2016 Intl IEEE Conferences on Ubiquitous Intelligence & Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People, and Smart World Congress

Deep learning based Affective Model for Speech Emotion Recognition

Xi Zhou[*], Junqi Guo (corresponding author)[*†], Rongfang Bie[*]
Email: zhouxi@mail.bnu.edu.cn
[*]College of Information Science and Technology, Beijing Normal University, Beijing, P.R. China, 100875
[†]Beijing Advanced Innovation Center for Future Education, Beijing Normal University, Beijing, P.R. China

***Abstract*—Considering the application value of emotion, increasing attention has been attracted on emotion recognition over the last decades. We devote ourselves to feasible speech emotion recognition research. We build two affective models based on two deep learning methods (a stacked autoencoder network and a deep belief network) respectively for automatic emotion feature extraction and emotion states classification. The experiments are based on a well-known German Berlin Emotional Speech Database, and the recognition accuracy reaches 65% in the best case. In addition, we validate the influence of different speakers and different emotion categories on recognition accuracy.**

***Index Terms*—Speech Emotion Recognition, Deep Learning, Affective Model**

I. Introduction

Speech conveys most of the information in human communication, and is companied by emotions of the speaker. Speech emotion recognition (SER) research has gain an increasing attention in last decades. The useful emotion information can be applied in various applications. For instance, it can be used for preventing fatigue driving accidents in automobilism. It is to identify the driver's emotional activation state and offer the warning for state of fatigue or low activation by collecting the real time conversations from the driver and interactive voice response system. In the remote teaching, SER attempts to estimate the knowledge mastery degree of students via the emotions hidden in their response to teachers. It is significant for teachers to understand class situation and adjust the teaching plan promptly. In addition, it is valuable for emotion management of teenagers. Collecting their speech signal and recognizing latent emotion, which tends to provide the teenager with warning of emotional outburst or manual intervention when the emotional activation is high such as fury.

Emotion has several expressions such as visual representation and vocal signal. The facial expression recognition has achieved great improvement [1], [2], [3], while the researches on speech emotion recognition are expanding gradually [4], [5], [6]. Emotion recognition was first put forward as machine and emotion problem [7] by Marvin Minsky, one of the founders of artificial intelligence in 1985. In 1997, the concept of affective computing [8] was first suggested by Massachusetts Institute of Technology. With the development of big data and artificial intelligence, there appeared amount of schemes for emotion recognition problems. In the earlier periods, researchers tended to apply hand-tuned features into traditional classifiers. The features containing emotional information that are mostly used for speech emotion recognition include prosodic features, spectral features and acoustic features [9]. The prosodic features include speech intensity, fundamental frequency, energy, etc. The spectral features contain linear predictor coefficient (LPC), log-frequency power coefficient (LFPC), linear predictor cepstral coefficient (LPCC) and mel-frequency cepstral (MFCC), etc. The acoustic features cover formant frequency and glottis parameters, etc. The features are then input to emotion classification model. The early classification model mainly based on traditional pattern recognition methods [10], [11] which achieved good performance. For example, Lin et. al. [12] introduced HMM and SVM systems using sequential forward selection method for feature selection, and both systems obtained relatively high accuracy (>95%) for five emotional states classification. Rieger et. al. [13] used two k nearest neighbors to accomplish the classification task of six emotion states with spectral features and an accuracy above 85% was obtained. Although the classification tasks based on hand-tuned features and pattern recognition methods perform well, there is a critical issue that artificial multi-dimension features extraction of each frame is complicated to implement. What's more, the features that are heavily affected by person factors will loss some most intrinsic characteristics.

Based on the above points of view, we introduce deep learning methods for feature extraction and state recognition of speech emotion. The feature extraction is managed automatically by deep networks so as to avoid the influence of subjective factor on the emotion features. The deeper the network is, the more essential the speech emotion features are. In addition, the automatic feature extraction much reduces the implementation complexity. The deep learning [14] was first introduced by Hinton in 2006. It has been an extremely hot research topic in machine learning and pattern recognition area. It has gained theoretical successes in the field of speech recognition, natural language processing and computer vision [15]. In speech emotion recognition researches, deep learning attracts increasing attention, and there has been a lot of research achievements. For instance, Kim et. al. [16] explicitly captured complex non-linear feature interactions in multimodal data. They proposed and evaluated a suite of deep belief network models, and demonstrated that these models shown improvements in emotion classification performance

978-1-5090-2771-2/16 $31.00 © 2016 IEEE
DOI 10.1109/UIC-ATC-ScalCom-CBDCom-IoP-SmartWorld.2016.42

over baselines which is obtained by traditional methods. Cibau et. al. [17] trained a deep autoencoder based on a multilayer perceptron to predict six different emotions and neutral emotional state. Performance of the classifier was over 70%, which promotes better results for this novel approach.

