Journal of Network Intelligence ©2025 ISSN 2414-8105
Taiwan Ubiquitous Information Volume **10**, Number **4**, November **2025**

Emotion Description Model Based on Speech Emotion Recognition for Museum Digital Exhibition Human-Computer Interaction

Wei-Wei Li, Ling Shi*

The School of Arts
Anhui Xinhua University
Hefei 230088, P. R. China
564543979@qq.com, 345124594@qq.com

Li-Xin Chen

University of Central Lancashire
Preston PR1 2HE, UK
LChen30@uclan.ac.uk

*Corresponding author: Ling Shi

Received August 28, 2024, revised March 29, 2025, accepted September 14, 2025.

Abstract. *This paper aims to study and propose a human-computer interaction method for digital exhibitions in museums based on speech emotion recognition. In digital exhibitions in museums, human-computer interaction is a key technology that can provide a more immersive and personalized visiting experience. As an important human-computer interaction method, speech emotion recognition can provide visitors with a more personalized and targeted exhibition experience by recognizing their emotional state. This article first analyzes the problems and challenges in digital exhibitions in museums, including the uniformity of visiting experience and the limitations of interactive effects. Subsequently, speech emotion recognition technology was introduced to recognize the emotional state of visitors by analyzing their speech features and emotional expressions. Aiming at the problems of low accuracy and insufficient generalization ability of current speech emotion recognition models, a speech emotion recognition method architecture based on Time-frequency Feature Dual Fusion (TF-DF) is proposed. There is complementarity between time and frequency features, and fully utilizing their information can help improve feature representation ability and avoid feature redundancy. A feature dual fusion module has been designed. Applying this model in a digital museum system can enhance the museum's service experience by accurately recognizing tourists' voice emotions.*

Keywords: speech emotion recognition; feature fusion; deep learning; museum

1. **Introduction.** Traditional museums follow the digital trend and widely apply Internet technology, mobile network technology, big data and other technologies to museums, making the original solemn and dignified museum lively, interesting, fashionable and diversified, "making cultural relics live in a real sense". Strengthening the display and expression of traditional cultural relics through modern technological systems can provide visitors with interesting cultural and interactive experiences [1].

With the continuous development of museums, they have shifted from being "collection oriented" to "experience oriented", and are gradually moving towards "emotion oriented" development, combining the educational mission of museums with the emotional participation ability of audiences, so that audiences can discover new content in a "game" way

[2]. This is the concept of a new museum developed since the ”New Museum Movement” broke the old traditional museum model in 1980. Based on new technology driven interactive experiences, it emphasizes and advocates providing visitors with experiences that involve physical, psychological, and emotional aspects. In addition to its mission of museum education, digital museums also resonate emotionally with visitors, providing them with a profound experience and impression, and meeting their deeper emotional needs will be an inevitable development trend.

In interpersonal communication, appropriate empathy is more conducive to the advancement and deepening of relationships, and can make the other party perceive your genuine understanding and empathy. Museums, as one of the important places that easily generate empathy, use digital means to vividly restore historical scenes, allowing audiences to immerse themselves in understanding the characters and events in the historical process, generating collisions of ideas and emotional resonance, thus leaving a deeper impression on the audience through the digital experience of museums.

At present, the interaction between tourists and tourist destinations transcends the limitations of time and space. Studying and analyzing the digital experience of museums, and promoting the guidance path for audiences to generate online compatible behaviors, can not only open up the visibility and expand the influence of museums, but also play a good role in cultural dissemination and exchange [3]. The online audience engagement behavior triggered by the digital experience of museums mainly involves the audience’s evaluation, feedback, suggestions, and recommendations on various services provided by the museum. The audience can gain a deeper understanding of the exhibition content and enhance their visiting experience by interacting with digital exhibits and exploring virtual exhibition halls; On the other hand, online audience engagement behavior effectively increases the visibility and exposure of museums, drives cultural exchange and dissemination, triggers interaction and discussion on social media, and strengthens the close connection between museums and audiences. In addition, audience feedback and suggestions provide valuable information to museums, helping them improve exhibitions, services, and facilities to meet the expectations and needs of more consumers, and enhance the quality and attractiveness of museums. In the field of Human Computer Interaction (HCI), research is not limited to recognizing individual phonemes or sentences of specific speakers, and recognizing hidden emotional states in speech has become a new trend in speech recognition research [4]. The core goal of Speech Emotion Recognition (SER) technology is to automatically recognize the emotional state of speakers [5], which is achieved through the analysis of emotional information in human speech.

1.1. **Related work.** Speech emotion recognition technology occupies a crucial position in the field of speech, and its research can be traced back to the 1970s. In 1972, Williamsd and colleagues first demonstrated [6] that emotional states can be effectively recognized through the speaker’s speech features, such as the display of fundamental frequencies, average facial expressions, accuracy of pronunciation, and waveform changes. This discovery has sparked a deeper exploration of the relationship between speech features and emotions. Subsequently, scholars in this field have held multiple conferences, journals, and events to discuss in-depth topics related to speech emotion computing and other emotion computing. During this period, universities in many countries around the world actively engaged in research on speech emotion recognition [7]. Among them, some influential research institutions include: the Emotion Speech Research Group established by Professor R. Cowie in 2000 at Queen’s University Belfast in the UK, which focuses on the study of psychology and speech analysis; The research team at the University of Southern California, founded by Professor Narayanan, focuses on acoustic analysis, recognition, and synthesis of emotions through laughter; At the Massachusetts Institute of Technology in the United States, a large institution led by Professor Picard focuses on research on emotions and emotion computation; Emotional Robotics Research Group at the Free University of Brussels; And the Emotion Research Laboratory at the University of Geneva in Switzerland. In the research of Speech Emotion Recognition (SER), extracting speech emotion features is a key step, which is the core component of pattern recognition. Its main purpose is to extract key information of emotion representation from speech data. Usually, acoustic features are divided into two categories: low-level features and deep features. Low level features are mainly obtained through time-domain and frequency-domain algorithms, including three types of features: sound quality features, spectral features, and prosodic features; Deep features are extracted using deep learning techniques, further learning high-level features from the original speech signal or extracted low-level features. Early research on speech emotion recognition focused on how to extract acoustic features that can best reflect emotions, as the quality of feature extraction directly affects the accuracy of recognition [8]. In early research, researchers relied on artificially designed sets of various acoustic features, mainly including rhythm, timbre, and sound quality [9]. However, this method of evaluating speech features from a human perspective has certain limitations. When computers use these artificially designed features for emotion recognition through deep learning models, if these acoustic features cannot fully understand the subtle differences in speech, the cognitive gap between the machine and the human brain may gradually widen, leading to the failure to achieve the expected results. Currently, many researchers have converted raw speech into spectrograms and inputted them into various recognition models, achieving good results. However, in order to be applied in the commercial field, speech emotion recognition technology still needs to be further improved and perfected.

Compared with a single feature, feature fusion shows superior and stable performance on multilingual datasets and various classifiers. Feature fusion encompasses the integration of traditional features, such as Rao et al. [10], which achieves higher accuracy than a single feature by fusing prosodic features that contain both local and global information. Fusion features also include the combination of traditional acoustic features and deep features. For example, Wang

et al. [11] fused deep features extracted by deep neural networks with manual features, significantly improving the overall performance of emotion recognition. Sun et al. [12] achieved fusion between shallow and deep features extracted by convolutional neural networks at different levels, and achieved good recognition results on most publicly available datasets.

In order to explore the general applicability of ensemble classifiers in speech emotion recognition tasks, Zehra et al. [13] developed a multi classifier ensemble method based on majority voting mechanism and applied it to the field of cross corpus multilingual speech emotion recognition. The Urdu corpus was used as the training and testing sets. Through comparison with existing literature, it was found that this method can improve recognition rate by up to 15%. This result indicates that using ensemble classifiers for cross corpus speech emotion recognition is a relatively effective strategy. This research work fully validates the advantages of integrating multiple classifiers and provides a new solution for speech emotion recognition tasks. By appropriately selecting and combining different types of classifiers, it is expected to further improve the system's generalization ability, enhance its applicability and robustness in multiple languages and contexts.

The latest research work focuses more on improving and integrating these deep learning models in order to further enhance the recognition ability of the system. It is worth mentioning that convolutional neural networks perform well in speech signal processing due to their sensitivity to local features; Recurrent neural networks are adept at capturing temporal information and have a natural advantage in processing serialized data. By improving the existing network structure or integrating multiple models, researchers are constantly exploring more efficient and robust speech emotion recognition methods, in order to achieve high applicability and generalization ability under multilingual and multi scenario conditions. Overall, deep learning models have brought new opportunities and development space for speech emotion recognition tasks, and related research work is steadily advancing.

As early as 2014, Mao et al. [14] proposed using CNN to learn significant emotional features in speech emotion recognition. Experimental results showed that this method outperformed existing speech emotion recognition feature extraction methods in complex scenarios. DiasIssa et al. [15] designed a novel speech emotion recognition framework that extracts various feature parameters such as MFCC, chromaticity diagram, Mel scale spectrogram, Tonnetz, and spectral contrast from speech files, and inputs them into a one- dimensional convolutional neural network for training. In experimental validation on public datasets IEMOCAP, EMODB, and RAVDESS, this method achieved accuracies of 64.3%, 86.1%, and 71.61%, respectively, demonstrating the excellent performance of convolutional neural networks in speech emotion recognition tasks. Overall, CNN based speech emotion recognition models can efficiently learn emotion related feature representations from raw speech signals, avoiding the limitations of manually designed features and laying a solid foundation for the sustainable development of this field. Researchers are constantly exploring and improving the structure of CNN models in order to achieve more accurate and robust emotion recognition in more complex speech scenarios.

To address this challenge, Long Short-Term Memory (LSTM) networks have emerged. Wollmer et al. [16] first introduced LSTM into the field of speech emotion recognition. In 2010, they proposed a multimodal emotion recognition framework based on a Bidirectional LSTM (Bi-LSTM), which fused speech and video data at the feature level. Experimental results

showed that the performance of this model was superior to the popular HMM and SVM models at that time. Hsu et al. [17] used Support Vector Machines (SVM) to recognize speech and non semantic sounds. After separating these two types of sounds through a prosodic feature extractor, deep residual networks were used to extract features separately and input them into an LSTM model with attention mechanism for classification. The recognition accuracy of this method exceeds that of traditional fusion techniques. Recently, Andayani et al. [18] proposed a hybrid LSTM network and Transformer model aimed at learning long-term temporal dependencies in speech signals and classifying emotions to address the challenge of strong temporal dependence of speech emotions.

1.2. **Motivation and contribution.**

This article mainly studies the problems of low accuracy and insufficient generalization ability of current speech emotion recognition models, and proposes a speech emotion recognition method architecture based on Time-frequency Feature Dual Fusion (TF-DF). Firstly, parallel convolutional neural networks are used to extract time- frequency feature representations from Mel Frequency Cepstral Coefficients (MFCC). However, not all information in the time-frequency features contributes to the emotion recognition task, and some features may interfere with the model's judgment. To this end, a channel attention mechanism is adopted to assign different weights to each channel, automatically focusing on features containing emotional knowledge and ignoring irrelevant features. On the other hand, there is complementarity between temporal and frequency features, and fully utilizing their information can help improve feature representation ability and avoid feature redundancy. Therefore, a feature dual fusion module was designed to capture complementary features in both time and frequency dimensions, and to achieve feature simplification and optimization through feature fusion, providing high-quality feature representation for subsequent emotion recognition tasks. The model achieved competitive performance on the IEMOCAP and RAVDESS datasets. Afterwards, the model will be used in a digital museum system to enhance the museum's service experience by accurately recognizing tourists' voice emotions.

2. **Analysis of relevant principles.**

2.1. **Emotional description model.**

The former divides emotions into several discrete categories, such as "happiness", "sadness", etc., while the latter represents emotions as continuous values of multiple dimensions, such as scores on dimensions such as "pleasure unpleasantness" and "proactivity passivity" [19]. Different emotion description models can characterize and represent emotional states from different perspectives, providing a theoretical basis for the design and construction of speech emotion recognition systems. Choosing an appropriate emotion description model is crucial for improving the accuracy and interpretability of speech emotion recognition.

In the research of emotion recognition, the discrete emotion description model is a commonly used method that classifies emotional states into multiple discrete categories, each labeled with a specific adjective. In the history of emotional cognition research, over 300 different emotional labels have been identified, which poses significant challenges to the study. To simplify this classification process, psychologist Ekman first proposed six basic emotional labels: happiness, sadness, disgust, surprise, anger, and fear [20]. These six basic emotions are considered the most fundamental and universal emotional experiences of humans, with cross- cultural consistency. Ekman's theory laid the foundation for subsequent research on speech emotion recognition, and currently the vast majority of emotion corpora use his proposed six basic emotion labels for data annotation. Although the six basic emotional labels have some representativeness, they cannot fully cover the rich emotional experiences of humans. Some researchers believe that more refined emotional categories should be considered to more accurately depict the diversity of emotions. For example, in addition to basic emotions, some complex emotional categories such as shame, jealousy, pride, etc. can also be added. At the same time, differences in emotional intensity should also be taken into account, as the same emotion can produce different experiences and expressions under different intensities.

The dimensional emotion description model adopts different ways to represent emotional states and maps them to points in multidimensional space. Each dimension represents an emotional attribute or feature, and combining multiple dimensions can depict complex emotional experiences. This method aims to describe emotions in a more continuous and detailed way, avoiding simplification in discrete label models. Common dimensional emotional description models include the two-dimensional Arousal Valence spatial model and the three- dimensional Motivation Evaluation Dominance spatial model. The valence dimension describes the positive and negative values of emotions, that is, the subjective evaluation of an individual's emotional experience, ranging from a positive valence of extreme pleasure to a negative valence of extreme pain; The arousal dimension measures an individual's physiological arousal or psychological activity level, measuring the arousal or relaxation state accompanied by an emotion. Low arousal state is usually manifested as fatigue and lethargy, while high arousal state is manifested as tension or excitement.

2.2. **Speech signal preprocessing.**

Preprocessing is the primary and crucial step in the process of speech emotion recognition. Preprocessing usually includes steps such as pre-emphasis, frame segmentation, and windowing. Proper preprocessing of the original speech signal is crucial for subsequent feature extraction and emotion recognition.

2.2.1. *pre-emphasis.* Pre emphasis, as an important part of speech preprocessing, essentially involves processing the original speech signal through a high pass filter, aiming to flatten the speech spectrum. This step can compensate for the energy attenuation caused in the high frequency range, thereby highlighting the performance of resonance peaks in the high frequency region. Pre emphasis is implemented by a high pass filter, and its mathematical expression is:

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

where μ is 0.97. After pre emphasis processing, the spectral distribution of the speech signal is more balanced compared to the original signal.

2.2.2. *Frame segmentation and windowing.* After pre emphasizing the speech signal, in order to perform Fourier transform, it is necessary to perform frame segmentation and windowing on the signal. Due to the variability of speech signals over larger time scales and their short-term stability over smaller time scales, speech signals are typically divided into short time periods of 10 to 30 milliseconds to analyze their short-term stability during that time period. However, traditional signal processing techniques often require the input signal to be continuous, while through frame processing, the speech signal is segmented into several short-term segments. To achieve smooth transitions between frames, it is necessary to leave overlapping areas between adjacent frames, which is called "frame shift".

Due to the segmentation of continuous speech signals into a series of short time periods by framing operations, this processing method can result in discontinuous speech signals between frames. These window functions are designed so that their edge values are close to zero, so when multiplied with the speech signal, the boundary values of the speech signal will also be close to zero, which helps to improve the continuity of the speech signal. The impact of different window functions on speech signals varies, and it is crucial to choose a reasonable window function to preserve the true features of the speech signal. Therefore, this affects the efficiency of subsequent feature extraction. Through extensive theoretical analysis and experimental comparison, this study used Hamming window for windowing treatment, and its mathematical expression is shown below:

$$w(n) = \begin{cases} 0.54 - 0.64\cos\left(\dfrac{2\pi n}{N-1}\right), & 0 \le n \le N-1, \\ 0, & \text{others}, \end{cases} \tag{2}$$

where N represents the frame rate. Hanming window not only effectively suppresses spectral leakage, but also reduces boundary distortion and preserves the main features of speech signals. Compared to other window functions, Hamming windows exhibit superior performance in terms of the ratio of main and side latent values to the minimum and maximum values.

2.3. **Speech emotion recognition model based on deep learning.** Since the concept of deep learning was first proposed, it has received widespread academic attention and has been applied in multiple research fields. Especially in the field of speech emotion recognition, deep learning network models have demonstrated excellent performance and have achieved a series of significant research results. This section will provide a detailed introduction to three commonly used deep learning network models in SER.

2.3.1. *Deep neural network model.* When it comes to the concept of deep learning and its widespread attention in academia and application fields, Hinton and Salakhutdinov's [?] research work in 2006 was a milestone. This literature explains two core points: (1) deep neural networks are optimized through a layered initialization strategy, which helps to seek the best solution during the training process; (2) Neural networks with multiple hidden layers can more effectively extract the representation features of raw data, which is very beneficial for solving visual and classification problems. Deep learning networks are composed of a large number of neurons that are interconnected and constantly adjust their connection weights during training, forming a complex network structure. In the field of deep learning, this complex structure is commonly referred to as Deep Neural Networks (DNNs).

In the architecture of DNN, each level is composed of multiple neurons that transmit features through fully connected connections to facilitate information exchange between layers. The input received by each neuron is the result of a linear transformation of the output of its previous layer of neurons, which must be nonlinearly transformed through an activation function before being passed on to the next step. There are three main types of common activation functions, which will be introduced in detail below.

For the Sigmoid function, as the input value decreases, its output value gradually approaches 0, and its rate of change also tends to be gentle; on the contrary, as the input value gradually increases, its output value approaches 1, and its change also appears slow. This function is typically used in the output layer, with output values ranging from $[0, 1]$, and is commonly used to perform binary classification tasks:

$$\text{Sigmoid}(x) = \frac{1}{1 + e^{-x}}. \tag{3}$$

It has been widely used in various deep learning networks, known as the rectified linear unit:

$$\text{ReLU}(x) = \max(0, x). \tag{4}$$

The Tanh function, commonly used in hidden layers, has the main advantage that its output range is between $[-1, 1]$, which makes the average value of the data 0 and does not cause saturation problems. Although the curve shape of the Tanh function is similar to that of the Sigmoid function, Tanh changes faster near its limit, thereby improving computational efficiency:

$$\text{Tanh}(x) = \frac{1 - e^{-x}}{1 + e^{-x}}. \tag{5}$$

Lee et al. used DNN as sentiment classifiers and further compared the performance of DNN with the results obtained using SVM as sentiment classifiers. The comparison results show that DNN outperforms SVM in classification accuracy, which verifies that applying DNN on manual feature sets can effectively improve the accuracy of emotion recognition. MFCC features were extracted from the original language and DNN was applied for emotion recognition. The research results show that the recognition accuracy of seven emotions is over 95%. One key advantage of DNN is that it can enhance the model's expressive power by increasing the number of network layers. However, this method consumes more computing resources and a large amount of training data, resulting in longer training time and increased computational costs.

2.3.2. *Convolutional neural network model.* CNN is a typical representative of feedforward neural networks. The Convolution Layer is mainly responsible for feature extraction and is a crucial component in CNN. It utilizes local connections to reduce the required computational load, typically located near the input layer of the network. By using fixed size convolution kernels to process feature maps, multiple different feature representations can be obtained. The different levels of convolutional layers can capture semantic features at different levels; lower level features are usually more vague and simple, while higher-level features contain more semantic information. The size and stride of the convolution kernel can be adjusted according to different application requirements. Although larger convolution kernels help extract higher-level features, they may lose some detail information in the process; smaller convolution kernels can capture more detailed features, but doing so requires more computational resources. The mathematical expression for convolution operation is shown below:

$$Z_{i,j,k} = \sum_{p=0}^{P-1} \sum_{q=0}^{Q-1} \sum_{c=0}^{C-1} X_{i+p,\, j+q,\, c}\, W_{p,q,c,k} + b_k, \tag{6}$$

where K, P, Q, and C represent the number, height, width, and number of channels of the convolutional kernels; $X_{i+p,\, j+q,\, c}$ refers to the value of the c-th channel of X at $(i{+}p, j{+}q)$; b_k represents the paranoid parameter of the k-th convolution kernel; $W_{p,q,c,k}$ represents the weight parameters of the k-th convolution kernel at (p, q) and the c-th channel; and $Z_{i,j,k}$ represents the reaction value of the k-th convolution kernel at the i-th and j-th positions of the input data X.

The pooling layer can also improve the model's fault tolerance to a certain extent and help reduce the risk of overfitting. This layer is also commonly referred to as the downsampling layer. The commonly used pooling methods include Average Pooling and Max Pooling, as expressed mathematically below:

$$Y_{i,j,k} = \max\bigl(Z_{s\cdot i+p,\, s\cdot j+q,\, k}\bigr), \qquad 0 \le p \le P,\ 0 \le q \le Q, \tag{7}$$

where $Y_{i,j,k}$ represents the (i, j, k)-th pixel in the feature map after pooling processing; the maximum pooling operation is represented by max, s represents the pooling step size, and Z represents the input raw feature map.

The fully connected layer (FC Layer) is structurally similar to multi class neural networks and typically uses Sigmoid or Softmax functions as its activation function. After the data is processed by the primary layer of the network and high-level features are extracted, these features will be passed on to the fully connected layer for further processing. However, if there are too many neurons in the fully connected layer, it may lead to overfitting issues. To alleviate this issue, