Scholar Publishing Group
International Journal of Neural Network
https://doi.org/10.38007/NN.2020.010206
ISSN 2790-2692 Vol. 1, Issue 2: 40-47

Speech Emotion Recognition with Deep Belief Network

Sahil Verma*

National Polytechnic Institute of Cambodia, Cambodia

**corresponding author*

Keywords: Deep Belief Network, Speech Emotion, Emotion Recognition, Boltzmann Machine

Abstract: With the continuous development of information technology and deep learning, more and more methodological ideas have penetrated into the application field of speech emotion recognition, and how to recognize the emotion expressed in human speech more scientifically and effectively has become an important issue in speech emotion recognition and neural network research. In order to solve the shortcomings of the existing research on speech emotion recognition by fusing deep confidence networks, this paper briefly introduces the sample data and parameter settings for the application of speech emotion recognition model by fusing deep confidence networks, based on the functional equations of Boltzmann machine and the training steps of deep confidence networks and the types of speech emotion recognition. The experimental data show that the recognition accuracy of deep confidence networks is higher than that of (SVN) and (RNN) models, and the average recognition accuracy of deep confidence networks reaches 96%, while the average recognition accuracy of (SVN) and (RNN) models reaches 91%, respectively. The average accuracy of (SVN) and (RNN) recognition reached 91% and 92%, respectively, thus verifying the feasibility of fusing deep confidence networks for speech emotion recognition.

1. Introduction

Speech emotion recognition is a method of using computer technology to classify and organize natural speech information to obtain different emotional characteristics. Using language emotion recognition method can allow computer systems to autonomously identify the speaker's emotional state in speech information.

Nowadays, more and more scholars pay attention to the research of various computer technologies and system tools in speech emotion recognition, and through practical research, they have also achieved certain research results. Mustaqeem's research mainly focuses on speech features and traditional convolutional neural network models for extracting speech emotion features

from speech spectrograms to improve speech emotion recognition accuracy and reduce the complexity of speech emotion recognition. Mustaqeem proposes a new SER model that uses the construction of key sequence segments based on redialing network groups. The speech emotion features are identified by the STFT algorithm and passed to the CNN model to identify and classify the obvious features of speech emotion. Mustaqeem speech emotion processing is to extract key parts of speech, so that speech emotion information can be easily recognized [1]. Morgan M M proposed a new emotion recognition algorithm, which is completely separated from the original method of speech acoustic features and combined with speaker gender characteristics. The goal of Morgan M M is to obtain information of emotional characteristics from the gender characteristics of the initial speaker, which is a form of recognition based on intelligent technology. Morgan M M uses deep learning algorithms to automatically extract key information from the speaker's initial speech information for the classification layer to perform speech emotion recognition. This method can avoid errors in speech data recognition [2]. Shoiynbek A evaluated the accuracy of neuroacoustic emotion recognition models in human-computer interaction. Shoiynbek A influences the recognition performance of the model by assuming various things that can happen in the speaker's ambient noise, room layout, and basic information about the speaker. Shoiynbek A conducted three tests on the Cub robot system and proposed several effective methods to reduce the error between the test value and the true value of the model in acoustic emotion recognition. Furthermore, Shoiynbek A demonstrated the necessity of introducing data augmentation techniques to improve the effect of model recognition [3]. Although there are many existing researches on speech emotion recognition, the research on speech emotion recognition integrated with deep belief networks still has certain limitations.

Therefore, in order to solve the existing problems of speech emotion recognition based on fusion of deep belief networks, this paper firstly introduces the mathematical model of Boltzmann machine, the training steps of deep belief networks and the concept of speech emotion recognition types, and then discusses the fusion of The parameter settings and sample data in the design and application of the speech emotion recognition model of the deep belief network. Finally, the speech emotion recognition model architecture integrated with the deep belief network is designed, and the experimental test is carried out through the specific application effect of the designed model. The final experiment shows that this paper the effectiveness of the designed speech emotion recognition model incorporating deep belief networks.

2. Speech Emotion Recognition with Deep Belief Network

2.1. Boltzmann Machines

Boltzmann machine is a two-layer structure of neural network, the visible layer is the input layer, which is used to accept the input data, and the other layer is the hidden layer, which is the mapping of the feature data of the input layer [4].

The Boltzmann machine draws on the idea of simulated annealing, and the purpose of model training is to find the probability $f(r|\beta)$ of the speech emotion recognition sample under the model parameters, where r is the speech data node, and β is the model parameter [5]. From the Bayesian formula, we know that:

$$f(r|\beta) = \sum_{k} f(r,k|\beta) \tag{1}$$

Among them, k is the speech emotion hidden layer node, and $f(r,k|\beta)$ is the joint

probability between the speech emotion node and the speech emotion hidden layer node [6]. First define the energy function of the speech emotion recognition model as:

$$G = -(\sum_{u<v} p_{uv} z_u z_v + \sum_u a_u z_u) \quad 2)$$

In the above formula, z_u represents the speech emotion recognition accuracy of node u, p_{uv} represents the connection weight between node u and node v, and c_u represents the bias weight of the node. It can be seen from the above formula that for node u, when it is converted from an inactive state (value 0) to an active state (value 1), the change of system energy can be expressed as:

$$\Delta G_u = G(z_u = 0) - G(z_u = 1) = \sum_v p_{uv} z_v + a_u \quad (3)$$

2.2. Deep Belief Network Training

The training process of the deep confidence network consists of the following two steps.

(1) Pre-training

Pre-training is essentially an unsupervised training process of neural networks [7]. Since the deep belief network cannot fully simulate the emotional features of speech, a higher-level network is needed to simulate the emotional characteristics of speech, and the training method of deep belief network is effective [8].

(2) Fine-tuning

After pre-training, each level of the deep belief network model has been initialized [9]. All RBM modules are connected according to the training order, and together with a single-layer neural network constitute a deep belief network [10]. to adjust the energy loss function based on the input data and the reconstructed data during the training process [11].

2.3. Speech Emotion Recognition

Generally speaking speech emotion features can be divided into three kinds, and this paper focuses on rhyme and tone quality features [11].

(1) Prosody features

The prosodic feature mainly reflects the prosody changes of a person's speech rate, pitch, and volume. In calm situations, these data are relatively flat, so they can be used as a measure of emotional state [14].

(2) Sound quality characteristics

The sound quality feature is a feature type related to the way of human pronunciation, and its influencing factors are the length tension of the human vocal tract and the pressure in the middle of the vocal tract [15]. In the types of emotions with strong emotions, it is not enough to use prosodic features to separate them, and the discrimination of such emotions can be improved by adding sound quality features [16].

3. Investigation and Research on Speech Emotion Recognition Integrating Deep Belief Network

3.1. Sample Data

This paper selects the Danish Speech Emotion Database as the experimental data, and uses four emotional scenes that are often expressed by happy, angry, fearful and sad humans in the database

[17]. There are 60 samples for each type of emotional speech, a total of 300 sample data. And 100 training samples, 100 validation samples and 100 test samples for experiments [18]. The specific sample datadistribution is shown in Table 1:

Table 1. Sample distribution

Emotion	Training samples	Test sample	Test numbers	Total
Happy	37	52	86	175
Angry	86	28	23	137
Fear	82	28	17	127
Sad	79	23	13	115

3.2. Experimental Parameter Setting

In the experimental tests of the three recognition models selected in this paper, the only difference between the three models is the difference in the classifier connected to the final output of the recognition [18]. The sigmoid function was used as the recognition activation function of the deep confidence network in the experiments, with the parameters shown in Table 2.

Table 2. Model parameters

Model parameters	Output layer	Hidden layer
Learning rate	0.02	0.05
Penalty coefficient	0.002	0.004
Cycles	200	200
Degeneration factor	0.002	0.003
Number of iterations	500	500

4. Application Research of Speech Emotion Recognition Based on Deep Belief Network

4.1. Establishment of Speech Emotion Recognition Model Integrating Deep Belief Network

Deep confidence networks are a common model for deep learning, which consists of a visible layer, a hidden layer, and an output layer, where the hidden layer is composed of multiple layers of restricted Boltzmann machines, and the role of the hidden layer is to be trained to capture the relevance of the data in the visible layer. For the specific recognition framework structure is shown in Figure 1.

The specific training process is as follows.

(1) In the speech processing block, the visible layer obtains the initial speech input feature vector, and the language sample is quantified and selected through the input feature vector of the known visible layer and transferred to the hidden layer. It is passed to the input layer to reconstruct the feature vector of the speech.

(2) In the speech sample training and recognition section, the deep belief network is fine-tuned through the reconstructed speech sample data, and the speech recognition results are obtained at the output layer through layer-by-layer evaluation, and the recognition results are compared with the real results. Through the back propagation of the error value, the weights of each layer of neurons are fine-tuned again.

(3) In the speech feature extraction section, after fine-tuning using the deep belief network, the

weights between neurons have been successfully constructed, and the speech emotion feature parameters are extracted through the deep belief network.

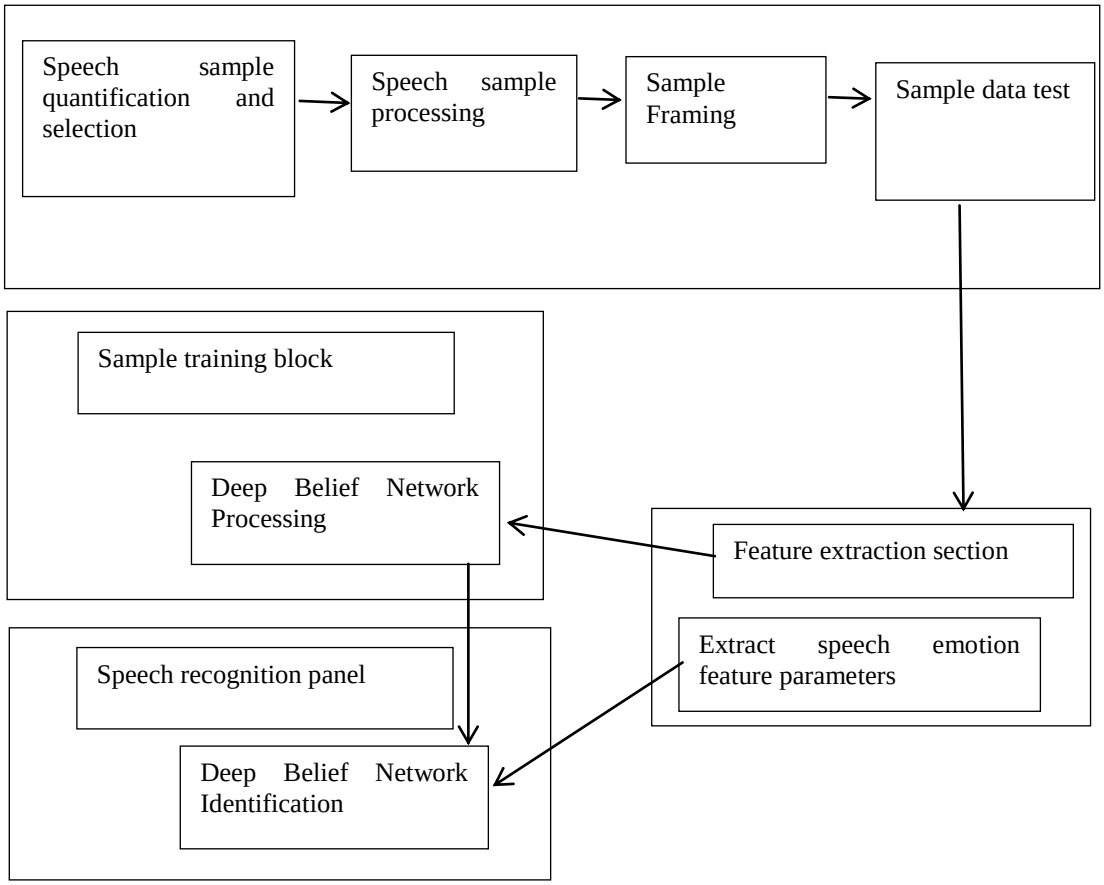

Figure 1. Framework diagram of speech emotion recognition model fused with deep belief network

4.2. Application of Speech Emotion Recognition Integrating Deep Belief Network

In order to verify the recognition effect of the fused deep confidence network, 150 test samples from the investigated dataset were used to compare the accuracy of the proposed algorithm model with the other two models (SVN) and (RNN) for the recognition of four emotions in the sample speech: happy, annoyed, fearful and sad.

Table 3. Model identification results

Emotion	Deep Belief Network	SVN	RNN
Happy	98.56%	92.26%	94.19%
Angry	96.89%	93.25%	92.45%
Fear	97.28%	91.75&	90.85%
Sad	95.17%	90.18%	91.71%

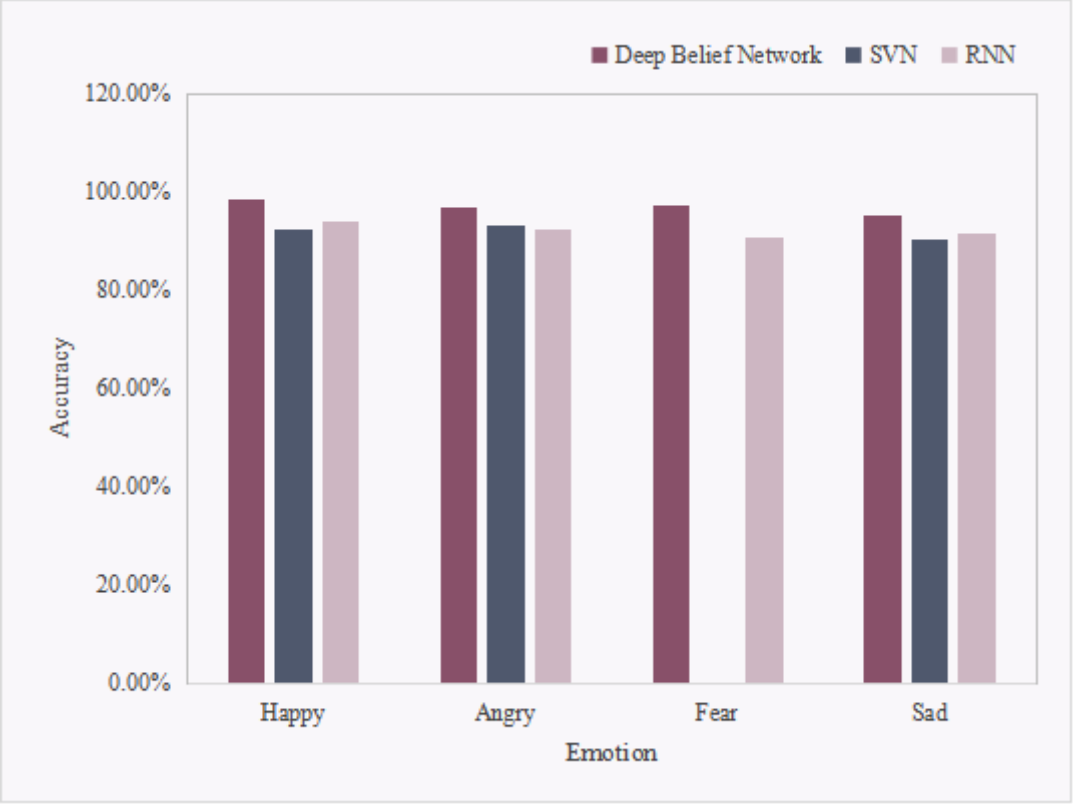

Figure 2. Comparison of model recognition results

Figure 2 shows the prediction results of the deep confidence network and the (SVN) and (RNN) models for the speech emotion categories, where the vertical direction indicates the real emotion category of the speech. The accuracy rate of the model recognition results is shown horizontally. In the comparison between the deep confidence network model and the (SVN) and (RNN) models, the accuracy rate of the deep confidence network for the recognition of the four emotions is 98.56%, while the accuracy rate of the (SVN) and (RNN) models is lower than that of the deep confidence network for the recognition of the emotions, but the accuracy rate of the recognition is also above 90%. The accuracy of the (SVN) and (RNN) models is lower than that of the (SVN) and (RNN) models, but the accuracy of the (SVN) and (RNN) models is above 90%, and the accuracy of the (SVN) and (RNN) models is still higher than that of the (SVN) and (RNN) models in the recognition of annoyed emotion, with 96.89%, 93.25% and 92.45%, respectively. This indicates that the deep confidence network has a strong recognition effect on speech emotion.

5. Conclusion

In this paper, we describe the technical basis of speech emotion recognition model building, including the functional equation of Boltzmann machine and two emotion features of rhythm and tone quality, as well as the training steps of pre-training and fine-tuning, and analyze the parameter setting and sample data of recognition model building. The designed recognition model is compared with two other models to demonstrate the superiority of deep confidence networks in speech emotion recognition.

Funding

This article is not supported by any foundation.

Data Availability

Data sharing is not applicable to this article as no new data were created or analysed in this study.

Conflict of Interest

The author states that this article has no conflict of interest.

References

[1] Mustaqeem, Kwon S. 1D-CNN: Speech Emotion Recognition System Using a Stacked Network with Dilated CNN Features. Computers, Materials and Continua, 2020, 67(3):4039-4059. https://doi.org/10.32604/cmc.2020.015070

[2] Morgan M M, Bhattacharya I, Radke R, et al. Automatic speech emotion recognition using deep learning for analysis of collaborative group meetings. The Journal of the Acoustical Society of America, 2019, 146(4):3073-3074. https://doi.org/10.1121/1.5137665

[3] Shoiynbek A, Sultanova N. Speech Emotion Recognition for Kazakh and Russian Languages. Applied Mathematics & Information Sciences, 2020, 14(1):65-68. https://doi.org/10.18576/amis/140108

[4] Mohammed S N, Karim A. Speech Emotion Recognition Using MELBP Variants of Spectrogram Image. International Journal of Intelligent Engineering and Systems, 2020, 13(5):257-266. https://doi.org/10.22266/ijies2020.1031.23

[5] Gunawan T S, Noor A, Kartiwi M. Development of english handwritten recognition using deep neural network. Indonesian Journal of Electrical Engineering and Computer Science, 2018, 10(2):562-568. https://doi.org/10.11591/ijeecs.v10.i2.pp562-568

[6] Jaratrotkamjorn A. Bimodal Emotion Recognition Using Deep Belief Network. ECTI Transactions on Computer and Information Technology (ECTI-CIT), 2020, 15(1):73-81. https://doi.org/10.37936/ecti-cit.2020151.226446

[7] Ocquaye E, Mao Q, Song H, et al. Dual Exclusive Attentive Transfer for Unsupervised Deep Convolutional Domain Adaptation in Speech Emotion Recognition. IEEE Access, 2019, PP(99):1-1. https://doi.org/10.1109/ACCESS.2019.2924597

[8] Kumar Y, Mahajan M. Machine Learning Based Speech Emotions Recognition System. International Journal of Scientific & Technology Research, 2019, 8(7):722-729.

[9] Haritha C V, Thulasidharan P P. Multimodal Emotion Recognition using Deep Neural Network- A Survey. International Journal Of Computer Sciences And Engineering, 2018, 06(6):95-98. https://doi.org/10.26438/ijcse/v6si6.9598

[10] Praseetha V M, Vadivel S. Deep Learning Models for Speech Emotion Recognition. Journal of Computer Science, 2018, 14(11):1577-1587. https://doi.org/10.3844/jcssp.2018.1577.1587

[11] Zvarevashe K, Olugbara O O. Recognition of speech emotion using custom 2D-convolution neural network deep learning algorithm. Intelligent Data Analysis, 2020, 24(5):1065-1086. https://doi.org/10.3233/IDA-194747

[12] Shao H, Jiang H, Zhang H, et al. Rolling bearing fault feature learning using improved convolutional deep belief network with compressed sensing. Mechanical Systems and Signal Processing, 2018, 100(FEB.1):743-765. https://doi.org/10.1016/j.ymssp.2017.08.002

[13] Kumar G, Dr S M, Dr A N. An Ensemble of Feature Subset Selection with Deep Belief Network Based Secure Intrusion Detection in Big Data Environment. Indian Journal of Computer Science and Engineering, 2020, 12(2):409-420. https://doi.org/10.21817/indjcse/2020/v12i2/211202001

[14] Et. A. Breast Cancer Detection Using Deep Belief Network by Applying Feature Extraction on Various Classifiers. Turkish Journal of Computer and Mathematics Education (TURCOMAT), 2020, 12(1S):471-487. https://doi.org/10.17762/turcomat.v12i1S.1909

[15] Jaratrotkamjorn A. Bimodal Emotion Recognition Using Deep Belief Network. ECTI Transactions on Computer and Information Technology (ECTI-CIT), 2020, 15(1):73-81. https://doi.org/10.37936/ecti-cit.2020151.226446

[16] Lalithadevi B. Novel Technique for Price Prediction by Using Logistic, Linear and Decision Tree Algorithm on Deep Belief Network. International Journal of Psychosocial Rehabilitation, 2020, 24(5):1751-1761. https://doi.org/10.37200/IJPR/V24I5/PR201846

[17] Vankdothu R. Efficient Detection of Brain Tumor Using Unsupervised Modified Deep Belief Network in Big Data. Journal of Advanced Research in Dynamical and Control Systems, 2020, 12(SP4):338-347. https://doi.org/10.5373/JARDCS/V12SP4/20201497

[18] Annamalai P. Automatic Face Recognition Using Enhanced Firefly Optimization Algorithm and Deep Belief Network. International Journal of Intelligent Engineering and Systems, 2020, 13(5):19-28.https://doi.org/10.22266/ijies2020.1031.03