In this paper, we train two affective model based on a stacked autoencoder network (SAE) and a deep belief network (DBN) to identify seven emotion states (Anger, Boredom, Disgust, Anxiety, Happiness, Sadness, Neutral state) on German Berlin Emotional Speech Database (EMO-DB). First of all, we preprocess the wav files in the database by sampling framing with a hamming window. The data of each frame is regarded as a sample. And we split the training dataset and testing dataset according to the proportion 7:3. Next, input the training dataset into the proposed affective model for automatic feature extraction and emotion states recognition. Finally, model evaluation is shown by experimental design.

The rest of this paper is organized as follows. Section II describes the profile of the affective model including system design and data preprocessing. In Section III, the deep learning algorithms including stacked autoencoder network and deep belief network are described more in detail. Section IV is to show the experiment setting, experiment design and experimental results of speech emotion recognition. Finally, some conclusions are drawn in Section V.

II. Affective Model

A. System Design

An affective model is used to recognize emotion state from speech signals. Basically, the architecture of the model contains four layers: input layer, data preprocessing layer, training layer and output layer. The input is the speech signal which is wav file. The data preprocessing layer is to sample and frame the audio signals. The detailed procedure is described later. The training layer attempts to apply two deep learning methods, stacked autoencoder network (SAE) and deep belief network (DBN) to automatically extract a set of salient features. The output layer gives out a predicted emotion state via a classification function. Fig 1. illustrates the system design.

B. Data preparation

In this paper, the obtained raw data is stored in wav format. The first thing is to transform analog speech signal to digital speech signal by sampling. The sampling frequency is 16kHz. And then, we obtain series of sample points. For the reason that the audio signals have short-time stationarity (0 50ms), it is necessary to frame the digital audio signals. The framing operation is implemented by a hamming window with length of 256, 50% overlap. Each frame is regarded as a sample with emotion label. The label is expressed by a binary vector with an element which represents corresponding emotion category equals 1. After that, we obtain a large amount of sample set. In order to train deep learning algorithms, the samples are separated to training dataset and testing dataset with the proportion of 7:3. The training and testing dataset are then used for training and evaluating the proposed affective model. The processing procedure is shown in Fig. 2.

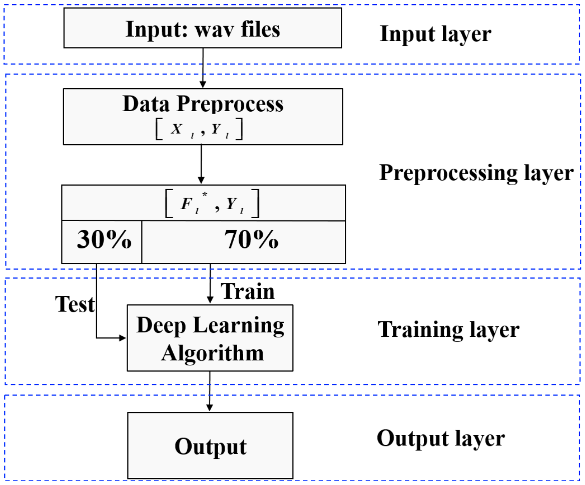

Fig. 1. The system design for emotion recognition.

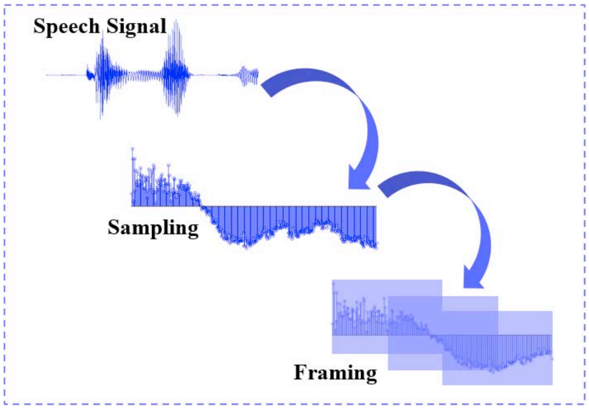

Fig. 2. The data preprocessing procedure.

III. Deep learning Methods

A. Stacked Autoencoder (SAE)

An autoencoder is a network including three layers: input layer (Layer 1), encoding layer (Layer 2) and decoding layer (Layer 3), as shown in Fig. 3. The stacked rule is setting the first autoencoder's output of the coding layer as the input vector of the next autoencoder. Assume that there is a stacked autoencoder network as shown in Fig. 4.

The encoding steps for a stacked autoencoder network are given by running the encoding step of each layer in forward order.

$$\mathbf{C}^k = f(z^k) \tag{1}$$

where z^k is defined as $z^k = \mathbf{W}^{(k,1)}\mathbf{C}^{k-1} + \mathbf{b}^{(k,1)}$. $\mathbf{W}^{(k,1)}$ between the input layer and encoding layer of the $k-th$ autoencoder. $\mathbf{C}^{k-1}$ is the encoding result from the $k-1-th$ autoencoder. $\mathbf{b}^{(k,1)}$ is the bias parameter of input layer of the $k-th$ autoencoder.

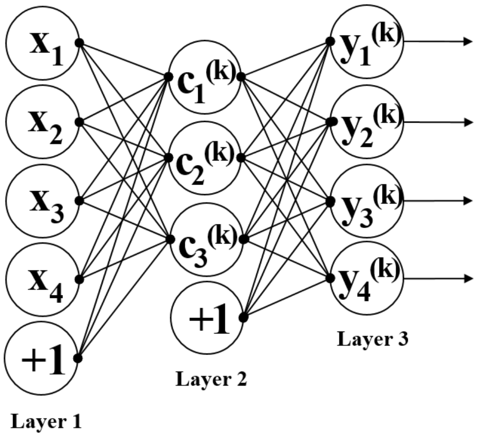

Fig. 3. The model of an autoencoder.

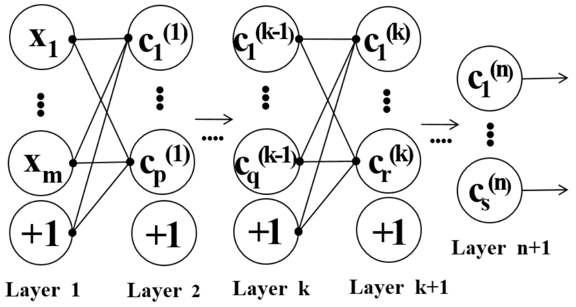

Fig. 4. The model of a stacked autoencoder network.

The decoding steps are to run the decoding step of each autoencoder in a reverse order.

$$\mathbf{y}^l = f(z^l) \tag{2}$$

where z^l is defined as $z^l = \mathbf{W}^{(l,2)}\mathbf{C}^l + \mathbf{b}^{(l,2)}$. $\mathbf{W}^{(l,2)}$ is the weight between the encoding layer and decoding layer of the $l-th$ autoencoder. $\mathbf{C}^l$ is the encoding result from the $l-th$ autoencoder. $\mathbf{b}^{(l,2)}$ is the bias of the encoding layer of the $l-th$ autoencoder. $\mathbf{C}^n$, which is the activation of the deepest layer of hidden units or the final result of encoding step contains the interest information. For a given input vector **x**, a stacked autoencoder network is to find a n-order encoding result that is features of the input. Define **y** as the final decoding result, then **y** can be regarded as the reconstruction of the input vector **x**. In order to better reconstruct **x** from **y**, the training of a stacked autoencoder network is performed by minimizing the reconstruction error, that is to say, optimizing the parameter set. In the standard autoencoder model, the reconstruction error is defined as the minimize square error (MSE) between raw input vector **x** and decoding vector **y**. So, the reconstruction loss function can be defined as follow:

$$L(\mathbf{y}, \mathbf{x}) = \|\mathbf{y} - \mathbf{x}\|^2 \tag{3}$$

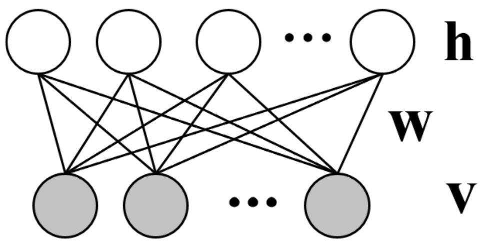

Fig. 5. The expression of graph model of a RBM.

