Speech emotion recognition using multiple classification models based on MFCC feature values

Hanwen Chen[1, 5]**, Shengping Wu**[2]**, Xinrui Wu**[3]**, Ziran Lin**[4]

[1]Electrical and electronic engineering, Leeds of university, Leeds, LS2 9JT, UK
[2]Faculty of innovation engineering, Macau university of science and technology, Macau, 999078, China
[3]Chengdu Foreign Languages school, Chengdu Foreign Languages school, Chengdu, 610000, China
[4]Victoria Hill School, Victoria Hill School, Kunming, 650000, China

[5]Corresponding author email:1821219686@qq.com
[2]1909863ui011010@student.must.edu.mo
[3]Lily_xinrui@outlook.com
[4]631864407@qq.com

Abstract. In everyday life, people are often influenced by emotions in our behaviour and language. When people use different emotions to articulate the same text, it can have a completely different effect. With the increasing demand for speech emotion recognition (SER), more machine learning and deep learning methods are being used to perform SER. matlab was used as the experimental tool in this study. The Berlin Database of Emotional Speech was used as the database. Feature extraction was performed by Mel Frequency Cepstrum Coefficients (MFCC) and based on these feature values, Support Vector Machines (SVM), K-Nearest Neighbors Algorithm (KNN), Semi-supervised graph-based classifier, ECOC classification model, Naive Bayes model and long short-term memory for predictive classification. The results surface that among the six classifiers, the best sentiment recognition method is LSTM, with 93.2% and 73.03% accuracies in the training and test groups respectively.

Keywords: **s**peech emotion recognition, Matlab, mel frequency cepstrum coefficient, support vector machine, LSTM, machine learning, deep learning, semi-supervised graph-based classifier.

1. Introduction

With the continuous development of communication devices, speech has become one of the most common methods of everyday communication. Emotion is an important component of speech. Emotion recognition plays an active role in medical [1], robotics [2], self-help systems [3] and teaching systems [4]. In the context of evolving artificial intelligence and machine learning techniques, there are a variety of options for implementing speech emotion recognition (SER).

Extracting emotional features is an important part of the SER model. It includes important speech factors such as pitch, volume and frequency. Some of the current popular methods are LPCC formant, line spectral frequency, and Modulation Spectral Features (MSFs) [5][6]. In this study, the sentiment

© 2023 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

extraction method I chose was Mel frequency cepstrum coefficient (MFCC) [7] and used it in conjunction with a variety of classifiers.

In the SER study, there are three main steps.

(1) Selecting a suitable set of databases

(2) Labeling and feature extraction of the data

(3) Using machine learning (ML) or Deep learning (DL) models for classification and prediction, and performance analysis by confusion matrix.

In this study, I used the Berlin Database of Emotional Speech [6] as the database. Emotional features were extracted from four classes of speech by Mel frequency cepstrum coefficient (MFCC). After that, six classifiers such as SVM and LSTM were used for prediction analysis, and the classifier with the best performance in combination with MFCC was finally obtained.

