WEP Warwick Evans Publishing

International Journal of Computer Science and Information Technology

ISSN: 3005-9682 (Print), ISSN: 3005-7140 (Online) | Volume 3, Number 2, Year 2024
DOI: https://doi.org/10.62051/ijcsit.v3n2.32
Journal homepage: https://wepub.org/index.php/IJCSIT/index

Research on Deep Learning-based Speech Emotion Recognition System

Hui Wang

Hui WangSchool of Electronic Information, Southwest University for Nationalities, China

ABSTRACT

Speech, as one of the primary means of human communication, conveys not only rich semantic information but also the emotional cues of the speaker. With the rapid advancement of deep learning, speech emotion recognition technology has been increasingly integrated into various aspects of daily life, such as telecommunications, automotive systems, and psychological health monitoring, highlighting the critical importance of research in this field. In this study, we propose a parallel architecture for multimodal feature fusion in speech emotion recognition. We design and implement a speech emotion recognition system that addresses challenges such as limited feature diversity and insufficient classification accuracy. To tackle these issues in speech emotion recognition, we introduce a method that integrates multiple features. Spectrograms, capturing local and global speech features through Convolutional Neural Networks (CNNs), are combined with Mel-Frequency Cepstral Coefficients (MFCCs), which extract dynamic features correlated with context using Long Short-Term Memory networks (LSTMs). Our proposed CNN+LSTM parallel structure (CL) fuses spatial and temporal features, yielding significant improvements in accuracy compared to models relying solely on spatial or temporal features, as demonstrated through experiments on the EMO-DB and CASIA databases, with accuracy gains of 6.88% and 7.20%, respectively. Finally, we validate the practicality and efficiency of the entire speech emotion recognition system by porting it to the NVIDIA Jetson Xavier NX platform.

KEYWORDS

Multimodal Feature Fusion; MFCC; Speech Emotion Recognition

1. INTRODUCTION

Speech emotion recognition (SER) is a rapidly advancing and significant research field aimed at endowing intelligent systems with the ability to perceive and accurately interpret human emotions expressed through speech in different contexts. Since the late 20th century, international research on SER has formally commenced. In recent years, Pengcheng Li et al. [1] proposed a novel emotion recognition method based on attention mechanisms, aimed at effective recognition through learning emotional representations in speech. Ziping Zhao and their team [2] introduced a solution that integrates LSTM (Long Short-Term Memory) and CNN (Convolutional Neural Network), extracting deep spectral features from spectrograms and feeding them into a deep neural network (DNN) for precise prediction and recognition of corresponding emotional states. Ho Ngoc-Huynh et al. [3] developed a multimodal emotion recognition model combining CNN and RNN (Recurrent Neural Network), leveraging Mel-Frequency Cepstral Coefficients (MFCC) for spectral feature extraction from speech signals, and using an autoencoder to extract parameter features from text data. These multimodal features are integrated as model input parameters to achieve effective emotion state recognition through precise classification. Song Wenjun cleverly fused CNN with Gated Recurrent Unit (GRU) to deeply mine temporal information inherent in speech signals, achieving a model

Content from this work may be used under the terms of CC BY-NC 4.0 licence (https://creativecommons.org/licenses/by-nc/4.0/).
Published by Warwick Evans Publishing.

accuracy of 80.21% [4]. Li Wenjie et al. addressed the issue of large parameter scales in neural network models by proposing the integration of Depthwise Separable Convolution with LSTM for speech emotion recognition tasks, achieving higher accuracy on the CASIA dataset [5].

2. DATABASE

The EMO-DB dataset is a high-quality emotional speech database recorded at the Berlin Institute of Technology using professional soundproof facilities [6]. It features recordings from 10 professional actors (5 male and 5 female) simulating seven different emotional states: happiness, sadness, disgust, fear, boredom, surprise, and neutrality. In total, there are 535 speech segments collected. These speech files are stored in both WAV and PCB formats, with a sampling frequency of 16 kHz, 16-bit quantization, and are all in mono format.

The CASIA Chinese Emotional Speech Corpus is a significant resource developed by the Institute of Automation, Chinese Academy of Sciences, and has become one of the most widely used standard datasets in the field of Chinese speech emotion recognition. It covers six basic emotional categories: anger, happiness, sadness, fear, surprise, and neutrality. In its construction, four professionally trained speakers recorded under excellent acoustic conditions with no background noise, speaking unified texts in six different emotional states. This ensured a large number of speech samples for each emotion. Overall, it comprises 9,600 speech recordings performed by these four speakers [7].

Both datasets, EMO-DB and CASIA, were divided into training and test sets in an 8:2 ratio for conducting emotion recognition experiments.

3. SPATIO-TEMPORAL NETWORK MODEL

3.1. Preprocessing

First, perform pre-emphasis to reduce imbalance between high and low-frequency components. Next, segment the continuous speech signal into short-time frames. Finally, apply a window function to each frame signal to reduce edge effects.

Pre-emphasis: Boosts the high-frequency components of the speech signal, flattening the overall spectrum. This process effectively reduces the influence of vocal cord vibration and lip-oral effects, thus facilitating subsequent spectral analysis tasks. Essentially, it involves applying a specific high-pass filter, as shown in Equation (1).

$$H(z) = 1 - \mu z^{-1} \tag{1}$$

The range of the pre-emphasis coefficient is 0.9 to 1, typically chosen as 0.97.

Frame Segmentation: In speech signal processing, to better analyze and process the speech signal, it is common practice to segment it into a series of short time intervals called "frames."

To maintain the accuracy and coherence of the analysis, the step betweensuccessive frames (frame shift) is usually set to half the frame length.

Windowing: The essence of windowing lies in performing mathematical convolution between a specifically designed window function ω(n) and the pre-emphasized speech signal, resulting in a new, windowed speech signal. The significance of this process lies in delicately adjusting the weight distribution of the edges of each frame signal, effectively mitigating signal discontinuities caused by frame segmentation issues, as shown in Equation (2).

$$S_{\omega}(n) = \omega(n)*s(n) \tag{2}$$

Where $\omega(n)$ is the window function, s(n) is the speech signal, and $S_\omega(n)$ is the speech signal after convolution

The Hamming window, to a certain extent, better preserves the content of the high-frequency components of the speech signal. This plays a positive role in reducing or even overcoming energy leakage phenomena. The formula for the Hamming window is shown in Equation (3).

$$\omega(n) = \begin{cases} 0.54 - 0.46\cos[2\pi n(N-1)], & 0 \le n \le N-1 \\ 0 & , \quad \text{others} \end{cases} \tag{3}$$

3.2. Spatio-temporal Network Model

Through in-depth exploration, it was discovered that MFCC features, which are continuously distributed in the time domain, can represent speech characteristics in the form of continuous time-domain sequences, fully considering the correlations between contexts. Therefore, in this chapter, MFCC features are introduced as another important input for emotion recognition. However, when dealing with sequential data, Recurrent Neural Networks (RNNs) exhibit higher adaptability due to their ability to capture temporal dependencies. Nevertheless, RNNs commonly face issues such as vanishing or exploding gradients when handling long sequences, severely impacting model recognition and accuracy. Therefore, this chapter also introduces a variant of RNNs—Long Short-Term Memory networks (LSTMs). LSTMs, with their ingeniously designed memory cells and gate mechanisms, effectively store and extract long-term dependency features from sequential data, thus successfully overcoming the gradient vanishing or exploding problems of RNNs. Figure 1 illustrates the CL network architecture.

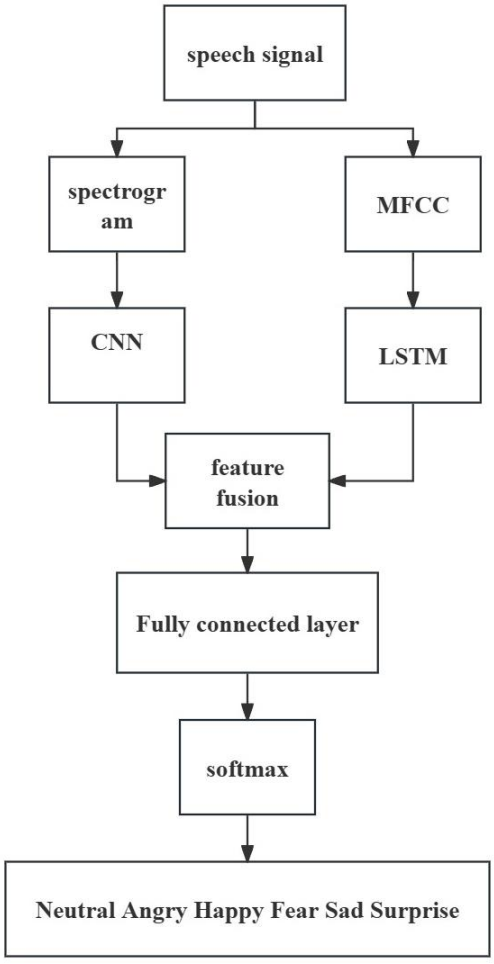

Figure 1. CL Network Architecture

3.3. Feature Extraction

Firstly, spectrogram and MFCC features are extracted separately from the speech signal. These two types of features are then inputted into CNN and LSTM models respectively for feature extraction. The extracted features are normalized using Min-Max normalization to ensure uniformity and comparability of the fused features for subsequent comprehensive analysis and emotion recognition.

Spectrogram Feature: The speech signal is preprocessed and then subjected to Fast Fourier Transform (FFT) to convert the signal from the time domain to the frequency domain. The magnitude squared spectrum is computed for each frame to obtain the power spectral density. These power spectral densities from each frame are arranged in chronological order to form a complete spectrogram. The formula for Fast Fourier Transform is shown in Equation (4).

$$X_i(k) = \sum_{n=0}^{N-1} x_i(n) e^{-\frac{j2\pi kn}{N}} \qquad 0 \le k \le N-1 \tag{4}$$

Through spectrograms, one can intuitively observe the entire process of how frequency components in audio data evolve over time. This reveals the changing patterns of sound over time and the various frequency components it contains. As shown in Figure 2 below, spectrograms corresponding to six emotions are generated from speech data in the CASIA emotional speech database used in this study.

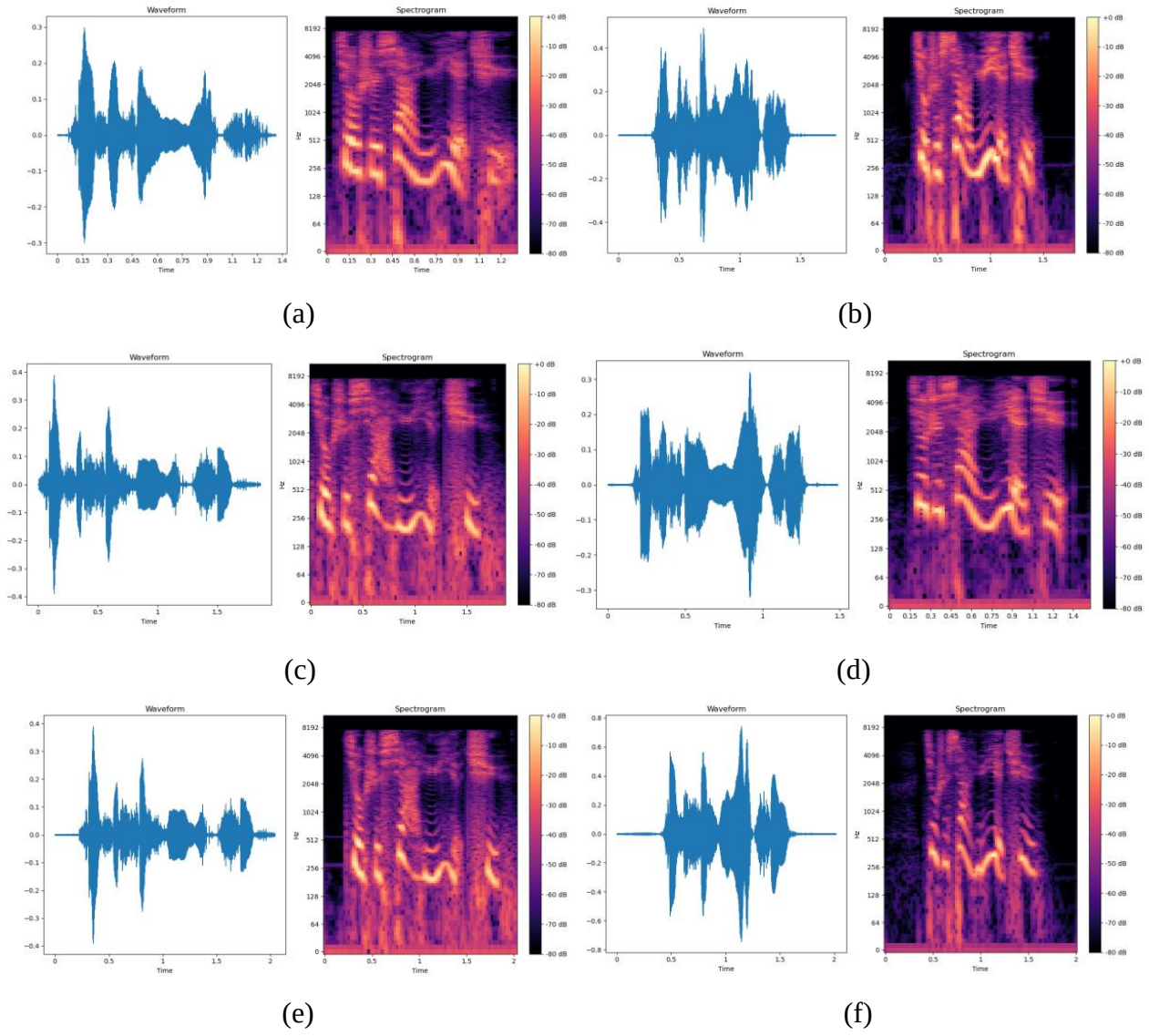

Figure 2. Shows waveform and corresponding spectrogram examples of six emotions from the CASIA Emotional Speech Database.

Panel (a) depicts the waveform and spectrogram for neutral emotion, (b) for angry emotion, (c) for fearful emotion, (d) for happy emotion, (e) for sad emotion, and (f) for surprised emotion.

MFCC Feature: MFCC performs remarkably well in speech recognition and emotion analysis due to its focus on capturing the core features of speech signals within the frequency range sensitive to the human ear, rather than all minor details. It can partly ignore high-frequency information that is not critical for speech recognition or emotion analysis, enhancing its ability to distinguish the essential features of speech signals. Even in the presence of noise interference, MFCC maintains robustness and efficient recognition capabilities. The computation of MFCC involves preprocessing steps such as pre-emphasis, framing, windowing, and Fast Fourier Transform (FFT). Subsequently, the linear frequency spectrum is transformed into a Mel-scale spectrum through a set of Mel-scale filter banks. The Mel-frequency spectrum is then logarithmically scaled and subjected to Discrete Cosine Transform (DCT), with most higher-order coefficients discarded to retain a low-dimensional coefficient set, which forms the MFCC feature vector. Figure 3 illustrates the process of extracting MFCC features.

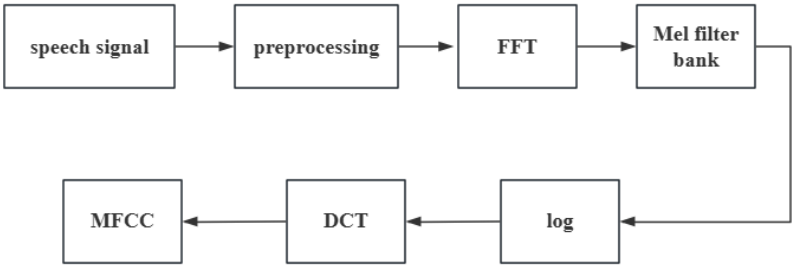

Figure 3. MFCC Feature Extraction Process

As shown in Figures 4 and 5, MFCC feature diagrams extracted from the CASIA database for the emotion 'happy'.

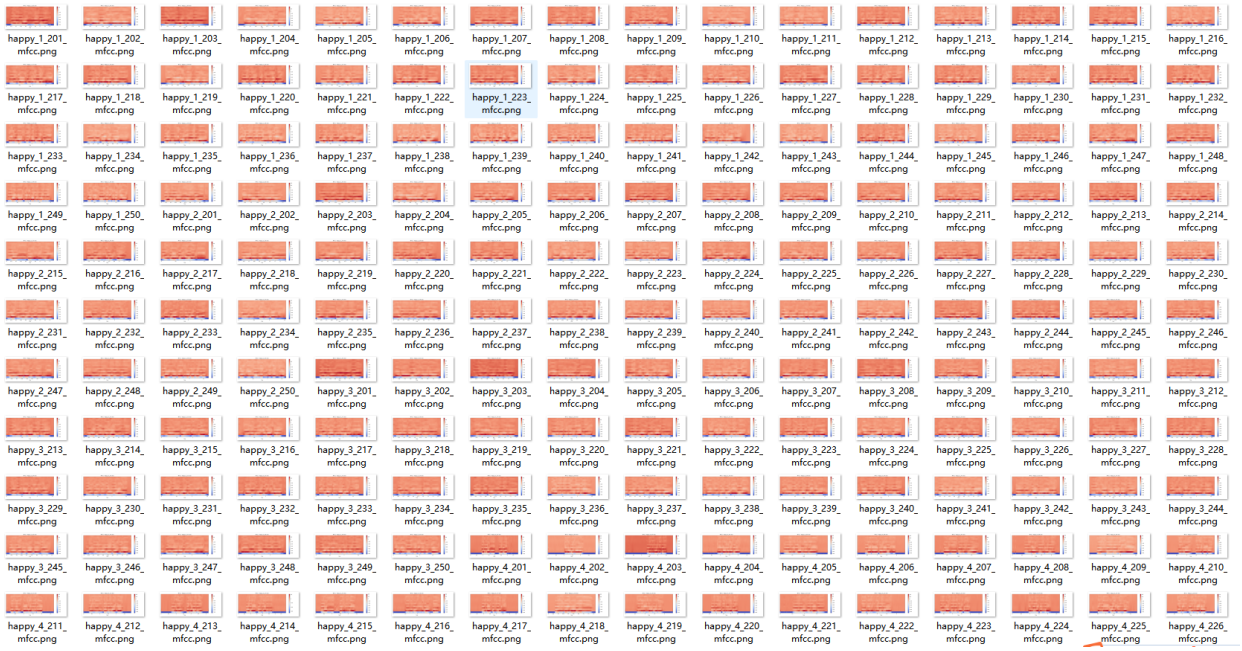

Figure 4. extracts the MFCC feature map of happy

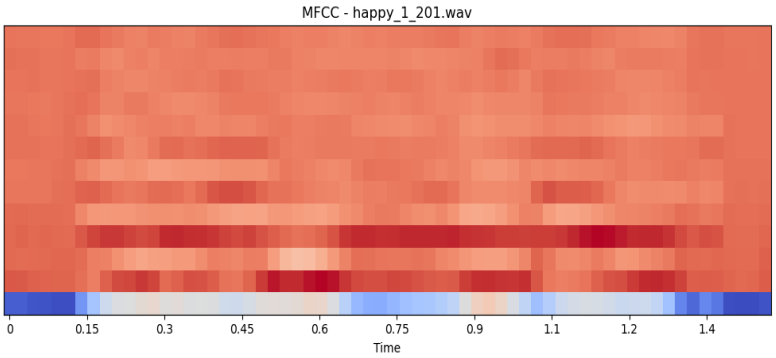

Figure 5. MFCC Feature Diagram for 'happy'

3.4. Model Training Parameters

Based on a GPU platform, training was conducted using the PyTorch framework. The model architecture consists of a 2-layer LSTM with 128 hidden units per layer, which strikes the optimal balance between model complexity and recognition performance. The network training parameters are as follows: batch size set to 5, training epochs set to 200. The gradient optimization algorithm used is Adam with a learning rate (Lr) of 0.0001, and dropout is set to 0.2.As shown in Table 1, these are the parameters of the CL network model.

Table 1. CL Network Model Parameters

Input Layer	Convolution Kernel	Stride	Number of Channels (Depth)
Convolutional Layer 1	3×3	1×1	32
Max Pooling Layer	2×2	2×2	32
Convolutional Layer 2	3×3	1×1	64
Max Pooling Layer	2×2	2×2	64
Convolutional Layer 3	3×3	1×1	64
Max Pooling Layer	2×2	2×2	64
Recurrent Layer	2 layers of LSTM with each layer having 128 hidden units		
Fully Connected Layer 1	128		
Fully Connected Layer 2	64		
Output Layer (Softmax)	Number of emotion categories		

4. EXPERIMENTAL RESULTS

In this experiment, the spectrogram matrices obtained had dimensions of [128, 128], while the MFCC features were represented as [128, 13]. Subsequently, the normalized spectrogram was fed into a Convolutional Neural Network (CNN) model, and the MFCC features were input into a Long Short-Term Memory (LSTM) model. The features extracted from each model were then flattened into feature vectors, normalized, and concatenated for fusion. Finally, the fused features were input into a fully connected layer for further processing and final output.

To validate the effectiveness of this fusion strategy, the experiment designated the spectrogram input to the CNN model and the MFCC features to the LSTM model as the experimental group. Additionally, three control groups were established: Control Group 1 utilized only the spectrogram input, extracting features through CNN and LSTM models separately before fusion. Control Group 2 employed only MFCC features, similarly extracting and fusing features through CNN and LSTM models. Control Group 3 used the spectrogram input for the LSTM model and MFCC features for the CNN model, followed by feature fusion.

By comparing and analyzing the results of these experimental and control groups, we can evaluate the comprehensive performance of the P-CLSTM model in speech emotion recognition tasks and assess its advantages over traditional methods. This preparation is crucial for subsequent deployment on embedded platforms. Tables 2 and 3 show the recognition results of different input features on the EMO-DB and CASIA databases, respectively. Figures 6 and 7 depict the confusion matrices of the experimental group on the EMO-DB and CASIA databases.

Table 2. Recognition results of different feature inputs on the EMO-DB library

Model	Accuracy for Each Emotion Category (%)							Overall Accuracy (%)
	happy	sad	angry	boredom	fear	disgust	neutral	
Control Group 1	77	86	85	82	87	85	83	83.11
Control Group 2	69	83	82	76	82	87	84	81.23
Control Group 3	74	72	83	79	87	85	83	80.83
Experimental Group	85	72	87	89	81	92	88	85.72

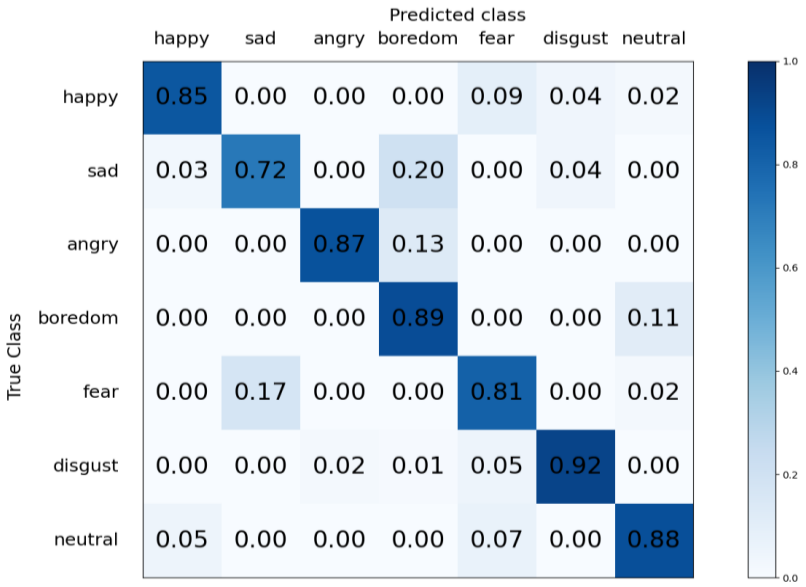

Figure 6. Confusion Matrix of the Experimental Group on the EMO-DB Database

Table 3. Recognition results of different features input on the CASIA database

Model	Accuracy for Each Emotion Category (%)						Overall Accuracy (%)
	happy	angry	fear	sad	neutral	surprise	
Control Group 1	90	75	77	79	82	74	79.35
Control Group 2	80	83	75	77	86	75	78.67
Control Group 3	75	76	76	82	77	75	77.57
Experimental Group	90	69	85	88	81	79	82.86

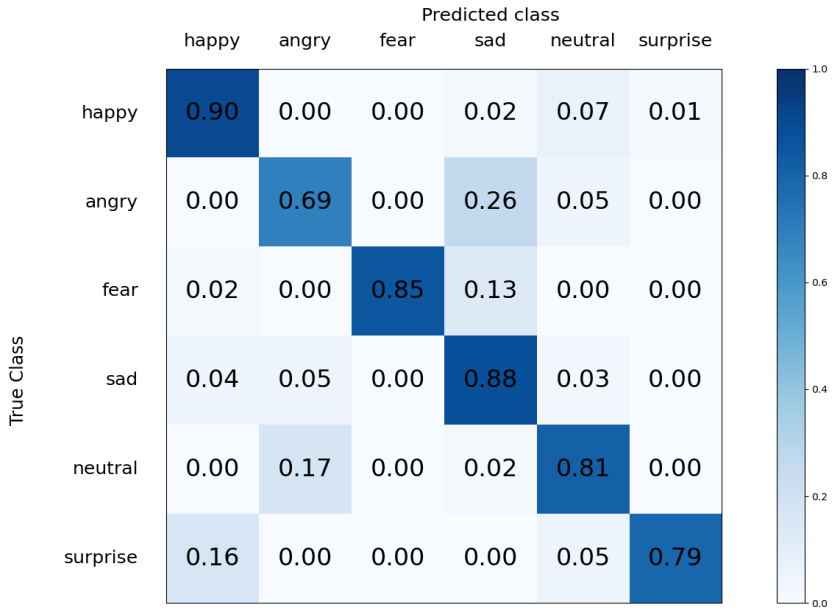

Figure 7. Confusion matrix of the experimental group on the CAASIA library

5. APPLICATION

A speech emotion recognition system incorporating the CL model has been designed and developed. Initially, the speech emotion recognition model was constructed using the PyTorch deep learning framework. Subsequently, an intuitive and user-friendly system interface was designed and implemented using PyQt5 tools. Finally, the trained and optimized model was deployed on the Jetson Xavier NX embedded hardware platform, and its practicality was validated through experimental comparisons.

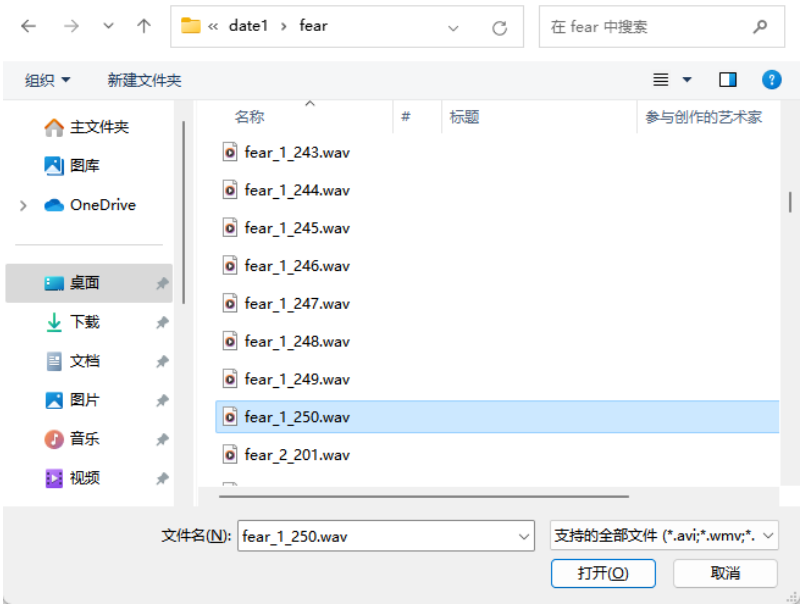

Figure 8. Interface of Speech Selection Testing CASIA Database

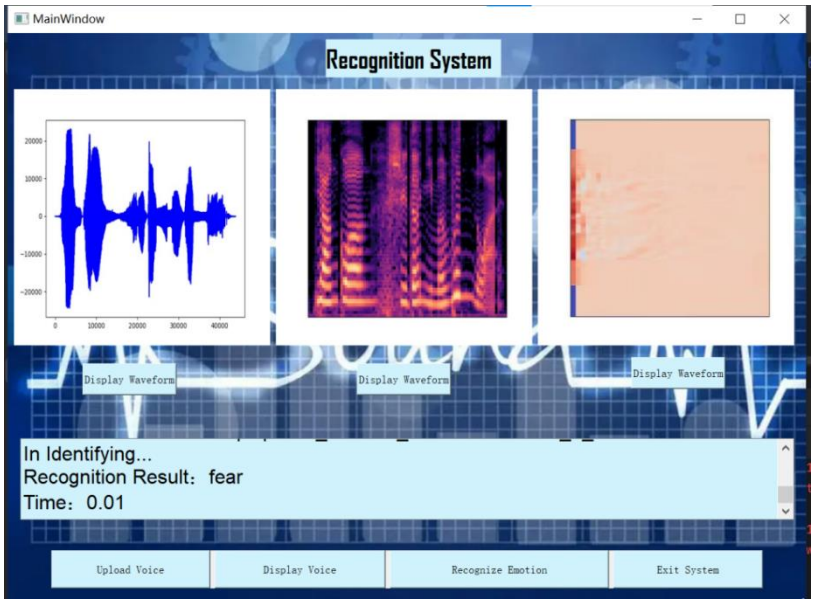

Figure 9. Results Display (CASIA Database)

Figure 8 shows the interface for selecting speech from the CASIA database. In Figure 9, the final recognition result is "fear," achieved in 0.01 seconds. It matches the selected true emotion, indicating accurate test results.

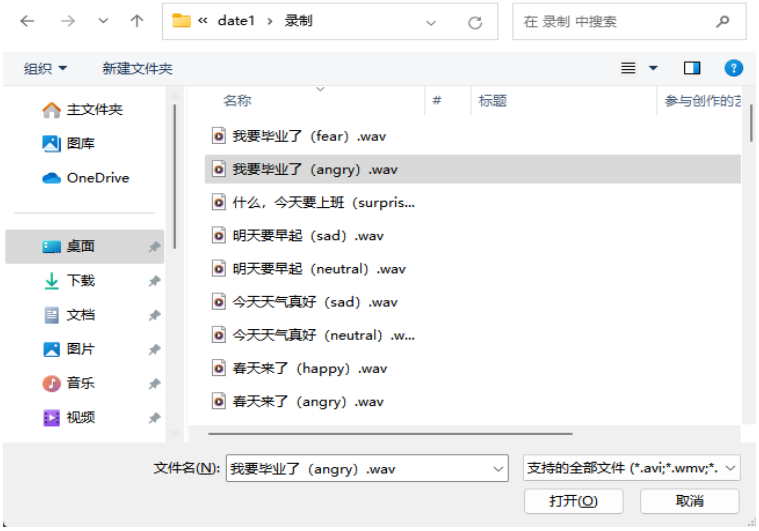

Figure 10. Interface for Selecting Sample Speech from the Database

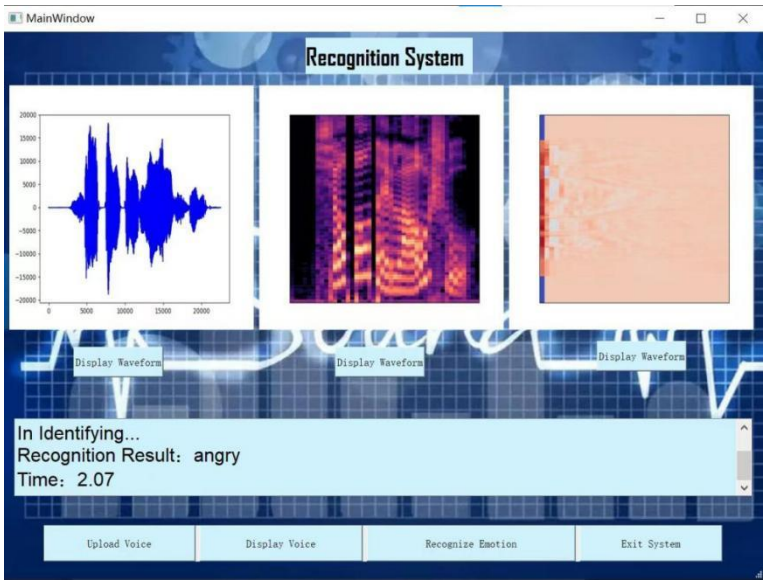

Figure 11. Results Display (Custom Database)

Figure 10 shows the interface for selecting sample speech from the database. In Figure 11, the results display shows the recognition result as "angry," achieved in 2.07 seconds. It matches the selected true emotion, indicating accurate test results. Figure 12 depicts the desktop display of the Jetson Xavier NX system.

Figure 12. Jetson Xavier NX System Desktop Display

Table 4. Comparison of the runtime of voice emotion recognition on the desktop and Jetson Xavier NX

Test Device	Recognition Time for 120 Sentences	Average Recognition Time
P-CLSTM (On desktop)	6.82s	0.057s
P-CLSTM (Jetson Xavier NX)	5.76s	0.048s

Table 5. Comparison of speech emotion recognition rates between desktop and Jetson Xavier NX

Test Device	Angry	Happy	Fear	Surprise	Neutral	Overall Accuracy
P-CLSTM (On desktop)	14/20	19/20	18/20	15/20	15/20	81.6%
P-CLSTM (Jetson Xavier NX)	12/20	17/20	15/20	17/20	17/20	79.2%

6. CONCLUSION

Considering the inherent temporal characteristics of speech signals, this study further incorporates MFCC features and Long Short-Term Memory (LSTM) models. MFCC features enable more accurate capture and representation of temporal information in speech, while LSTM networks excel in efficiently extracting key features from time-series data. Based on this, a parallel CNN+LSTM network model is designed and implemented.

By applying this model, a visualized speech emotion recognition system is constructed and successfully deployed on the Jetson Xavier NX embedded device. It is noteworthy that even on such an embedded platform, the system maintains stable and efficient operational performance.

REFERENCES

[1] LI P, SONG Y, MCLOUGHLIN I V. An attention pooling based representation learning method for speech emotion recognition [J]. 2018.

[2] ZHAOZ, ZHAOY, BAOZ. Deep spectrum feature representations for speech emotion recognition [C]//Proceedings of the Joint Workshop of the 4thWorkshop on Affective Social Multimedia Computing and first Multi-Modal Affective Computing of Large-Scale Multimedia Data. 2018:27-33

[3] Ho N H, Yang H J, Kim S H, et al. Multimodal Approach of Speech Emotion Recognition Using Multi-Level Multi-Head Fusion Attention-Based Recurrent Neural Network [1]. IEEE Access, 2020, 8(8):61672-61686.

[4] Song Wenjun. Research on Speech Emotion Recognition Based on Neural Networks and Attention Mechanism [D]. Changchun: Changchun University of Science and Technology, 2021: 26-28.

[5] Li Wenjie, Luo Wenjun, Li Yiwen. Research on Speech Emotion Recognition Based on Separable Convolution and LSTM [J]. Information Technology, 2020, 44(10): 61-66.Burkhardt F, Paeschke A, Rolfes M, et al. A database of German emotional speech [C]//Ninth European Conference on Speech Communication and Technology. 2005.

[6] Burkhardt F, Paeschke A, Rolfes M, et al. A database of german emotional speech [C]. INTERSPEECH, Lisbon, 2005:1517-1520.

[7] CHEN M, ZHAO X. A multi-scale fusion framework for bimodal speech emotion recognition [C] // Proc. Interspeech. 2020:374-378.

[8] CAI R, GUO K, XU B.Meta Multi-task Learning for Speech Emotion Recognition [J]. Proc.Interspeech 2020, 2020:3336-3340.

[9] Peng Z, Lu Y, Pan S, et al.Efficient Speech Emotion Recognition Using Multi-Scale CNN and Attention [C]. 2021 EEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada. 2021:3020-3024.

[10] Hema C, Marquez F PG. Emotional Speech Recognition Using CNN and Deep Learning Techniques [J]. Applied Acoustics, 2023, 211:109492.

[11] Wang Shui Hua, Phillips P, Sui Yuxiu, et al. Classification of Alzheimer's disease based on eight-layer convolutional neural network with leaky rectified linear unit and max pooling [J]. Journal of medical systems, 2018, 42(5):85.

[12] J. S. Wang, X. F. Du, L. L. He. Evaluation and modeling of automotive transmission whine noise quality based on MFCC and CNN [J]. Applied Acoustics, 2021(12):172-184.

[13] Li H, Zhang X, Duan S, et al. Speech emotion recognition based on bi-directional acoustic–articulatory conversion [J]. Knowledge-Based Systems, 2024, 299112123-112123.

[14] Samson A, Serestina V, Adekanmi A. An enhanced speech emotion recognition using vision transformer [J]. Scientific Reports, 2024, 14(1):13126-13126.

[15] Yan J, Li H, Xu F, et al.Speech Emotion Recognition Based on Temporal-Spatial Learnable Graph Convolutional Neural Network [J]. Electronics, 2024, 13(11):

[16] Kishor B, Mohanaprasad K. Speech Emotion Recognition Using Generative Adversarial Network and Deep Convolutional Neural Network [J]. Circuits, Systems, and Signal Processing, 2023, 43(4):2341-2384.

[17] Masum M B, Likhon M S, M. A. H. A. KBES: A dataset for realistic Bangla speech emotion recognition with intensity level [J]. Data in Brief, 2023, 51109741-109741.

[18] Bhanusree Y, Kumar S S, Rao K A. Neural network-based blended ensemble learning for speech emotion recognition [J]. Multidimensional Systems and Signal Processing, 2022, 33(4):1323-1348.

[19] Vasuki P. Design of Hierarchical Classifier to Improve Speech Emotion Recognition [J]. COMPUTER SYSTEMS SCIENCE AND ENGINEERING, 2023, 44(1):19-33.

[20] Wang Rui. Research on Speech Emotion Recognition Method Based on Deep Learning [D]. Beijing University of Posts and Telecommunications, 2023. DOI: 10.26969/d.cnki.gbydu.2023.002773.