A good way to obtain good parameters for a stacked autoencoder network is to use greedy layer-wise training [18]. To do this, first of all, we need to train the first autoencoder using raw input data to obtain a parameters set $\{\mathbf{W}^{(1,1)}, \mathbf{b}^{(1,1)}, \mathbf{W}^{(1,2)}, \mathbf{b}^{(1,2)}\}$. Then, encode the raw input data into a coding vector **C**, consisting of activation of the hidden units. Next, train the second autoencoder by using this vector to obtain parameters set $\{\mathbf{W}^{(2,1)}, \mathbf{b}^{(2,1)}, \mathbf{W}^{(2,2)}, \mathbf{b}^{(2,2)}\}$. Repeat this implementation for subsequent autoencoders, using the output of each encoding layer as the input of the next encoding layer. The method we presented trains parameters of each layer individually fixing other parameters of the model. To achieve better results, we can apply fine-tuning using backpropagation to improve the results which changes parameters of all layers simultaneously. The training steps are summarized in Table I.

TABLE I
THE TRAINING STEPS OF A STACKED AUTOENCODER NETWORK

Training steps of a stacked autoencoders network
1. Train the first autoencoder and obtain parameters by minimizing reconstruction loss;
2. Put the hidden units of the autoencoder as the input of the next autoencoder's encoding layer;
3. Iterate and initialize parameters of each layer;
4. Apply the output of the last hidden layer as input of a output layer, and initialize parameters;
5. Fine-tune parameters of all layers, and stack all autoencoders into a stacked autoencoder network.

B. Deep Belief Network (DBN)

A Restricted Boltzmann Machine (RBM) [19] is an undirected graph model which is as shown in Fig. 5. The **v** denotes the visible layer, which expresses the observation data. The **h** is the hidden layer which can be regarded as the feature extractors. The **W** is the weight between two layers.

The deep belief network (DBN) is a kind of deep network piled up by multiple RBMs. A DBN model containing three RBMs is illustrated in Fig. 6.

The general stacked rules of a DBN are demonstrated in Table II. When it comes to the more detailed and explicit learning and training methods of RBM and DBN, Hinton's paper [18] provides perfect explanation.

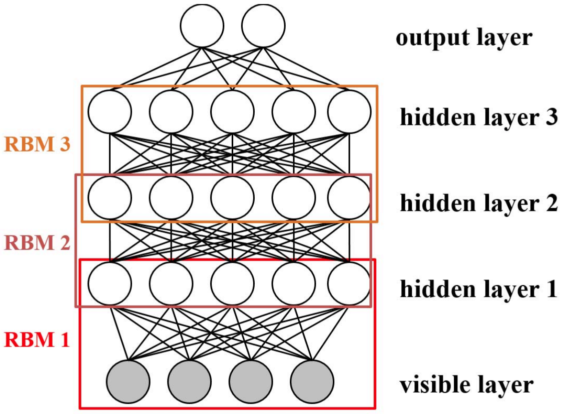

Fig. 6. The model of a DBN stacked by three RBMs.

TABLE II
THE GENERAL STACKED RULES OF A DBN.

The general stacked rules of a DBN
1. Train a RBM at first;
2. Fix the weight and bias of the first RBM, and then make the states of hidden units be the input of the second RBM;
3. Train the second RBM, stack the second RBM above the first RBM;
4. Repeat the above steps several times to obtain a DBN with specific number of RBMs;

In the paper, the input units are sampling data points of each frame. The stacked autoencoder network and deep belief network models based on deep learning algorithm are used to extract speech emotion features automatically. The classification of emotion states is achieved by sigmoid function which is expressed as follow.

$$sigmoid(x) = \frac{1}{1+e^{-x}} \tag{4}$$

IV. EXPERIMENT & RESULTS

A. Experiment Setting

1) Dataset: The recognition rate of emotion usually depends on the features and the database. The database we choose is the German Berlin Emotional Speech Database (Emo-DB), which is recorded by Burkhardt et. al [20]. There are seven kinds of emotion which are anger, boredom, disgust, anxiety, happiness, sadness and neutral state, respectively. The speech materials including ten German utterances are simulated by ten (five females, five males) actors according to each emotion. The database is comprised of 535 utterances with different length. The information of the ten actors is shown in Table III where the speaker number is in accordance with the database's..

2) Deep learning toolbox: To accomplish the affective models described in Section 3, we introduce a deep learning toolbox developed by Rasmus on Matlab to use the examples of stacked autoencoder and deep belief network.

TABLE III
THE INFORMATION OF THE SPEAKERS IN THE DATABASE.

Speaker num.	03	08	09	10	11	12	13	14	15	16
Gender	M	F	F	F	M	M	F	F	M	M
Age	31	34	21	32	26	30	32	35	25	31

3) Evaluation Criteria: Accuracy is a common and widely used criteria to evaluate the classification tasks. It is to indicate the proportion of correctly predicted samples in the whole dataset. The definition of accuracy in mathematics is as follow.

$$Accuracy = \frac{N_c}{N} \tag{5}$$

where N_c denotes the number of samples which are correctly predicted. N denotes the number of all samples to predict.

B. Experimental Results

For the purpose of evaluating the proposed deep learning based affective model, we design the following experiments. The baseline is the accuracy 22.48% which is obtained by Huang [21] using raw spectral features and support vector machine.

1) Classification performance under different sizes of training dataset: We generate ten different sizes of training and testing datasets from all of the labeled frames/samples to observe the seven emotions classification accuracy obtained by SAE and DBN based affective model. We conduct the experiment to show the influence of training dataset size on emotion classification performance. The parameters of affective model are shown in Table IV. The network structure 256+100+100+7 means that the network contains four layers: input layer including 256 units, hidden layer including 100 units, hidden layer including 100 units, and output layer including 7 units. The network structure 256+100+7 means the network contains three layers: input layer including 256 units, hidden layer including 200 units, and output layer including 7 units. Fig. 7 is the experimental result.

TABLE IV
THE PARAMETERS OF DBN AND SAE.

	network structure	batch size	epochs	learning rate
DBN	256+100+100+7	100	20	0.03 ± 0.02
SAE	256+200+7	100	20	0.1 ± 0.05

As the figure shows, the SAE based classification model obtains an accuracy around 25%, the DBN's is about 38%, and both of them are higher than baseline. The reason why the DBN obtains higher accuracy than SAE is that the network structure of DBN is deeper than SAE one hidden layer. It is obvious that the accuracy is relatively low and it changes a little with the increasing of the training dataset size. Therefore, we suggest that the training dataset size is not the critical factor to the classification performance in the experiment.

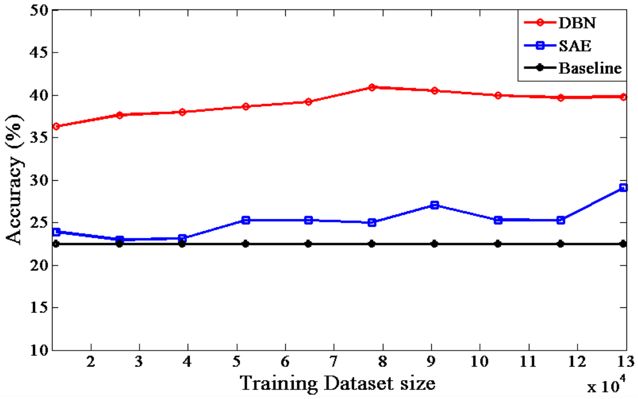

Fig. 7. The accuracy of different training datasets with increasing size.

TABLE V
THE PARAMETERS OF DBN AND SAE.

	network structure	batch size	epochs	learning rate
DBN	256+100+100+7	100	20	0.1
SAE	256+200+7	100	20	0.1

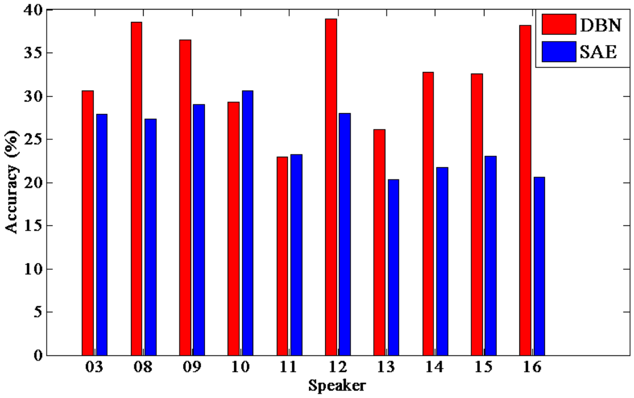

Fig. 8. The accuracy with different speakers' data.

2) Classification performance under different speakers: We split the speech data of each speaker to generate ten datasets for training and testing DBN and SAE. The related parameters of the affective model are explained in Table V. The size of each training dataset of corresponding speaker is listed in Table VI. Fig. 8 expresses the experimental results.

As we can see from the figure, SAE achieves an accuracy about 29%, while DBN reaches an accuracy 39% in the best case. The DBN obtains higher accuracy than SAE averagely, as well. The reason is also that DBN is deeper than SAE. For the specific model, DBN or SAE, the accuracy has no much change. Therefore, we hold the opinion that speaker is also not the crucial factor to classification performance.

3) Classification performance under different kinds of emotion states: The last experiment is designed as follow. We select different numbers of emotion states to evaluate the proposed affective model. The emotions categories and training dataset size are shown in Table VII. Fig. 9 illustrates the experimental result. The relative parameter settings are as the same as the first experiment.

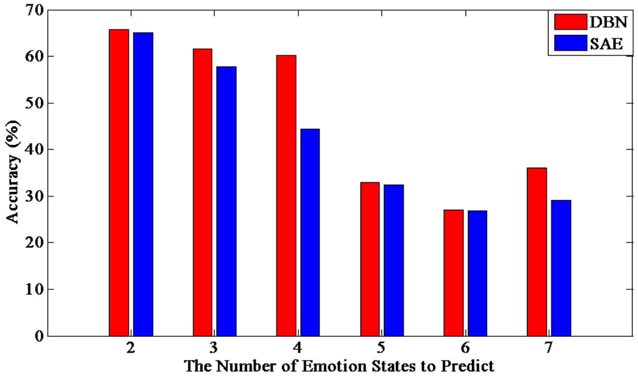

Fig. 9. The accuracy of different number of emotion states.

The figure expresses that the DBN's performance is also better than SAE. What's more, the classification performance is reduced rapidly with the increase of emotion state number. We infer that the emotion category is the key issue for classification performance. We conclude that the features automatically extract by DBN and SAE is not robust and salient enough for recognizing more categories of emotion precisely. It is important to extract more robust, salient, and discriminative features to identify some similar emotions.

V. CONCLUSION

In this paper, we have proposed two affective model based on deep leaning algorithms for speech emotion feature automatic extraction and emotion state recognition. The database used in the paper is German Berlin Emotional Speech Database with seven emotion states which are anger, boredom, disgust, anxiety, happiness, sadness and neutral state, respectively. We apply sampling and framing operation to implement the data preprocessing of speech signals and obtain a set of labeled samples. The samples are split to different training datasets and testing datasets for experimental verification. In the part of experiment, we conduct three experiments using different sizes of training dataset, different speakers' datasets, and different kinds of emotions' datasets, respectively. We compare the accuracy of emotion recognition and find out that the accuracy changes from highest 65% to the lowest 23%. In order to reinforce the performance of recognizing similar emotion states, we need to use deeper networks to extract more robust, salient and discriminative features for emotion recognition. This is also what our future work focuses on.

ACKNOWLEDGMENT

This research is sponsored by National Natural Science Foundation of China (No.61401029, 61171014, 61472044, 61472403, 61371185), Beijing Advanced Innovation Center for Future Education (BJAICFE2016IR-004), the Fundamental

TABLE VI
THE SIZE OF EACH TRAINING DATASET OF CORRESPONDING SPEAKER

Speaker num.	03	08	09	10	11	12	13	14	15	16
Training dataset size	11200	15400	10200	7900	13800	8700	13900	16800	12100	18900

TABLE VII
THE EMOTION CATEGORIES AND TRAINING DATASET SIZE IN THE EXPERIMENT.

Emotion numbers	2	3	4	5	6	7
Emotion categories	anger happiness	anger happiness sadness	anger happiness sadness anxiety	anger happiness sadness anxiety disgust	anger happiness sadness anxiety disgust boredom	anger happiness sadness anxiety disgust boredom neural state
Training dataset size	44900	66800	80200	80200	80200	80200

Research Funds for the Central Universities (No.2012LYB46), and Beijing Youth Excellence Program (YETP0296).

REFERENCES

[1] M. F. Valstar, T. Almaev, J. M. Girard, G. McKeown, M. Mehu, L. Yin, M. Pantic, and J. F. Cohn, "Fera 2015-second facial expression recognition and analysis challenge," in *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, vol. 6. IEEE, 2015, pp. 1–8.

[2] S. Happy and A. Routray, "Automatic facial expression recognition using features of salient facial patches," *Affective Computing, IEEE Transactions on*, vol. 6, no. 1, pp. 1–12, 2015.

[3] Y. Tian, T. Kanade, and J. F. Cohn, "Facial expression recognition," in *Handbook of face recognition*. Springer, 2011, pp. 487–519.

[4] H. P. Martinez, Y. Bengio, and G. N. Yannakakis, "Learning deep physiological models of affect," *Computational Intelligence Magazine, IEEE*, vol. 8, no. 2, pp. 20–33, 2013.

[5] V. Sethu, J. Epps, and E. Ambikairajah, "Speech based emotion recognition," in *Speech and Audio Processing for Coding, Enhancement and Recognition*. Springer, 2015, pp. 197–228.

[6] 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.

[7] M. Minsky, *The Society of Mind*, ser. A Touchstone Book. Simon & Schuster, 1986. [Online]. Available: https://books.google.com.hk/books?id=veVOAAAAMAAJ

[8] R. W. Picard and R. Picard, *Affective computing*. MIT press Cambridge, 1997, vol. 252.

[9] N. A. Zaidan and M. S. H. Salam, "A review on speech emotion features," *Jurnal Teknologi*, vol. 75, no. 2, 2015.

[10] S. Casale, A. Russo, G. Scebba, and S. Serrano, "Speech emotion classification using machine learning algorithms," in *Semantic Computing, 2008 IEEE International Conference on*. IEEE, 2008, pp. 158–165.

[11] C.-N. Anagnostopoulos, T. Iliou, and I. Giannoukos, "Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011," *Artificial Intelligence Review*, vol. 43, no. 2, pp. 155–177, 2015.

[12] Y.-L. Lin and G. Wei, "Speech emotion recognition based on hmm and svm," in *Machine Learning and Cybernetics, 2005. Proceedings of 2005 International Conference on*, vol. 8. IEEE, 2005, pp. 4898–4901.

[13] S. A. Rieger, R. Muraleedharan, and R. P. Ramachandran, "Speech based emotion recognition using spectral feature extraction and an ensemble of knn classifiers," in *Chinese Spoken Language Processing (ISCSLP), 2014 9th International Symposium on*. IEEE, 2014, pp. 589–593.

[14] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[15] X.-W. Chen and X. Lin, "Big data deep learning: challenges and perspectives," *Access, IEEE*, vol. 2, pp. 514–525, 2014.

[16] Y. Kim, H. Lee, and E. M. Provost, "Deep learning for robust feature generation in audiovisual emotion recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 3687–3691.

[17] N. E. Cibau, E. M. Albornoz, and H. L. Rufiner, "Speech emotion recognition using a deep autoencoder," *Proceedings of the XV Reunión de Trabajo en Procesamiento de la Información y Control (RPIC 2013), San Carlos de Bariloche*, 2013.

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

[19] G. E. Hinton and T. J. Sejnowski, "Learning and releaming in boltzmann machines," *Parallel distributed processing: Explorations in the microstructure of cognition*, vol. 1, pp. 282–317, 1986.

[20] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of german emotional speech." in *Interspeech*, vol. 5, 2005, pp. 1517–1520.

[21] Z.-w. Huang, W.-t. Xue, and Q.-r. Mao, "Speech emotion recognition with unsupervised feature learning," *Frontiers of Information Technology & Electronic Engineering*, vol. 16, pp. 358–366, 2015.