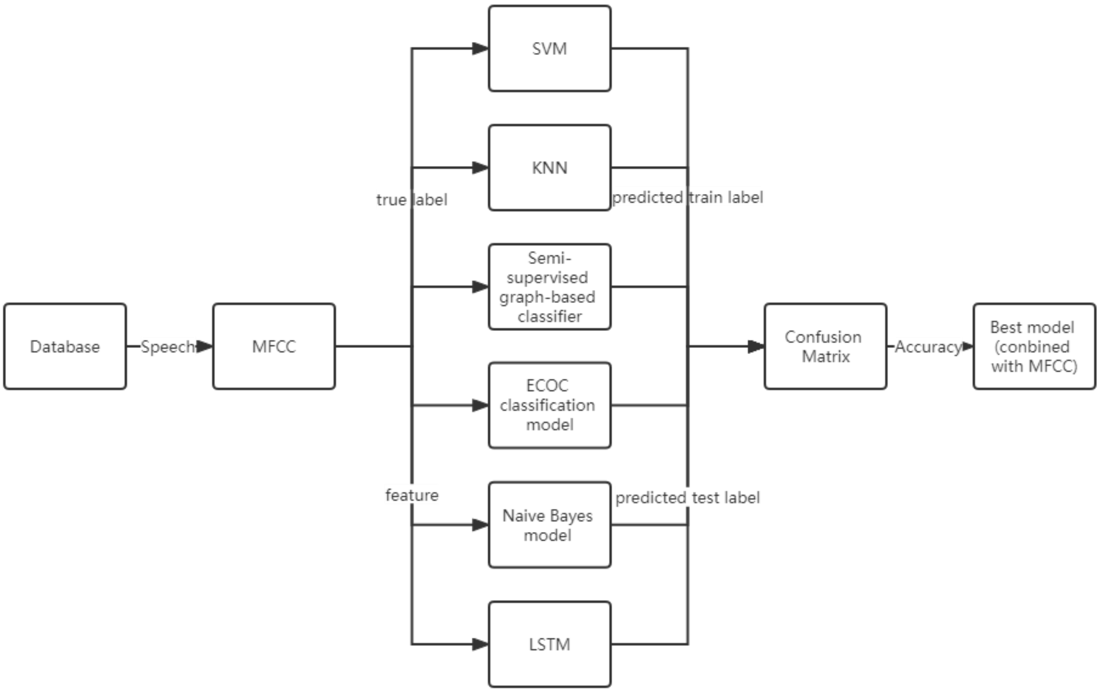


Figure 1. Principle diagram.

2. Background

As technology continues to evolve, so do the expectations for artificial intelligence. In 2017, M. S. Likitha et al. proposed that the acoustic features of the speech signal are unique to each individual. Feature extraction is the process of extracting a small amount of data from a speech signal that can then be used to represent each person speaking. There are many feature extraction methods available, the most widely used is the Mel Frequency Completion Factor (MFCC). In their experiment, data extracted from the speaker's speech signal was used to determine the speaker's feelings. Frequency slope coefficient (MFCC) technology is used to identify the speaker's feelings from the speaker's voice. [8]

In 2019, C. Caihua proposed a multimodal speech emotion recognition method based on SVM. Then, the SVM method was applied to a standard database and continuously optimized, and finally the method was applied to speech emotion recognition with good results. [9] In 2020, Sung-Lin Yeh et al. proposed a Dialogue Emotion Decoding (DED) algorithm which performs emotion decoding in a conversational manner. It processes dialogues as sequences and decodes the mood of each speech in turn using a given recognition engine. The decoder is trained by combining the emotional influence between the speaker and the discourse in the conversation. [10] In 2021, Ainurrochman et al. presented the development of an application capable of recognizing speech emotions using an Extreme Learning Machine (ELM).

They used a dataset from the Toronto Emotional Speech Scale (TESS). The data set contains a total of 2800 data points (audio files) and features high quality audio focused on female voices to ensure data reliability. Speech emotion recognition app was designed as a web application using Golang and Python and built with Extreme Learning Machine and Random Forest to recognize speech emotions. [11] in the same year, M. Sakurai et al. showed that a combination of linguistic and acoustic features is effective in recognizing emotions. The performance difference between the decrypted text and the ASR results was small. This is attributed to the performance of vocal characteristics, primarily linguistic characteristics. To improve the overall performance of the features, you need to improve the performance of sentiment recognition using language features. [12]

3. Materials

This Emo-DB dataset was created by ten actors—five female and five male—faked emotions while producing ten German expressions—five short and five long—that could be utilized in normal conversation and understood in all simulated moods.

It has roughly 500 lines of dialogue spoken by performers in neutral, happy, angry, sad tones. Ten different performers and ten different texts are available to select from.

High-end recording tools and an anechoic chamber were used to create the recordings. Electro-glottograms of sound were recorded in addition to vocals. In a perceptual test about the recognition ability and naturalness of the emotions, the entire database was assessed.

Its identification rate was above 80%, and more than 60% of listeners thought his statements sounded natural. These lines were labeled in a precise transcription with unique markers for voice quality, articulation and articulation settings, and articulation features. [13]

The public has access to the database through the internet (http://www.emodb.bilderbar.info/download/).

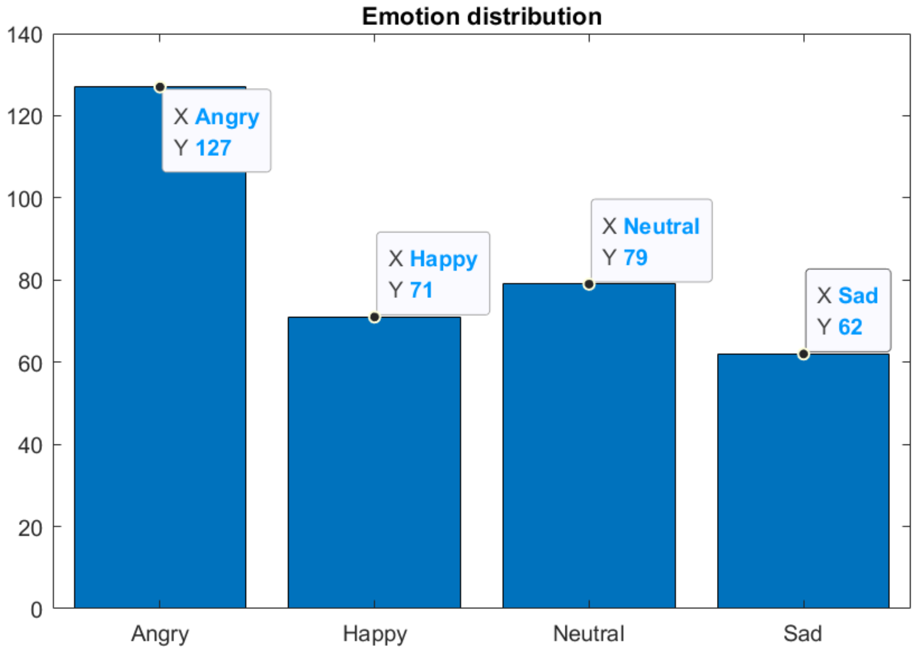


Figure 2. Emotion distribution.

4. Methods

4.1. Feature extraction

The Mel-frequency cepstrum [14] (MFC) differs from the cepstrum since the frequency bands are evenly spaced on the Mel scale, which more slightly approximates the response of something like the human auditory system than the linearly-spaced frequency bands used in the conventional spectrum. Since the Mel-frequency bands are distributed quite uniformly in MFCC (Mel-frequency cepstral coefficients) [15], just like in human speech, MFCC is a set of feature vectors created by storing the physical information (spectral envelope and details) of speech in the field of speech recognition.

Pre-processing, rapid Fourier transformation, Mei filter bank, logarithmic operation, discrete cosine transform, dynamic feature extraction, and other processes are included in the MFCC extraction process.

The specific operations for extracting MFCC are as follows: [16]

(1) Perform a signal's Fourier transform on a windowed extract.

(2) Use triangle overlapping windows or, as an alternative, cosine overlapping windows to map the powers of the spectrum acquired in step one onto the Mel scale.

(3) Calculate the power logs for each of the Mel frequencies.

(4) Consider the list of Mel log powers' discrete cosine transform as a signal.

(5) The amplitudes of the resulting spectrum are the MFCCs.

And basic steps in MFCC calculation

(1) To acquire spectral envelopes in dB, outputs from a logarithmic filter bank are generated and multiplied by 20.

(2) The Discrete Cosine Transform (DCT) of the spectral envelope is used to obtain MFCCs.

(3) Cepstrum the coefficients are calculated as:

$$ci = \sum_{n=1}^{Nf} Sn\,cos[\,i(n-0.5)(\frac{\pi}{Nf})],i = 1,\ 2,\ ..$$

Where the ci is MFCC coefficient, The number of triangular filters in the filter bank is Nf , the number of MFCC coefficients we wish to calculate is L, and the log energy output of the nth filter coefficient is S_n .

About the application of MFCC, applying the MFCC to find elements of bird song is one example of how it is frequently used to extract features in speech recognition systems. [17]

Since normalizing the values reduces the impact of noise in speech recognition systems, MFCC also has the drawback of having values that are not very stable in the face of additive noise.

4.2. Classifier

1) SVM

A supervised learning model and related learning algorithm called SVM are used in machine learning to evaluate data for regression and classification. [18] In order to generate the optimal separating hyperplane in this space, the input vector must first be translated into a high-dimensional feature space, this is how the SVM works. SVM creates a hyperplane or collection of hyperplanes in a high-dimensional or infinite-dimensional space that may be utilized for regression analysis, classification, or other activities. [19] Strong generalization capabilities allow the support vector machine model to process high-dimensional data fast and effectively. As a result, it is frequently employed in the processing and analysis of nonlinear, high-dimensional, small sample data. [20]

2) KNN

The K-nearest neighbors algorithm is one of the simplest machine learning techniques. The nearest neighbor algorithm is a nonparametric statistical technique for classification and regression in the pattern recognition field. [21] By using a vector space model and the closest neighbor algorithm, instances are classified into groups that have a high degree of similarity to one another, and the likelihood that cases in an unknown category will be classified similarly to cases in the known category can be assessed. [21] K is a modest positive integer. If K is equal to 1, the sample is then simply classified in the class of its closest neighbors. Individual numbers under 10 can be used as the K value to prevent combining two classes in order to score the same number of points.[22] Three, five, and seven are the ideal K values.[22]

Since it is effective, nonparametric, and easy to apply, the k-nearest neighbor method can be used in a variety of situations.

3) Semi-supervised graph-based classifier

Semi-supervised learning techniques can be used when only a portion of the data is labelled and the true label of all the data is determined. Semi-supervised graph-based learning can assign labels to unlabelled data. It can estimate a non-linear classification line that is closest to linear, making the otherwise linear approach more flexible [23].

4) ECOC classification model

The Error Correcting Output Code (ECOC) classification model can transform a multi-class learning problem into a binary classification problem.[24]

5) Naive Bayes model

The Naive Bayes model is an algorithm based on Bayes' theorem and the assumption of conditional independence of features, which can be used for dodge classification. For a given training data it first assumes the joint probability distribution of the learned inputs and outputs and obtains new instances, after which the maximum posterior probability is calculated using Bayes' theorem [25].

6) LSTM

Long short-term memory (LSTM) is a special kind of RNN [26], which is mainly designed to solve the gradient disappearance and gradient explosion problems during the training of long sequences. Due to the unique design structure, LSTM is suitable for processing and predicting important events with very long intervals and delays in time series. Simply put, LSTM can perform better in longer sequences than normal RNN.

5. Results

The current study used feature values extracted by MFCC from the Berlin Database of Emotional Speech dataset. The feature values were predicted by Support Vector Machine (SVM), K-Nearest Neighbors Algorithm (KNN), Semi-supervised graph-based classifier, ECOC classification model, Naive Bayes model and long short-term memory (LSTM) for predictive classification of feature values. Afterwards, a comparison of the accuracy of the classifiers for sentiment recognition was carried out. The confusion matrix clearly identifies the strengths and weaknesses of each classifier in recognising the different emotions of joy, anger, neutrality and sadness. (1-Angry; 2-Happy; 3-Neutual; 4-Sad)

1) SVM

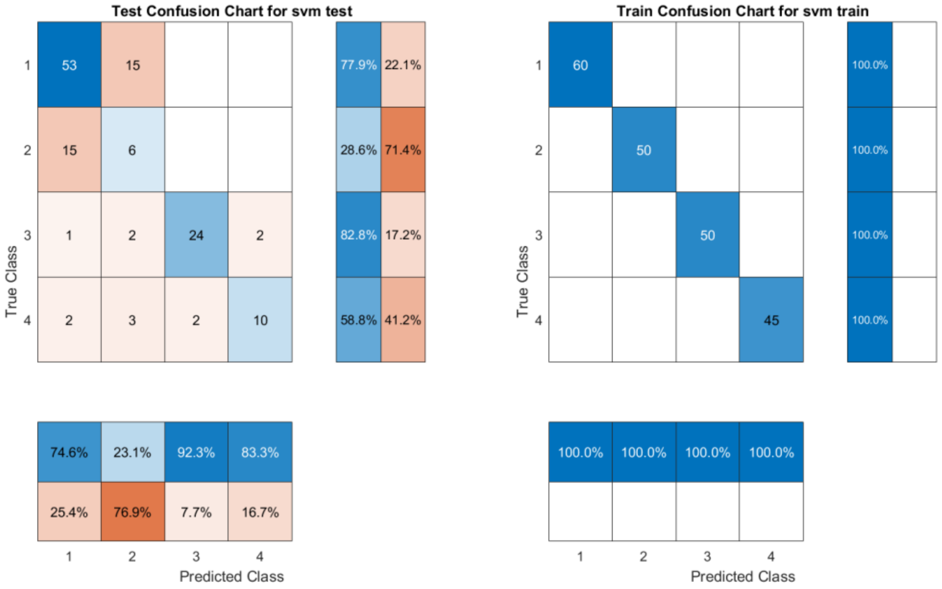


Figure 3. SVM confusion matrix.

2) KNN

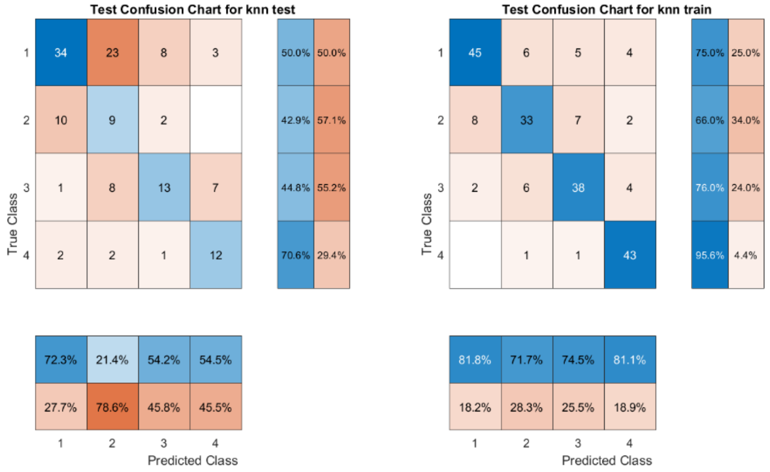


Figure 4. KNN confusion matrix.

3) Semi-supervised graph-based classifier

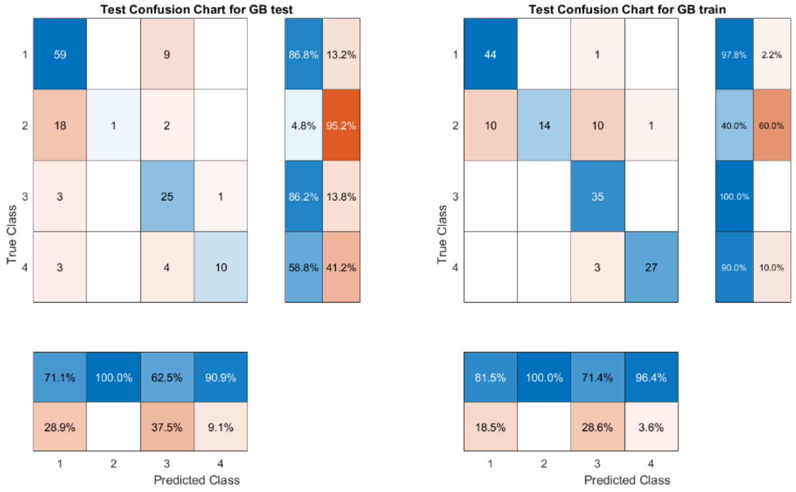


Figure 5. Graph-based classifier confusion matrix.

4) ECOC classification model

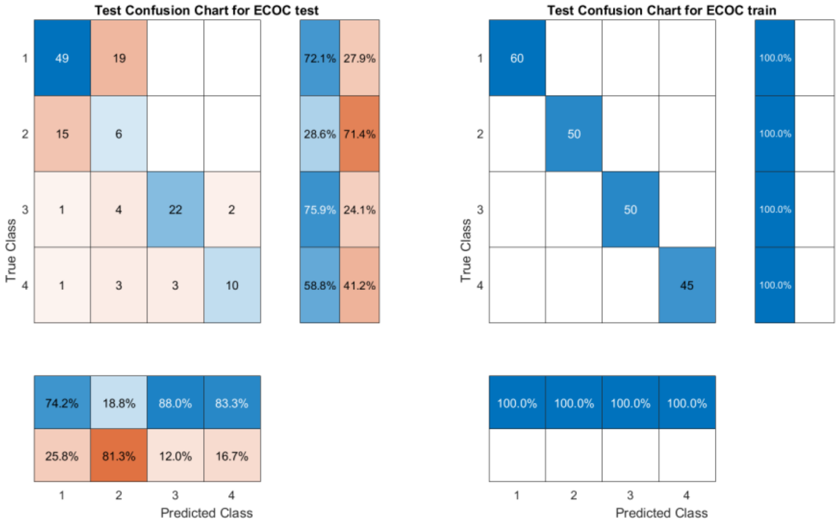


Figure 6. ECOC confusion matrix.

5) Naive Bayes model

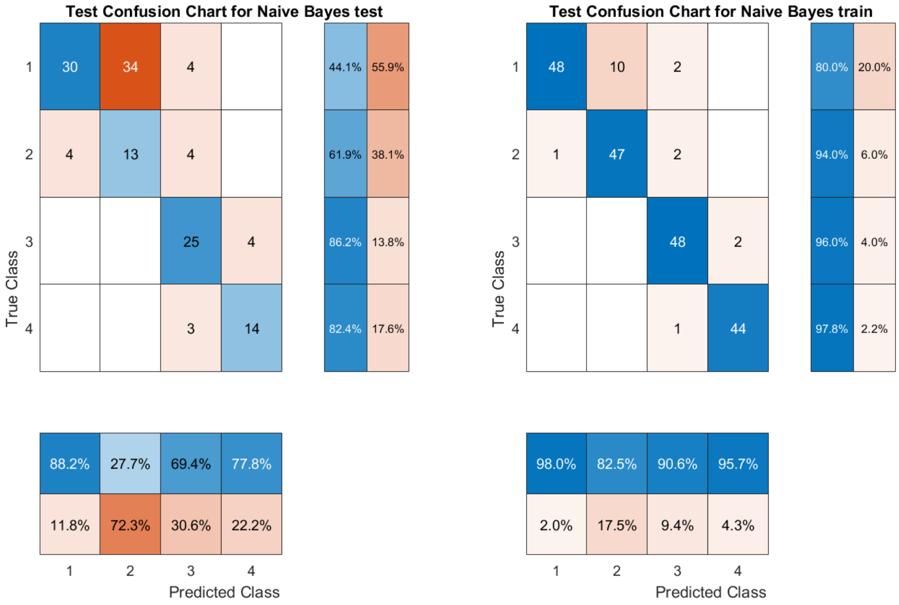
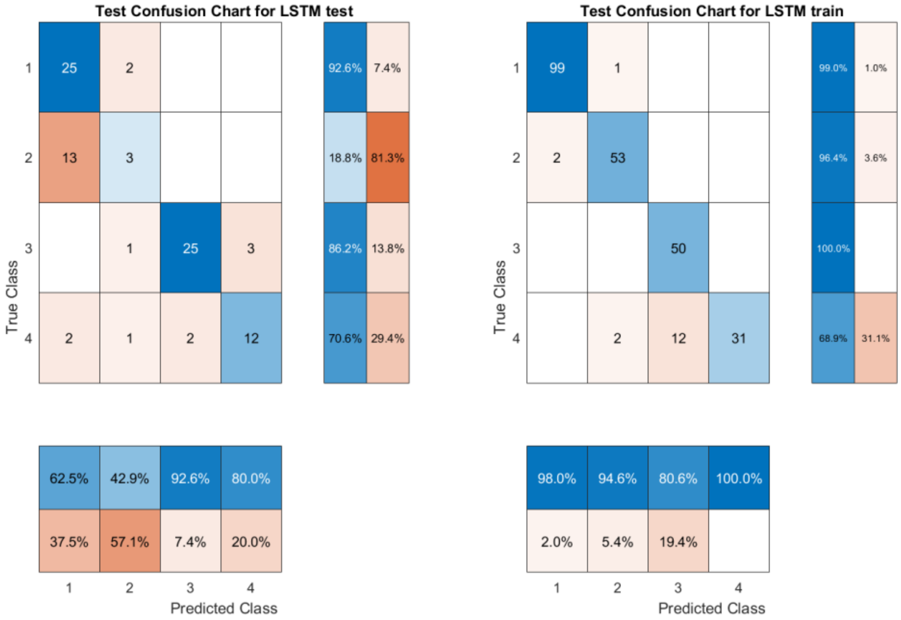


Figure 7. Naive bayes model confusion matrix.

6) LSTM



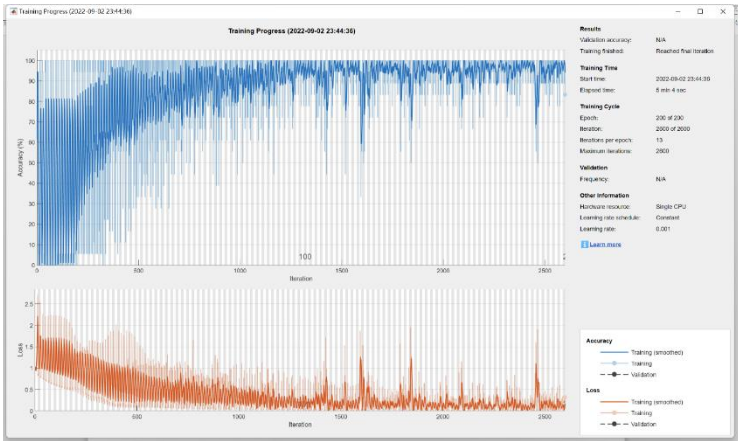
Test Confusion Chart for LSTM test

True Class	1	2	3	4		
1	25	2			92.6%	7.4%
2	13	3			18.8%	81.3%
3		1	25	3	86.2%	13.8%
4	2	1	2	12	70.6%	29.4%
	62.5%	42.9%	92.6%	80.0%		
	37.5%	57.1%	7.4%	20.0%		
Predicted Class	1	2	3	4		

Test Confusion Chart for LSTM train

True Class	1	2	3	4		
1	99	1			99.0%	1.0%
2	2	53			96.4%	3.6%
3			50		100.0%	
4		2	12	31	68.9%	31.1%
	98.0%	94.6%	80.6%	100.0%		
	2.0%	5.4%	19.4%			
Predicted Class	1	2	3	4		

Figure 8. LSTM confusion matrix.



Results
Validation accuracy: N/A
Training finished: Reached final iteration

Training Time
Start time: 2022-09-02 23:44:36
Elapsed time: 5 min 4 sec

Training Cycle
Epoch: 200 of 200
Iteration: 2600 of 2600
Iterations per epoch: 13
Maximum iterations: 2600

Validation
Frequency: N/A

Other Information
Hardware resource: Single CPU
Learning rate schedule: Constant
Learning rate: 0.001

Figure 9. LSTM Training progress.

Table 1.

Classifier	Test Accuracy	Train Accuracy
SVM	68.88%	100%
KNN	50.37%	77.56%
Semi-supervised graph-based classifier	70.37%	88.89%
ECOC classification model	64.44%	100%
Naive Bayes model	60.74%	91.21%
LSTM	73.03%	93.20%

The SVM and ECOC classification models have an accuracy of 100%, the LSTM and ECOC classification models have an accuracy of 93.2% and 91.21% respectively. Semi-supervised graph-based classifier and LSTM performed the best, with their accuracies of 70.37% and 73.03%, respectively. Overall, LSTM is the most suitable model for speech emotion recognition in this study.

Also according to the confusion matrix, we can find that the prediction models often confuse the emotions of Happy and Angry, which indicates that these two emotions directly share more similar speech characteristics. This study is more concerned with the comparison between classification models and the dataset used is not very large. It can be believed that the prediction accuracy of the test groups of all models would have improved somewhat if a larger dataset of data had been used.

For the training set of data to end up with almost 0 error predictions across multiple models, we make the realistic suggestion that for products that perform speech emotion recognition, it is desirable that the product should ideally have the ability to learn over its lifetime. As it works, it continually expands its own dataset based on the user's speech, so that the product's emotion recognition capabilities will continue to grow as the product continues to be used.

6. Conclusion

In this study, we combined MFCC with LSTM as a very reliable choice for the SER system. Compared to some traditional ML models, the DL model LSTM exhibits higher efficiency and performance. This study compares the characteristics of various classifiers in performing SER system design and provides a basis for selection.

However, there are several directions for improvement in this study and we look forward to further research by the reader as follows.

(1) the selected dataset is not large enough in terms of category and data volume, which to some extent affects the reliability of the final model

(2) Consider increasing the amount of feature values and using wavelet scattering, MSF and other methods for feature extraction in addition to MFCC.

(3) Consider using more DL models for SER system construction. Compared with traditional ML models, advanced DL models diversify the needs of emotion recognition. Semi-supervised and fully supervised classifiers will equally show higher efficiency.

There is currently a need for better SER systems in many fields. In scenarios where phone messages are being handled, SER systems can help customer service to help people who are in more urgent need of help more quickly. In hospitals, police stations, etc., SER systems can also provide nursing staff and police officers with a better understanding of the emotional state of patients and suspects in real time.

References

[1] C. Heth et al., "*Psychology: The Science of Behavior*", 2007.

[2] M. Anjum, (2019) "*Emotion Recognition from Speech for an Interactive Robot Agent*," 2019 IEEE/SICE International Symposium on System Integration (SII), pp. 363-368.

[3] Kunxia Wang, Ning An, Bing Nan Li, Yanyong Zhang and Lian Li. (2015) "*Speech emotion recognition using fourier parameters*", IEEE Transactions on Affective Computing, vol. 6, no. 1, pp. 69-75.

[4] L. Jie, Z. Xiaoyan and Z. Zhaohui, (2020) "*Speech Emotion Recognition of Teachers in Classroom Teaching*", 2020 Chinese Control And Decision Conference (CCDC), pp. 5045-5050.

[5] W. Dai, D. Han, Y. Dai and D. Xu, (2015) "*Emotion Recognition and Affective Computing on Vocal Social Media*", Inf. Manag.

[6] E. Bozkurt, E. Erzin, C. E. Erdem and A. T. Erdem, (2010) "*Use of Line Spectral Frequencies for Emotion Recognition from Speech*", 2010 20th International Conference on Pattern Recognition, pp. 3708-3711.

[7] M. S. Likitha, S. R. R. Gupta, K. Hasitha and A. U. Raju, (2017) "*Speech based human emotion recognition using MFCC*", 2017 International Conference on Wireless Communications,

Signal Processing and Networking (WiSPNET), pp. 2257-2260.

[8] M. S. Likitha, S. R. R. Gupta, K. Hasitha and A. U. Raju, (2017) "*Speech based human emotion recognition using MFCC*", 2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET), pp. 2257-2260.

[9] C. Caihua, (2019) "*Research on Multi-modal Mandarin Speech Emotion Recognition Based on SVM*", 2019 IEEE International Conference on Power, Intelligent Computing and Systems (ICPICS), pp. 173-176.

[10] S. -L. Yeh, Y. -S. Lin and C. -C. Lee, (2020) "*A Dialogical Emotion Decoder for Speech Emotion Recognition in Spoken Dialog*", ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6479-6483.

[11] Ainurrochman, I. I. Febriansyah and U. L. Yuhana, (2021) "*SER: Speech Emotion Recognition Application Based on Extreme Learning Machine*", 2021 13th International Conference on Information & Communication Technology and System (ICTS), pp. 179-183.

[12] M. Sakurai and T. Kosaka, (2021) "*Emotion Recognition Combining Acoustic and Linguistic Features Based on Speech Recognition Results*", 2021 IEEE 10th Global Conference on Consumer Electronics (GCCE), pp. 824-827.

[13] Burkhardt, Felix & Paeschke, Astrid & Rolfes, M. & Sendlmeier, Walter & Weiss, Benjamin. (2005). A database of German emotional speech. 9th European Conference on Speech Communication and Technology. 5: 1517-1520.

[14] Childers, D.G., Skinner, D.P., Kemerait, R.C. (1977). "*The cepstrum: A guide to processing*". Proceedings of the IEEE. Institute of Electrical and Electronics Engineers (IEEE). 65 (10): 1428–1443.

[15] "Mel-Frequency Cepstrum." Wikipedia, Wikimedia Foundation, 1 Aug. 2022, https://en.wikipedia.org/wiki/Mel-frequency_cepstrum#cite_note-2.

[16] Sahidullah, Md.; Saha, Goutam (2012). "*Design, analysis and experimental evaluation of block based transformation in MFCC computation for speaker recognition*". Speech Communication. 54 (4): 543–565.

[17] D. Lv et al., (2020) "*Birdsong Recognition Based on MFCC combined with Vocal Tract Properties*", 2020 5th International Conference on Mechanical, Control and Computer Engineering (ICMCCE), pp. 1523-1526.

[18] En.wikipedia.org. (2022). Support-vector machine - Wikipedia. [online] Available at: https://en.wikipedia.org/wiki/Support-vector_machine.

[19] Y. Lin and J. Wang, (2014) "*Research on text classification based on SVM-KNN*", 2014 IEEE 5th International Conference on Software Engineering and Service Science, pp. 842-844.

[20] D. Hao and D. Chai, (2021) "*Application of SVM-KNN intelligent classification prediction model in IT Vocational Education*", 2021 5th Asian Conference on Artificial Intelligence Technology (ACAIT), pp. 312-315.

[21] S. Andarabi, A. Nobakht and S. Rajebi, (2020) "*The Study of Various Emotionally-sounding Classification using KNN, Bayesian, Neural Network Methods*", 2020 International Conference on Electrical, Communication, and Computer Engineering (ICECCE), pp. 1-5.

[22] En.wikipedia.org. 2022. k-nearest neighbors algorithm - Wikipedia. [online] Available at: https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm#Algorithm.

[23] F. Dornaika and Y. El Traboulsi, (2016) "*Learning Flexible Graph-Based Semi-Supervised Embedding*", in IEEE Transactions on Cybernetics, vol. 46, no. 1, pp. 206-218.

[24] S. Shahtalebi and A. Mohammadi, (2018) "*Bayesian Optimized Spectral Filters Coupled With Ternary ECOC for Single-Trial EEG Classification*", in IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 26, no. 12, pp. 2249-2259.

[25] C. Bemando, E. Miranda and M. Aryuni, (2021) "*Machine-Learning-Based Prediction Models of Coronary Heart Disease Using Naïve Bayes and Random Forest Algorithms*", 2021 International Conference on Software Engineering & Computer Systems and 4th International

Conference on Computational Science and Information Management (ICSECS-ICOCSIM), pp. 232-237.

[26] S. Hochreiter and J. Schmidhuber. (1997) Long short-term memory. Neural Computation, 9(8):1735–1780.