of j and i could be obtained as follows:

$$\begin{aligned} P\left(h_j = 1 \mid v, \theta\right) &= \sigma\left(b_j + \sum_i v_i W_{ij}\right), \\ P\left(v_i = 1 \mid h, \theta\right) &= \sigma\left(a_i + \sum_j h_j W_{ij}\right). \end{aligned} \quad (11)$$

3.2. The Fast Learning Algorithm Based on Contrast Divergence. Gibbs Sampling algorithm is based on Markov Chain Monte Carlo (MCMC) strategy [18]. By getting a conditional probability distribution of the weight, which can begin from any state, the algorithm implements iteration sampling in turn for each component. Gibbs Sampling method is used to obtain the probability distribution, which is often necessary to employ a lot of sampling steps. In particular within the high-dimension data, the training efficiency of model may be greatly influenced. Therefore, Hinton proposed a fast learning algorithm of RMB called contrastive divergence (CD) [19]. Unlike Gibbs Sampling, this method (CD) uses the training data to initialize and just needs k steps (usually $k = 1$) to



(a) BM

(b) RBM

Figure 4: BM and RBM model.

gain a satisfactory approximation. At the beginning of the CD algorithm, the visible unit state is set to a training sample, and then formula (12) is used to calculate the unit state of the hidden layer. After that, the probability of the *i*st unit hidden values equaling 1 could be calculated according to formula (12). Furtherly, refactoring of visible layer is obtained.

The task of training RBM is to get parameters θ. The logarithm likelihood function is obtained through the training set that is maximized for parameter set θ, which may fit the given training data. If the number of training samples is T, there are

$$\theta^{*} = \arg\max_{\theta} L(\theta) = \arg\max_{\theta} \sum_{t=1}^{T} \log P\left(v^{(t)} \mid \theta\right). \tag{12}$$

Then, the Stochastic Gradient Ascent method is used to find the optimal parameters maximizing equation (12):

$$\begin{aligned} \frac{\partial}{\partial\theta} L(\theta) &= \frac{\partial}{\partial\theta} \sum_{t=1}^{T} \log \sum_{h} P\left(v^{(t)}, h \mid \theta\right) = \frac{\partial}{\partial\theta} \\ &\cdot \sum_{t=1}^{T} \log \frac{\sum_{h} e^{[-E(v^{(t)},h|\theta)]}}{\sum_{v}\sum_{h} e^{[-E(v^{(t)},h|\theta)]}} \\ &= \sum_{t=1}^{T} \left(\left\langle \frac{\partial}{\partial\theta} \left(-E\left(v^{(t)}, h \mid \theta\right)\right) \right\rangle_{P(h|v^{(t)},\theta)} \right. \\ &\quad \left. - \left\langle \frac{\partial}{\partial\theta} \left(-E(v, h \mid \theta)\right) \right\rangle_{P(v,h|\theta)} \right). \end{aligned} \tag{13}$$

In formula (13) $\langle\cdot\rangle_{P}$ is calculating mathematical expectation of the distribution P. The first item of the formula can be determined by the training sample, while $P(v, h \mid \theta)$ in the following item need to get the joint probability distribution of visible and hidden units first. And then, for calculating the distribution function $z(\theta)$, which cannot be directly calculated, the sampling method (such as Gibbs Sampling) is introduced to approximate the related value. When using "data" as the tag of $P(h \mid v^{(t)}, \theta)$ and "model" as $P(v, h \mid \theta)$, the offset on visible and hidden units of formula (13) is a_i and b_j, respectively, and the connection weight is W_{ij}. Then a partial derivative is available:

$$\begin{aligned} \frac{\partial}{\partial a_i} \left[\log P(v \mid \theta)\right] &= \langle v_i \rangle_{\text{data}} - \langle v_i \rangle_{\text{model}}, \\ \frac{\partial}{\partial b_j} \left[\log P(v \mid \theta)\right] &= \left\langle h_j \right\rangle_{\text{data}} - \left\langle h_j \right\rangle_{\text{model}}, \\ \frac{\partial}{\partial W_{ij}} \left[\log P(v \mid \theta)\right] &= \left\langle v_i h_j \right\rangle_{\text{data}} - \left\langle v_i h_j \right\rangle_{\text{model}}. \end{aligned} \tag{14}$$

Based on the criteria of formula (14), the method of Stochastic Gradient Rise is used to maximize the value of the logarithm likelihood function on the training data. Therefore, the updating criteria of parameters are

$$\begin{aligned} \Delta a_i &= \varepsilon \left(\langle v_i \rangle_{\text{data}} - \langle v_i \rangle_{\text{recon}} \right), \\ \Delta b_j &= \varepsilon \left(\left\langle h_j \right\rangle_{\text{data}} - \left\langle h_j \right\rangle_{\text{recon}} \right), \\ \Delta W_{ij} &= \varepsilon \left(\left\langle v_i h_j \right\rangle_{\text{data}} - \left\langle v_i h_j \right\rangle_{\text{recon}} \right), \end{aligned} \tag{15}$$

in which ε is the learning rate and $\langle\cdot\rangle_{\text{recon}}$ represents the distribution of model defined after one step refactoring.

