# Deep Learning Analysis Models for Speech and Emotional Recognition

Jun WU [1,2] , Tianliang Zhu [1] , Chengtian YU[1] and Chunzhi WANG[1], Xianjing ZHOU[2], Hu LIU[2]
[*]Tsinghua University, Beijing, China
1 School of Computer Science, Hubei University of Technology, Wuhan 430068, China
2 Wuhan Zall Information Technology Co. , Ltd. Wuhan 430000, China

***Abstract*—Today speech emotion recognition is one of the major technology usages for every areas in the world. There are many important aspects with speech emotion recognition which would be more convenient and intelligent. Furthermore it is necessary to optimize existing methods to be executable for speech emotion recognition. In this paper, build models based on CNN, SVM and LSTM for speech emotion recognition. Compared with the three methods and analyzed the shallow learning and deep learning models to find a method which has a better performance.**

## I. INTRODUCTION

Speech is one of the main means of information exchange between people, through the different intones and voices, we can express a variety of emotions. But in most cases, the speaker does not express the real emotion directly, so that the real emotion of the voice need to be excavated. With the development of artificial intelligence technology, accurate emotional analysis can bring better services, such as cheaper psychological guidance, more interactive intelligent escort.

Speech emotion recognition is a kind of technology that extracts the features of emotion signals through computer processing and infers the types of speech emotion through analysis. The task is to extract the features related to emotion from the speaker's speech, and find out the mapping relationship between these features and human emotion. In 1987, Professor Minsky put forward the concept of "computer's emotional ability" in his book the society of mind[1] ,since then, speech emotion recognition has attracted more and more attention; in the early 1990s, the Multimedia Laboratory of Massachusetts Institute of technology in the United States constructed an "emotion editor", which can collect speech signals, facial expressions, physiological signals for emotion recognition, and calculate internal emotions and make appropriate and simple reactions[2]; in 1999, Moriyama This paper proposes a linear correlation model of voice emotion, which will be applied to e-commerce for the first time[3], but the voice emotion analysis is not mature at this time. With the development of technology, many algorithms have been applied to speech emotion recognition.

Feature learning strategies always spend much time. In order to reduce these time and computational costs, ZT Liu, A Rehman and his comrade proposed pre-processing step[4] in which speech segments with similar formant characteristics are clustered together and labeled as the same phoneme.

In order to mine the relevance of signals in audios an increase the diversity of information, Bi-directional Long-Short Term Memory with Directional Self-Attention (BLSTM-DSA) is proposed by Kunxia Wang, Guoxin Su, Li Liu, Shu Wang.[5] There use autocorrelation of speech frames to deal with the lack of information, so that Self-Attention mechanism is introduced into SER. Thus, the algorithm can automatically annotate the weights of speech frames to correctly select frames with emotional information in temporal network.

Dias Issa, M. Fatih Demirci and Adnan Yazici introduced a new architecture, which extracts mel-frequency cepstral coefficients, chromagr[6], mel-scale spectrogram, Tonnetz representation, and spectral contrast features from sound files. The new architecture outperforms all previous works.

Therefore, starting from shallow learning and deep learning, this paper builds 3 models: support vector machine (SVM), long-term and short-term memory network (LSTM) and convolutional neural network (CNN) for comparative experiments, the goal of the study is to make full use of the advantages of shallow learning and deep learning so that we can improve the model is recognition rate and robustness.

## II. RELATED WORK

Speech signal-based emotional recognition is roughly divided into two categories, the first is classifier-based recognition, the emotions are divided into basic emotions, such as happiness, sadness, anger, fear, etc., the second is based on regression analysis, through two dimensions: Arouse (the level of arousal), Valence (the level of positive emotions), to build a two-dimensional space to describe human emotions. Compared with the discrete emotion database, the dimension emotion database only accounts for a small number. At present, there are mainly VAM[7] and semaine[8] The same emotional language information not only has similar acoustic characteristics, but also mixed with the individual style among different speaker. Speech emotion recognition is also a kind of pattern recognition in essence, which involves the theory of machine learning, so the traditional shallow learning model is still widely used as far as deep learning model. In terms of model selection, the support vector machine (SVM)[9], hidden Markov model (HMM)[10], Swain M and Sahoo S used with Hidden Markov model and support vector machines classifier, for classifying a speech into one of the seven discrete emotion classes ,Gauss hybrid model

(GMM)[11], convolutional neural network(CNN)[12], Z Huang and his comrade propose to learn affect-salient features for Speech Emotion Recognition (SER) using semi-CNN, and there experiment results on benchmark datasets show that our approach leads to stable and robust recognition performance in complex scenes, recursive neural network (RNN)[13] , With only prosodic acoustic features and SVM multi-classifier, T Zhang and J Wu[14] obtain a f-measure of 38.3%,adding the I-vector features and the RNN model, they achieve a better result of 48.9%. Compared with them, deep learning can extract high-level features but the network structure is complex, adjusting parameters are difficult and require a lot of training data, while the shallow learning model training speed is faster, the parameters are few, the extracted features are targeted but the performance of large amounts of data is not good compared with deep learning.

## III. SYSTEM MODEL

### A. Models

1) Support Vector Machine, SVM

This method is based on the theory of statistical learning, through learning algorithms, SVM will automatically find the classification support vectors which have a better ability to distinguish. The constructed classifier can maximize the interval between classes, and thus have better adaptability and high accuracy. This method only needs to determine the final classification result by the category of boundary samples for each domain. The purpose of the support vector machine algorithm is to find a hyperplane H(d), which separates the data in the training set, and get the largest distance from the class domain boundary perpendicular to the hyperplane direction, so that the SVM method is also known as the maximum edge (maximum margin) algorithm. For most samples in the sample set which are not support vectors, removing or reducing these samples has no effect on the classification results, and the SVM method has good classification results for automatic classification in the case of small samples.

2) *Convolutional Neural Network, CNN*

The applications of Convolutional Neural Network are extensive, including image and video recognition, speech recognition, medical image analysis and natural language processing. The convolution kernel structure of the CNN model can integrate local emotional features at the top level to obtain global emotional characteristics. The pooling operation in the model can effectively adapt to different speech speed, position changes and so on. It also can improve the accuracy of recognition. Compared with shallow neural networks, the parameter sharing of convolution neural networks can greatly reduce the number of parameters. The convolution neural network consists of several parts:

*(a) Input layer: data input.*

*(b) Reel layer*: extract the input feature and multiply each parameter in the co product core by the local pixel value corresponding to the upper input layer. Performing one layer of convolution operations can get results in the lower-level characteristics of the input, and performing multi-layered convolution operations can iterate from the lower-level features to extract more advanced features.

*(c) Linear rectification layer:* Non-linearity mapping of the structure of the convolution layer by excitation function.

*(d) Pooled layer*: reduce the amount of data operations.

*(e) Full connection layer*: expand the output of the upper layer. Connect to each neuron to reduce the loss of feature information.

*(f) Output layer:* be used to final output results.

*3) Long-Term Memory Neural Network , LSTM*

LSTM networks are an improvement on traditional recurrent neural networks that are less able to learn when encountering longer sequences intervals. The LSTM model can store information for a long time, and it is a kind of neural network with memory dynamic energy, which is suitable for modeling time series data. LSTM network unit consists of three nodes: the forget gate, the input gate, and the output gate. The forget gate normalizes the degree of oblivion of the last unit between 0 and 1 through the sigmoid function, the input door controls the addition of new information, and the output gate controls the information output of the current unit to the next unit. LSTM network node is shown below as Figure1.

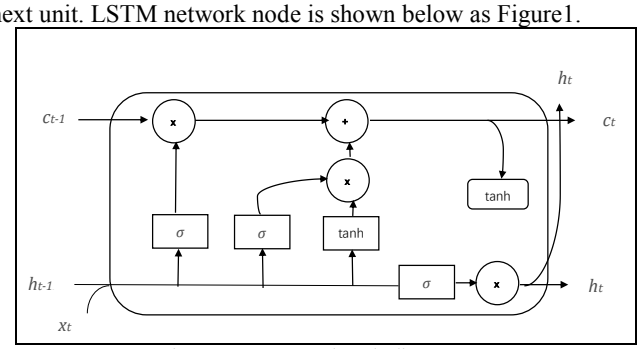



Fig. 1. LSTM network node diagram

The advantage of adding attention mechanism in the model as figure2 shown is that, the first we can learn the connections with the sequence, the second is to reduce the computation complexity of each layer, and the third is to effectively reduce the number of minimum units for parallel computation and training costs.

### B. Process

*1) Signal Processing And Feature Extraction*

Experiment used feature is MFCC. Research shows that the human ear is more sensitive to low-frequency signals. Mel frequency cepstral coefficient (MFCC) features is a nonlinear frequency unit which based on the auditory characteristics of the human ear. It has good robustness and accuracy in speech emotion classification. The relationship between MFCC and frequency is approximate:

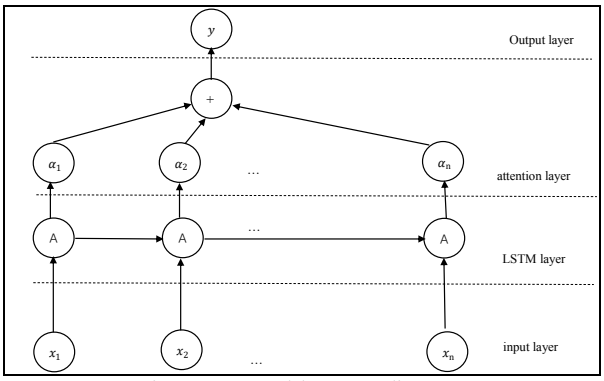



Fig. 2. LSTM model structure diagram

$$Mel(f) = 2595lg(1 + \frac{f}{700}) \quad (1)$$

f is the frequency rate and the single bit is Hz. The basic idea of extracting MFCC: first load the speech, take the speech for emotional analysis, and then determine a suitable speech frame length for Fourier transformation according to the sample rate of the speech. Then normalize audio length, to obtain N_FFT, unified sound range after extraction of MFCC characteristics. The extraction process is shown in Figure 3.

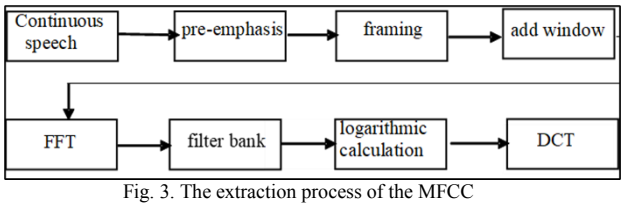



Fig. 3. The extraction process of the MFCC

*2) SVM Specific steps*

*a).Split individual emotions (neutral, happy, sad, angry) in equal training (80%) and test (20%) sets as figure4 shown.*

*b).Get final sets of training and tests waveforms.*

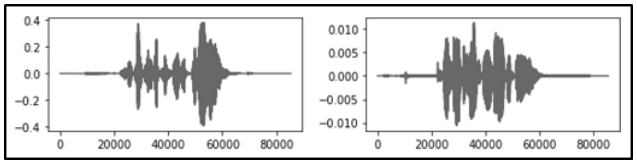


Fig. 4. The sets of training and tests waveforms

*c).Feature Extraction for each sets*

Tab. 1. The shape of the each set

| | Train | Test |
|---|---|---|
| **Audio x Features x Frames** | (1147, 17) | (293,17) |
| **Labels of audio** | (1147,) | (293,) |

*d). Normalize the training features, the result as figure5 shown.*

*e).Build and train the SVM model*

It is implemented with sklearn and C-Support Vector Classification，which is a support vector machine based on libsvm . The advantages of radial basis function: approaching ability, classification ability and learning speed are better than others, and it has the characteristic of simple structure, simple training, fast learning convergence speed. Also, it can approximate any nonlinear function and overcome the local minimum problem.

*3) CNN specific steps*

The specific steps: pre-processing of speech data, extracting 'MFCC' characteristics, building a convolution neural network model, training the characteristics of sample data into the model, and testing on the test set. The convolution layer can be regarded as a fuzzy filter. It performs convolution operation between the feature processed

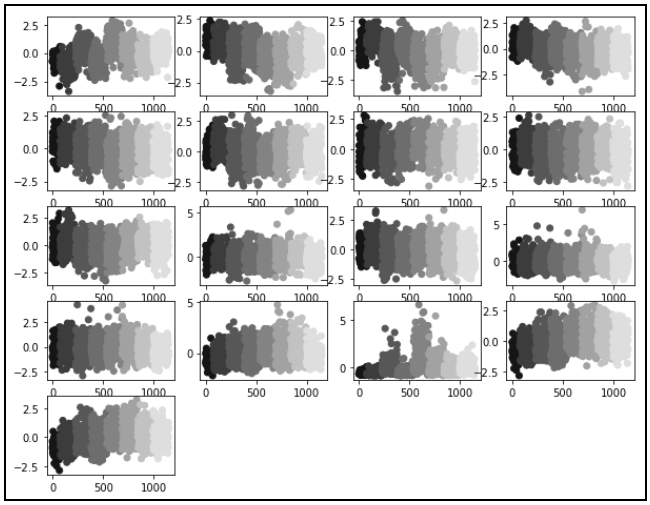


Fig. 5. The scatter of the normalized training features

by the upper layer and the convolution kernel of the current layer, which can enhance the characteristics of the original signal and reduce noise. Finally, the results of convolution calculation are given by activation function. Pooling process can achieve the feature selection and main information extraction of convolution layer information, which can reduce the number of outputs and improve the robustness and generalization performance of the system. In CNN network, the convolution kernel size is 5 and the number is 128. In this layer, the modified linear unit ReLU function is used as the activation function. In the maximum pooling layer, 8 and 3 are the index regions. The model used in this experiment is shown below. Specific steps: pre-processing of voice data, extraction of MFCC characteristics, construction of reel neural network model, sample data characteristics input model for training, testing on the test set, the experimental structure as shown in the figure.

*4) LSTM Specific steps*

*a).Pre-process the dataset, code the emotional characteristics in one-hot;*

*b). Integrate the attention mechanism after the output of the LSTM layer and construct the emotional classification model of the attention mechanism based on LSTM;*

*c). After adding the attention mechanism, connect the base of the fully connected feed-forward neural network;*

*d). Connect the Sigmoid layer to achieve probability output after the output of the fully connected layer;*

*f). Divide the data set for testing and validation.*

### C. Task

Starting from speech emotion recognition, this paper selects RAVDESS dataset as a speech database, first of all, pre-processing the dataset, extracting the characteristic parameters, dividing them into test sets and training sets, and then training them through three models, identifying the emotions. Finally analyzing the experimental results. The speech emotion recognition process for this experiment is shown in the following figure6.

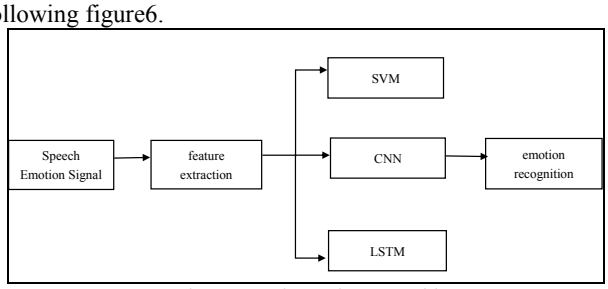



Fig. 6. Speech emotion recognition process

## IV. Experiments And Results Analysis

### A. AP

Three models are used in this experiment SVM, CNN and LSTM, which tested on the RAVDESS dataset. The accuracy of the three models is shown below.

Tab 2 Different model test results

| Method | accuracy |
|---|---|
| SVM | 56.31% |
| CNN | 60.50% |
| LSTM | 67.86% |

### B. *Experiment and Dataset*

This experiment uses the voice file from the RAVDESS dataset. The RAVDESS is a validated multimodal database of emotional speech and song[i]. The database is gender balanced consisting of 24 professional actors, vocalizing lexically-matched statements in a neutral North American accent. Speech includes calm, happy, sad, angry, fearful, surprise, and disgust expressions. Each expression is produced at two levels of emotional intensity, with an additional neutral expression. We use the emotional speech. It contains 1440 files for a total of 24 participants, each of whom conducted 60 experiments.

### *C. Result*

*1) The Result Of SVM*

Bring the test set into the model for testing. During figure7 shown about the generated confusion matrix, the X-axis is the Predicted label, and the Y-axis is the True label. The classifier has the highest classification accuracy of 'disgust', 77 % of which are correctly classified, followed is 'neutral' and 'calm', but the accuracy of the emotion of 'sad' is only 31 %.

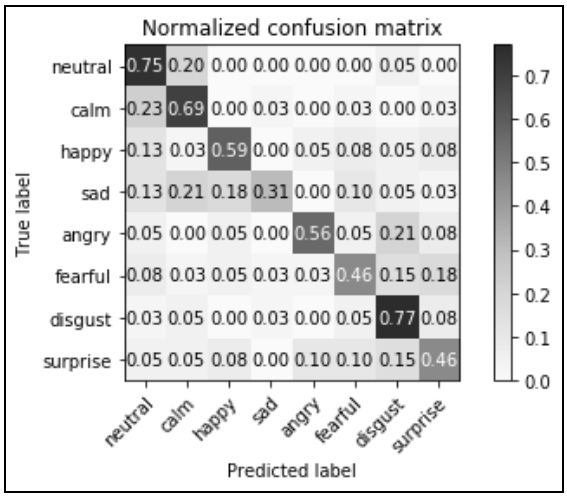



Fig. 7. Normalized confusion matrix

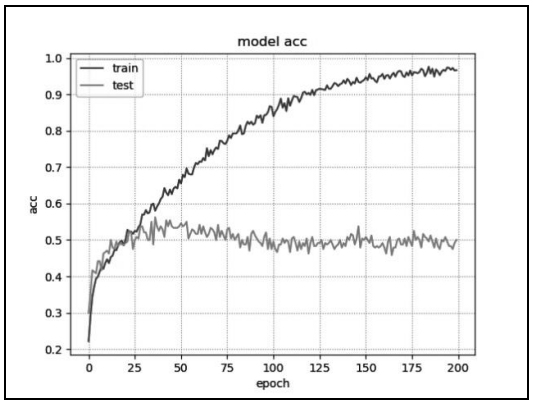



Fig. 8.The model accuracy

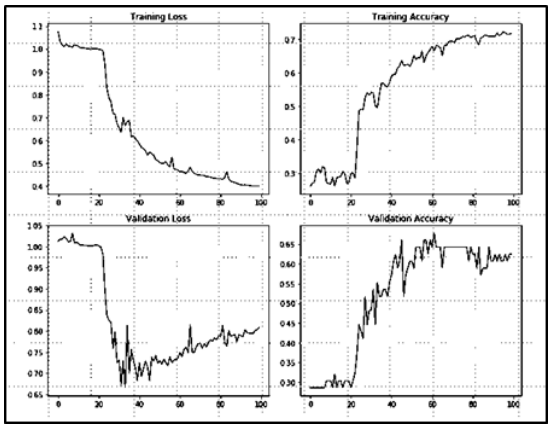


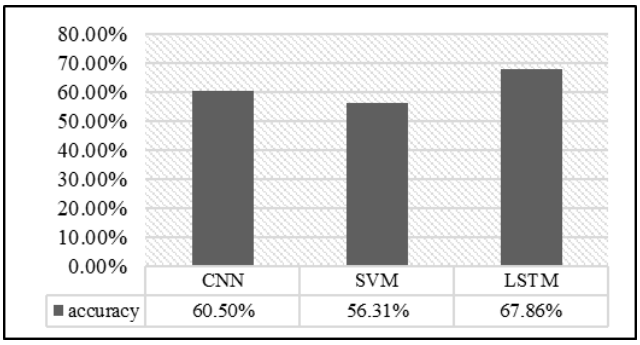

Fig. 9. LSTM model experimental structure

### *D. Analysis*

As Figure7-9 shown the result about SVM, CNN and LSTM. The first column corresponds to the classification model, and the second column corresponds to the accuracy. Based on the RAVDESS data set, it can be seen that LSTM has the highest accuracy, CNN second, and finally SVM. In this experiment ,

data set is small, the recognition accuracy of deep learning is still higher than the shallow learning model in table3.

Tab. 3. Bar graph of the accuracy of three method




### *E. Future Works*

Subsequent optimizations can be considered from the fusion of models, such as building CNN_LSTM-based multi-converse kernel neural networks, or by increasing the attention mechanism for LSTM optimization, making the model more robust and adapted to more data sets. I think if the numbers of the hidden layers as well hidden units are properly set, the DNN may extend the labeling ability of GMM-HMM. So trying to this way is available. In addition, feature selecting can be improved. The speech signal including short-term energy, pitch, and frame has 64 statistical features. The improved Correlation-based Feature Selection can employed to select the most influential feature set.

## VI CONCLUSION

It shows that the deep learning of recognition of speech emotion which done on the small data set has a better performance. Traditional machine learning method can adjust from the selection of parameters, such as the core function parameters of vector machines to improve accuracy. Compared to the classic machine learning algorithm, deep learning is powerful, but for the optimization of the model, the need for complex parameter adjustment is difficult.

## ACKNOWLEDGMENTS

This work is supported by the National Natural Science Foundation of China (Grant No. 61602161,61772180 ), Hubei Province Science and Technology Support Project (Grant No: 2020BAB012 ), The Mater Fundamental Research Funds of Hubei University of Technology(2021046),The Fundamental Research Funds for the Research Fund of Key Lab of Traffic and Internet of Things (WUT: 2015III015-A03 ).

Reference:

[1]Minsky M L.The society of mind[J].Personalist Forum，1987，3（1）：19-32.

[2] Cahn J E.The generation of affect in synthesized speech[J]. Journal of the American Voice I/O Society，1990，8：1-19.

[3] Moriyama T，Ozawa S.Emotion recognition and synthesis system on speech[C]//IEEE International Conference on Multimedia Computing and Systems，1999：840-844.

[4] Liu Z T，Rehman A，Wu M，et al. Speech Emotion Recognition Based on Formant Characteristics Feature Extraction and Phoneme Type Convergence[J]. Information Sciences, 2021.

[5] Kunxia Wang et al. Wavelet packet analysis for speaker-independent emotion recognition[J]. Neurocomputing, 2020, 398 : 257-264.

[6] Dias Issa and M. Fatih Demirci and Adnan Yazici. Speech emotion recognition with deep convolutional neural networks[J]. Biomedical Signal Processing and Control, 2020, 59

[7] Grimm M，Kroschel K，Narayanan S.The Vera am Mittag German audiovisual speech database[C]//Proceedings of the 2008 IEEE International Conference on Multimedia and Expo（ICME），2008：865-868.

[8] McKeow G，Valstar M F，Cowei R，et al.The semaine corpus of emotionally coloured character interaction[C]// Proceedings of the 2010 IEEE International Conference on Multimedia and Expo，2010：1079-1084

[9] Zhang W S，Zhao D H，Chai Z，et al.Deep learning and SVM- based emotion recognition from Chinese speech for smart affective services[J].Software- Practice & Experience，2017，47（8）：1127-1138.

[10] Qin Y Q，Zhang X Y.HMM- based speaker emotional recognition technology for speech signal[J].Advanced Materials Research，2011，230/231/232：261-265.

[11] Swain M，Sahoo S，Routray A，et al. Study of feature combination using HMM and SVM for multilingual Odiya speech emotion recognition[J]. International Journal of Speech Technology, 2015, 18(3):387-393.

[12] Pribil J，Pribilova A，Matousek J.Artefact determination by GMM- based continuous detection of emotional changes in synthetic speech[C]//2019 42nd International Conference on Telecommunications and Signal Processing，2019：45-48

[13] Huang Z，Ming D，Mao Q，et al. Speech Emotion Recognition Using CNN[C]// Acm International Conference. ACM, 2014.

[14] Zhang T，Wu J . Speech emotion recognition with i-vector feature and RNN model[C]// 2015 IEEE China Summit and International Conference on Signal and Information Processing (ChinaSIP). IEEE, 2015:524-528.