From the above contents, the training procedure of RBM algorithm is divided into a few steps:

(1) Firstly, initialization of RBM is necessary. Thus, mainly the following contents are included: sample training set S; the number of neurons n_h contained in hidden layer, the number of visible layer neurons n_v; the connection weight W_{ij} of visible and hidden layer; the unit biases a_i and b_j of visible and hidden layer; the learning rate ε and the training cycle J; the number of the algorithm steps k.

(2) Rapid sampling is carried out based on the CD-k algorithm. Furtherly, according to the updates of each parameter, the value of parameter set is refreshed.

(3) The sampling process is repeated in the whole training period, until the convergence of formula (12).

3.3. DBN21 and DBN22 Models. According to the RBM, two kinds of DBN models, respectively, DBN21 and DBN22,

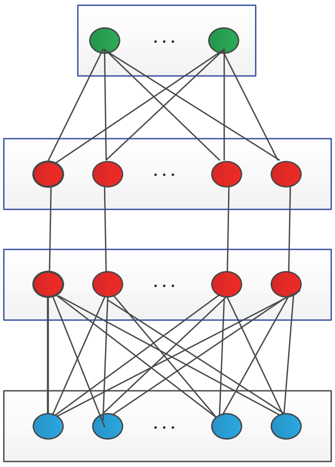


Figure 5: DBN21 model.

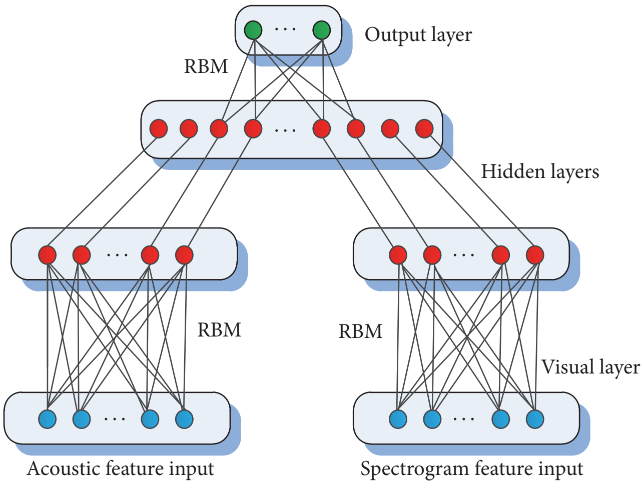


Figure 6: DBN22 model.

are structured for SER experiments on cross-database. As shown in Figures 5 and 6, (1) the DBN21 model is proposed for separate layer feature fusions with spectrogram features and traditional acoustic features (by the international general extracting method mentioned in Section 4.1.3); (2) the DBN22 model is constructed for integration of the spectrogram and traditional acoustic features in the feature layer. Because the speech emotion features extracted for SER experiments are real number data, it is not appropriate to apply the binary RBM for modeling. As a result, we chose the Gaussian-Bernoulli RBM (GRBM) [20] to build the bottom structure. The energy function of GRBM is

$$E(v,h\mid\theta)=-\sum_{i=1}^{V}\sum_{j=1}^{H}\frac{v_i}{\sigma_i}\omega_{ij}h_j+\sum_{i=1}^{V}\frac{(v_i-b_i)^2}{2\sigma_i^2}-\sum_{j=1}^{H}a_jh_j. \quad (16)$$

Formula (16) represents the Gaussian noise variance of the visible neurons. Due to the change of the energy function, conditional probability is also changed, which should be amended as

$$p(h_j=1\mid v)=\text{logistic}\left(a_j+\sum_{i=1}^{V}\omega_{ij}\frac{v_i}{\sigma_i}\right), \quad (17)$$

$$p(v_i=1\mid h)=N\left(b_i+\sigma_i\sum_{j=1}^{H}\omega_{ij}h_j,\sigma_i^2\right). \quad (18)$$

As shown in Figure 6, the input visual, hidden, and output layers are represented as blue, red, and green colorized rounds, respectively. The restructures of models show that the DBN22 input has two representations (although in practice DBNN21 and DBN22 have the same structure once the feature vectors are combined). The training process of the DBN22 network model is conducted in accordance with the following steps.

(1) The initialization of unsupervised learning is needed at the beginning of training. The initialization process is step by step completed in each layer by multiple RBM in accordance with the order of the bottom-up.

First of all, the feature vector extracted from the traditional features is considered as the visual layer of the first left side in the RBM; the spectrogram feature vector is considered as the visual layer of the first right side in the RBM. Then, the CD-1 algorithm is applied to training for the weight of each layer, denoted by LW_1 and RW_1. According to the weights obtained and the input visible layers, the weighted summations are conducted on all the input nodes. Then, the hidden layers LH_1 and RH_1 could be obtained by mapping [21].

After that, LH_1 and RH_1 are the input of visual layer in the second RBM. Also after CD-1 training, connection weight W_2 can be obtained. Then, the hidden layer H_2 is gained according to the input visual layer and weight W_2.

(2) The Deep Belief Networks are constituted. The trained RBMs in top-bottom order are overlapped layer by layer as shown in Figure 6. The uppermost level of RBM is in the form of a two-way connection while the others are connected by top-bottom.

(3) The kernel-KNNFLC classifier is added to the top of the above for classification.

(4) The network weights are fine-tuned. Before the final network parameters being obtained, the fine-tuning is necessary by the training results and BP algorithm so that the weights may be more accurate.

The training process of DBN21 model is similar to DNB22, while the bottom layer RBM only has the left half of DBN22. The data generation process of DBN is through the top RBM with Gibbs Sampling and completed by transferring from top to bottom. The Gibbs Sampling of the top RBM is divided into multiple alternate processes, which makes the sample distribution obtained balanced. Then, the data are generated by top-bottom DBN network. This way effectively saves feature information of cross-corpus samples, so as to improve the robustness of the SER system. The operation of weights adjustment is conducted after the pretraining. Then, based on the method of Error Back Propagation, the tag data are used to fine-tune the weights. This strategy searches the weight space locally in the process of running, which could accelerate the training speed effectively.

4. Experimental Results and Analysis

4.1. Experimental Preparation

4.1.1. Settings of Experiments. In this section, the fusion experiments are divided into three parts. First of all, DBN21 model is used to carry out layer fusion for SER across the databases, in which features are traditional acoustic (see Section 4.1.3). Then, the results of the experiments are contrasted with the traditional features without DBN fusion and this experiment group is marked as *Fusion 1*. DBN21 model is then used to extract spectrogram features of layer fusion based on selective attention mechanism. Also the results are contrasted to the features without DBN fusion, which demonstrates the cross-corpus SER ability. This group of experiments is marked as *Fusion 2*. Finally the DNB22 model is proposed to fuse the traditional and spectrogram features. The experimental results are compared with Fusion 1 and Fusion 2. To prove that DBN22 possesses the significant improvement of performance on features fusion for SER, this group of experiments is marked as *Fusion 3*.

4.1.2. Database Settings. The selection of appropriate emotional databases is also an important part of speech emotion recognition. In our research, we chose one common speech emotion database: ABC (Airplane behaviors Corpus) which is recorded in German [22]. Moreover, the Deep Learning technology is suitable for the situation with a huge number of data sets, while the international classic databases usually have less samples. Meanwhile, in order to verify the effect of the fusion method proposed on Chinese speech databases, two Chinese corpora which are widely researched in China domestic are introduced and combined. The following are the brief introductions of the 3 databases, respectively.

ABC is obtained on a holiday aircraft flight in the background of prerecorded announcement played. The flight contains 13 upcoming trip scenarios and 10 return scenarios. Eight targeted passengers are chosen to get through the set condition: false meals, aircraft navigation turbulence, sleep, and conversation with the neighbors. During this process 11.5 hours of video and 431 voices with 8.4-second length are recorded. Finally the collected segments are independent analyzed by three professional researchers, and then the selected samples are labeled in accordance with the "aggressive," "amused," "excited," "strained," "neutral," and "tired" 6 kinds of emotional categories.

Chinese corpora used in our SER experiments consist of two databases which are recorded by induced and acted speech, respectively. One of them is the Chinese Database (CNDB) created by the Key Laboratory of Underwater Acoustic Signal Processing of in Southeast University. It consists of two parts: practical speech emotion database [23] and Whisper emotion database [24]. The statement materials of practical speech emotion database are recorded by performers with histrionic or broadcast experience (8 males and 8 females, aged between 20 to 30 years, without a recent cold, standard Mandarin). The recording environment is indoors quiet. In order to guarantee the quality of the emotional corpora, the subjective listening evaluation is carried out. The statements selected with more than 85% confidence coefficient are in total 1410 from the male performers and 1429 from the female performers, including six basic emotional categories: "raged," "fear," "joy," "neutral," "sad," and "surprise." The Whisper database contains "happy," "angry," "surprise," "grieved," and "quiet" such five kinds of emotions. Then, the speech materials are divided into three types: word, phrase, and long sentence. The corpus contains 25 words, 20 phrases, and 6 long sentences for each emotion category. Each speaker repeats the whispers 3 times and with normal voice for 1 time (for later comparison), forming a total of 9600 statements. The research of whispered speech database has great significance: further improving the ability of human-computer interaction, combining with the semantic to judge the inner activities of speakers, and helping computers really understand the operators' thoughts, feelings, and attitudes. The analysis and processing of the emotional characteristics of whispered speech signals have important meaning of judgment and simulation of the emotion status from speakers in theory and application.

According to the recording criterion of corpora, the two Chinese databases are merged, ultimately forming 7839-statement CNDB Chinese corpora. The recording employs mono, 16-bit quantitative, and 11.025 kHz sampling rate. The selection of statements follows two principles: (1) the statements selected do not contain a particular emotional tendency; (2) the statements must have high emotional freedom, which could exert different emotions on the same statement.

Another Chinese corpus is the Speech Emotion Database of Institute of Automation Chinese Academy of Sciences (CASIA) [25]. The language of the database is Chinese, made by four actors. Database contains 1200 statements and is divided into 6 categories of emotions: "angry," "fear," "happy," "neutral," "surprise," and "sad."

In order to verify the effectiveness of the method proposed in this paper, in each group of experiments two kinds of schemes (*Case I and Case II*) are adopted, respectively, for testing. According to the theory of Emotion Wheel [26], mutual or similar 4 basic emotions in the three chosen databases, "angry (aggressive, raged)," "happy (joy, amused)," "surprise (excited, amazed)," and "neutral," are chosen for experimental evaluation. Because the DBN model could show the effectiveness of the fusion under the condition of large amount of data, we merge 3 Chinese speech emotion corpora into one called *Mandarin Database*. In *Case I*, Mandarin Database is as the training data set (known label), while ABC (unknown label) is as the testing set. The cross-corpus SER of this scheme adopts rotation experiment testing method: the data set is divided into 10 portions, in which the proportion of training/testing is 9 : 1. The set of this 10-fold cross-validation is intended to optimize the parameters within the source corpus [10]. After the cross-validation, the averages are obtained as the results for the cross-corpus experiments. In *Case II*, ABC corpus (known label) is as the training set, while Mandarin Database (unknown label) is as the testing set. Because the number of samples in German ABC corpus is less, for the sample balance of corpus in the process of SER evaluation, we join a part of the Mandarin samples (45% is the optimum by testing, known label) into ABC samples which are as the training set. Then the remaining Mandarin samples (55% corpora, unknown label) are as the testing set.

4.1.3. Settings of Feature Parameters and Classifier. With regard to the *traditional acoustic global features*, the common tool openSMILE is used for feature extraction whose tool number is set as 1 [27]. Then the feature set in the Interspeech 2010 SER competition [28], which contains a total of 1582 dimensions features, is introduced in our experiments. 38 acoustic low-level descriptors (LLDs) and their first-order differences are contained in the set. Through statistics of 21 class functions on the LLDs (16 features with 0 information are removed), we add the numbers and lengths of F0 to the feature set. In the contrast group without fusion, the feature sets extracted directly carried out the LDA dimension reduction, making its dimension match the fusion experimental group.

In this paper, the recognition experiment employs kernel-KNNFLC classifier, which may verify the SER ability of the fusion features. According to the gravity center criterion, Kernel-KNNFLC learns the sample distances and improves the K-neighbor with kernel learning method. The classifier optimizes the differentiation between kinds of the emotional feature vectors, which solves the problem about huge calculation caused by the features of prior samples. Based on the cross-corpus samples trained, the recognition model is established and then the different emotional categories are distinguished. The Gaussian Radial Basis kernel Function (RBF) is used in the classifier: $k(x, y) = \exp(-\|x - y\|^2/2\sigma^2)$, in which $\sigma = 4$. The KNNFLC classifier based on kernel has stable SER performance on high-dimensional data. In addition our experiments use 4 kinds of speech emotions, so the dimension dropped to 3 for achieving the best recognition rate. This is due to the solving of the generalized eigenvalue principle: the optimization is achieved when the minimum number of features is solved. With the K-nearest neighbor algorithm based on feature line centroid, the kernel function of RBF is improved and the optimum value is $K = 3$ [13].

4.2. Traditional Global Acoustic Feature Fusion Experiment (Fusion 1). The purpose of *Fusion 1* is comparing the fusion feature with DBN21 to features without DBN, so that the cross-corpus recognition performance of fusional traditional features could be revealed. The extracted acoustic features are as the input of DBN21 model. Then the optimization process is carried out by DBN. After that, combined with the kernel-KNNFLC classifier introduced before, the emotion recognition missions are proceeded on cross data sets.

The settings of RBM learning rate should be moderate, because too big or too small rates will both increase the reconstruction error. GRBM learning rate of bottom layer in *Fusion 1* is set as $\varepsilon = 0.001$ and training cycle is set as $J = 200$. The upper layer RBM is set as $\varepsilon = 0.01$ and $J = 70$. Since the numbers of visible layer unit and input dimension are the same, the input units number of visible layer in the experiment is $n_v = 60$ and the number of upper hidden units is set as $n_h = 20$. The weight is set according to the Gaussian random vector $N = (0, 0.01)$. The visible and hidden unit biases are $a_i = 0$ and $b_j = 0$. Because the number of hidden units in the middle layer may influence systemic performance, therefore we enumerate the 6 units' numbers of contrast experiments: 250, 500, 750, 1000, 1250, and 1500, in order to determine the optimal number of hidden units. The experimental comparison results are shown in Figure 7.

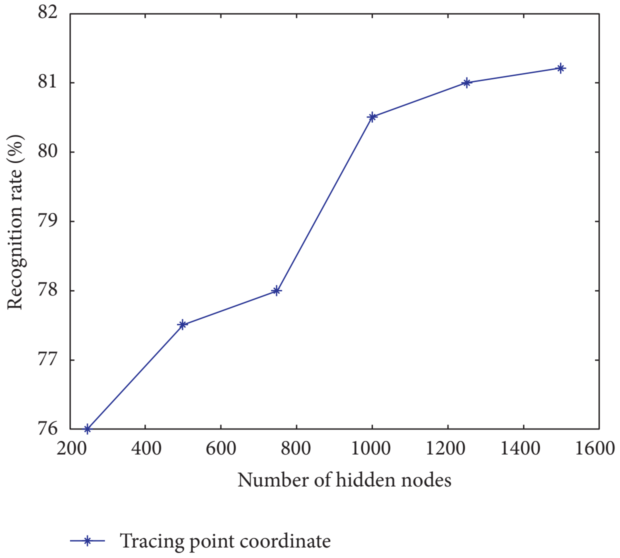


Figure 7: The influence of hidden node numbers to recognition rates *Fusion 1*.

Figure 7 reveals that, along with the increase of the hidden nodes number, the recognition efficiency of system is growing. However, the increased number of nodes may cause the extra amount of calculation. It is clear that when the nodes number rises from 750 to 1000, the recognition rate has greatly improved, and then it is steady. Hereby considering the time consumed and accuracy, the number of hidden nodes in *Fusion 1* is set as 1000.

The speech emotion recognition experiments are carried out through the DBN21 model proposed. In our testing strategy, the ABC and Mandarin databases are cross-validated, which is to verify the robustness of algorithm proposed under the cross-corpus SER task. Toward each kind of emotion in two cross-database cases, recognition rates of the traditional features which are before and after fusion are shown in Table 1.

The experimental results indicate the traditional features after optimization by DBN. The emotion recognition ability has greatly ascended, which rises by 4.6% on the average recognition rate. It reveals that the DBN model proposed in feature layer is effective for SER feature fusion research. After training on ABC and Mandarin databases in *Case I*, SER rates of ABC testing set on Mandarin training set reach 52.2%. Among them "happy" and "neutral" reach over 63% as the highest, whose recognition effect is superior to *Case II*. It related to the many similar types of samples in

Table 1: Recognition results (%) of *Fusion 1* in two cross-corpus cases.

Cross-corpus scheme	Happy	Angry	Surprise	Neutral	Average
Case I					
DBN feature fusion	63.7	38.6	41.6	64.9	52.2
Without fusion	52.8	34.3	39.7	63.7	47.6
Case II					
DBN feature fusion	57.0	37.5	40.8	62.7	49.5
Without fusion	50.1	29.7	35.3	60.4	43.9

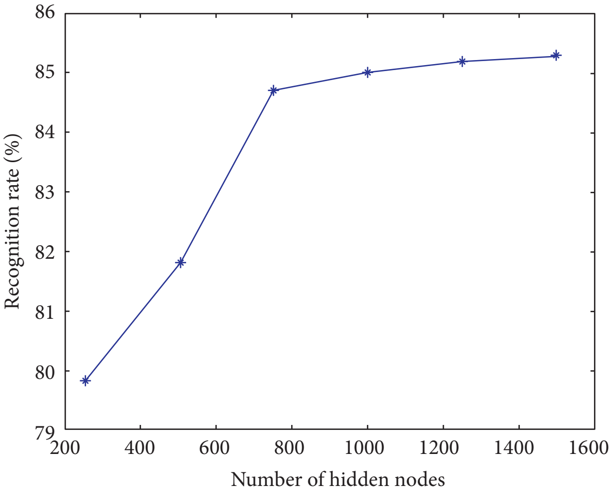
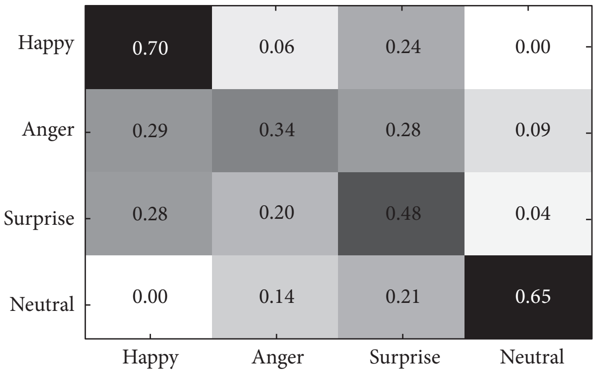


Figure 8: The influence of hidden node numbers to recognition rates in *Fusion 2*.



	Happy	Anger	Surprise	Neutral
Happy	0.70	0.06	0.24	0.00
Anger	0.29	0.34	0.28	0.09
Surprise	0.28	0.20	0.48	0.04
Neutral	0.00	0.14	0.21	0.65

Figure 9: Cross-corpus SER confusion matrix in *Fusion 2* with *Case I*.

3 training corpora. Through the great amount of emotional data training in various categories by Deep Learning, the model becomes highly mature while the matching degrees of the traditional emotional categories with high inner-class discrimination ("happy," "neutral") are high. The comparison of two experiment schemes shows that the small amount of training samples in ABC gives rise to information insufficiently. Thus, further testing in large data corpus may cause undermatching with model.

4.3. Spectrogram Feature Fusion Experiment (Fusion 2). Main aim of *Fusion 2* is to validate the feature effectiveness of spectrogram on cross-corpus. The feature sets abstracted are based on selective attention mechanism introduced in this paper. In order to reflect the promotional recognition performance on cross-databases, the experimental results after DBN21 fusion are compared with traditional features in *Fusion 1*.

The same as Fusion 1, GRBM learning rate of bottom layer in Fusion 2 is set as $\varepsilon = 0.001$ and training cycle is set as $J = 200$; the upper layer RBM is set as $\varepsilon = 0.01$ and $J = 70$. But the input units number of visible layer here is $n_v = 291$ and the number of upper hidden units is set as $n_h = 80$. The weight is set also as $N = (0, 0.01)$; meanwhile, the visible and hidden unit biases are $a_i = 0$ and $b_j = 0$. In consideration, the number of hidden nodes in layer RBM may cause the influence of system performance; this experiment still needs the discussion of the numbers of hidden nodes. The analysis of node numbers in traditional features is as in Figure 8.

As shown in the relationship in Figure 8, it is different from the traditional feature experiment; the recognition efficiency of spectrogram features has greatly promoted at 750 hidden nodes of point position. This is due to the traditional acoustic features compared to the spectrogram ones, which possess much higher input dimensions, so that, in the spectrogram feature fusion experiments, the number of hidden nodes in bottom RBM is set to 750.

According to the SER fusion model in feature layer, which is based on selective attention as shown in Figure 1, the cross-corpus experiments are carried out. In *Fusion 2*, ABC and Chinese databases are crossed training for cross-corpus testing. The SER confusion matrix in *Case I* by DBN21 fusion model is as shown in Figure 9.

It could be seen from the experimental results that the spectrogram features extracted integrally have strong ability of speech emotion recognition. When compared to traditional features, the spectrogram exhibits advantages in dealing with the cross data set tasks. This is because the traditional features contain only the local traits of common speech processing field, while the spectrogram, abstracted from the aspects of time and frequency domains, contains information between adjacent frames and the temporal features which can make up for a lack of global features. In spectrogram features, meanwhile, compared to traditional global features, the cascade vectors possess higher dimensions which contain more information for characterizing the emotions. Among them, the "happy," "angry," and experimental results improve significantly compared with the traditional fusion features. It reveals that spectrogram features have a relatively better distinction effect on the emotion category with high frequency domain correlation dependence.

4.4. DBN22 Feature Fusion Experiment on Cross-Corpus (Fusion 3). In experiment *Fusion 3* with DBN22 model, we conduct feature layer fusion of traditional global acoustic features and spectrogram features based on selective attention. After that the kernel-KNNFLC is combined with SER system for cross-corpus experiments. This method integrates image characteristics and acoustic characteristics, which is a novel

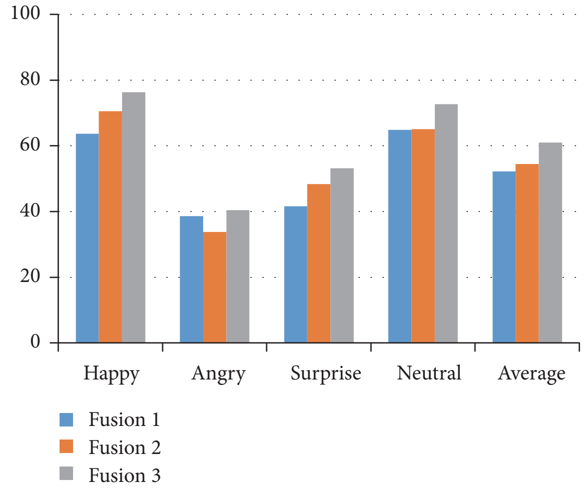


Figure 10: Recognition rates of 3 fusion models.

new attempt for data source extension in the field of speech emotion recognition. At the same time, the experiment may demonstrate that the features with thus fusion method have significant help for improving SER performance in cross-databases.

The settings of RBM in *Fusion 3* are as follow: GRBM learning rate of bottom layer is set as $\varepsilon = 0.001$ and training cycle is set as $J = 200$. The upper layer RBM is set as $\varepsilon = 0.01$ and $J = 70$. Since the numbers of visible layer unit and input dimension are the same, thus the input units numbers of visible layer in acoustic and spectrogram features are $Ln_v = 291$ and $Rn_v = 60$, respectively. The number of upper hidden units is set as $n_h = 100$. The weight is set according to the Gaussian random vector $N = (0, 0.01)$. The visible and hidden unit biases are $a_i = 0$ and $b_j = 0$. Because the number of hidden units in the middle layer may influence systemic performance, according to the two Fusion experiment advance, hidden units numbers in RBM of acoustic and spectrogram features are 1000 and 750, respectively.

After cross-database SER experiment, fusion features of traditional acoustic and spectrogram features are gained based on DBN22 network. Then the recognition results are compared with DBN21 groups in *Fusion 1* and *Fusion 2* by the bar plot (using *Case I* cross-corpus settings) (see Figure 10).

After the analysis of Figure 10, the cross-database recognition efficiency of the fused features in *Fusion 3* is the highest. Specifically "happy," "angry," "surprise," and "neutral" 4 emotional kinds compared with the traditional group rise by 12.6%, 1.8%, 11.6%, and 12.6%, respectively; the promotion of average recognition rate is 8.8%. Relative to the spectrogram features, the results increase by 5.8%, 5.8%, 6.6%, and 5.8%, respectively, and the elevation of average recognition rates up to 6.5%. The fusion by DBN22 of two kinds of features obtains excellent recognition effect in all of the emotion categories. The results benefit from the optimization of feature fusion layer in RBM stack of the DBN network, while there are also the factors of classifier and network parameters' settings. Experiments show that the DBN network model proposed successfully gains the fused features of traditional acoustic characteristics and the information of spectrogram images, which meanwhile effectively improve the cross-corpus efficiency of the SER system.

5. Conclusion

This paper mainly researches the feature layer fusion model on the strength of DBN for speech emotion recognition. First of all, based on the mechanism of selective attention, the system extracts three kinds of spectrogram features with both temporal information and global information, which are used for cross-corpus SER. The spectrogram features introduced solve the problem of information loss by the traditional feature selection methods. Further, it is a supplement to the types of emotional information under the cross-database. Then, the modified DBN models are proposed to reasonably optimize the high-dimension spectral features, to retain the useful information and to improve the robustness of cross-corpus SER system. In the subsequent simulation experiments, the DBN21 and DBN22 models designed are used in the feature layer to fuse the spectrogram and traditional acoustic traits. Furthermore, the experimental results are compared with those of the benchmark models. Through experiments in cross-databases containing three Chinese ones and a general German one, DBN networks with multilayer RBM are proved as robust feature layer fusion models for cross-corpus. Spectrogram traits, at the same time, are validated conducive to boost emotional distinguish ability after feature fusion. In this paper, on the basis of Deep Learning thought, the DBN22 model proposed effectively fuses the spectrogram and traditional acoustic emotion features. This progress realizes the features fusion of various data sources and provides a new direction for further research of SER in cross-corpus.

Competing Interests

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

Acknowledgments

This work has been supported by the national natural science foundation of china (NSFC) under Grants nos. 61231002 and 61375028.

References

[1] S. G. Koolagudi and K. S. Rao, "Emotion recognition from speech: a review," *International Journal of Speech Technology*, vol. 15, no. 2, pp. 99–117, 2012.

[2] S. Ramakrishnan and I. M. M. El Emary, "Speech emotion recognition approaches in human computer interaction," *Telecommunication Systems*, vol. 52, no. 3, pp. 1467–1478, 2013.

[3] E. Marchi, A. Batliner, B. Schuller et al., "Speech, emotion, age, language, task, and typicality: trying to disentangle performance and feature relevance," in *Proceedings of the International Conference on Privacy, Security, Risk and Trust (PASSAT '12) and International Conference on Social Computing (SocialCom '12)*, pp. 961–968, 2012.

[4] C. Parlak, B. Diri, and F. Gürgen, "A cross-corpus experiment in speech emotion recognition," in *Proceedings of the International Workshop on Speech, Language and Audio in Multimedia (SLAM '14)*, pp. 58–61, Penang, Malaysia, 2014.

[5] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: features, classification schemes, and databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, 2011.

[6] S. Kim, P. G. Georgiou, S. Lee, and S. Narayanan, "Real-time emotion detection system using speech: Multi-modal fusion of different timescale features," in *Proceedings of the IEEE 9th International Workshop on Multimedia Signal Processing (MMSP '07)*, pp. 48–51, Crete, Greece, October 2007.

[7] K. Han, D. Yu, and I. Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," in *Proceedings of the 15th Annual Conference of the International Speech Communication Association (Interspeech '14)*, pp. 223–227, Singapore, September 2014.

[8] H. Lee, C. Ekanadham, and Y. Ng A, "Sparse deep belief net model for visual area V2," in *Advances in Neural Information Processing Systems*, pp. 873–880, 2008.

[9] V. Nair and G. E. Hinton, "Rectified linear units improve Restricted Boltzmann machines," in *Proceedings of the 27th International Conference on Machine Learning (ICML '10)*, pp. 807–814, June 2010.

[10] B. Schuller, Z. Zhang, F. Weninger et al., "Selecting training data for cross-corpus speech emotion recognition: prototypicality vs. generalization," in *Proceedings of the 27th International Conference on Machine Learning (ICML-10 '11)*, pp. 807–814, 2011.

[11] T. A. Lampert and S. E. M. O'Keefe, "On the detection of tracks in spectrogram images," *Pattern Recognition*, vol. 46, no. 5, pp. 1396–1408, 2013.

[12] A. Borji, D. N. Sihite, and L. Itti, "Quantitative analysis of human-model agreement in visual saliency modeling: a comparative study," *IEEE Transactions on Image Processing*, vol. 22, no. 1, pp. 55–69, 2013.

[13] X. Zhang, C. Zha, X. Xu, P. Song, and L. Zhao, "Speech emotion recognition based on LDA+kernel-KNNFLC," *Journal of Southeast University (Natural Science Edition)*, vol. 45, no. 1, pp. 5–11, 2015.

[14] O. Kalinli and R. Chen, "Speech syllable/vowel/phone boundary detection using auditory attention cues," Google Patents, 2014.

[15] Y. Sun and R. Fisher, "Object-based visual attention for computer vision," *Artificial Intelligence*, vol. 146, no. 1, pp. 77–123, 2003.

[16] C. Stevens, B. Harn, D. J. Chard, J. Currin, D. Parisi, and H. Neville, "Examining the role of attention and instruction in at-risk kindergarteners: electrophysiological measures of selective auditory attention before and after an early literacy intervention," *Journal of Learning Disabilities*, vol. 46, no. 1, pp. 73–86, 2013.

[17] G. Evangelopoulos, A. Zlatintsi, A. Potamianos et al., "Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention," *IEEE Transactions on Multimedia*, vol. 15, no. 7, pp. 1553–1568, 2013.

[18] A. Smith, A. Doucet, N. de Freitas et al., *Sequential Monte Carlo Methods in Practice*, Springer Science & Business Media, 2013.

[19] G. E. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural Computation*, vol. 14, no. 8, pp. 1771–1800, 2002.

[20] A.-R. Mohamed, T. N. Sainath, G. Dahl, B. Ramabhadran, G. E. Hinton, and M. A. Picheny, "Deep belief networks using discriminative features for phone recognition," in *Proceedings of the 36th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2011)*, pp. 5060–5063, Prague, Czech Republic, May 2011.

[21] Y. Tsao and Y.-H. Lai, "Generalized maximum a posteriori spectral amplitude estimation for speech enhancement," *Speech Communication*, vol. 76, pp. 112–126, 2016.

[22] B. Schuller, D. Arsic, G. Rigoll, M. Wimmer, and B. Radig, "Audiovisual behavior modeling by combined feature spaces," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '07)*, pp. II-733–II-736, Honolulu, Hawaii, USA, April 2007.

[23] C. Huang, Y. Jin, Y. Zhao et al., "Design and establishment of practical speech emotion database," *Acoustic Technologies*, vol. 29, no. 4, pp. 396–399, 2010 (Chinese).

[24] Y. Jin, Y. Zhao, C. Huang et al., "The design and establishment of a Chinese whispered speech emotion database," *Technical Acoustics*, no. 1, pp. 63–68, 2010.

[25] Institute of Automation Chinese Academy of Sciences, The selected Speech Emotion Database of Institute of Automation Chinese Academy of Sciences (CASIA), 2010, http://www.chineseldc.org/resource_info.php?rid=76.

[26] T. Bänziger, V. Tran, and K. R. Scherer, "The Geneva Emotion Wheel: a tool for the verbal report of emotional reactions," in *Poster Presented at ISRE*, vol. 149, pp. 149–271, Bari, Italy, 2005.

[27] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM International Conference on Multimedia (MM '10)*, pp. 1459–1462, Firenze, Italy, October 2010.

[28] B. Schuller, S. Steidl, A. Batliner et al., "The INTERSPEECH 2010 paralinguistic challenge," in *Proceedings of the International Speech and Communication Association (INTERSPEECH '10)*, pp. 2794–2797, Makuhari, Japan, 2010.

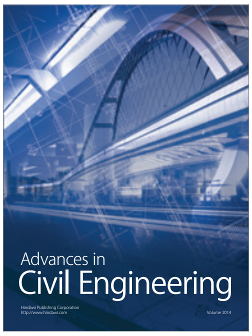
Journal of
Engineering

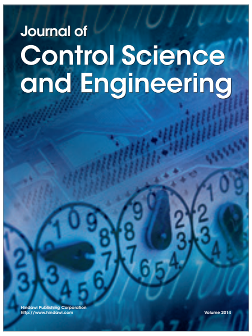
The Scientific World Journal

International Journal of
Rotating Machinery

Journal of
Sensors

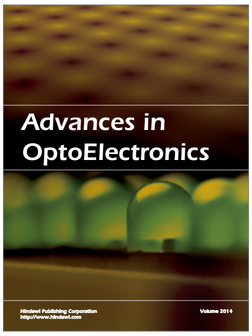
International Journal of
Distributed Sensor Networks


Advances in
Civil Engineering


Journal of
Control Science and Engineering

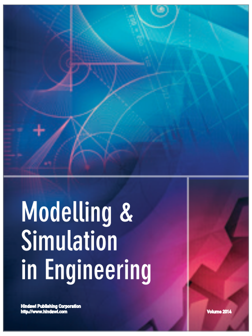
Journal of
Robotics

Journal of
Electrical and Computer Engineering

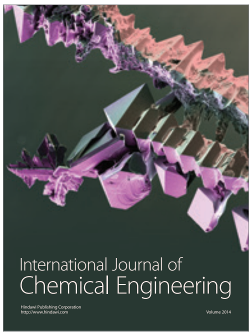

Advances in
OptoElectronics

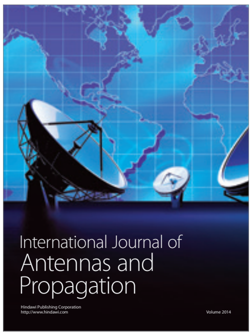
VLSI Design

International Journal of
Navigation and Observation

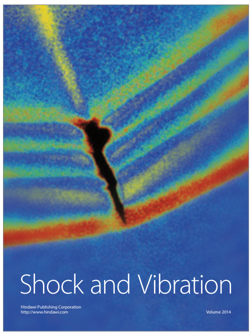

Modelling & Simulation in Engineering

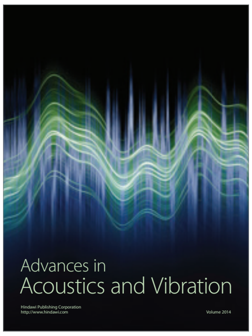
International Journal of
Aerospace Engineering


International Journal of
Chemical Engineering


International Journal of
Antennas and Propagation

Active and Passive
Electronic Components


Shock and Vibration


Advances in
Acoustics and Vibration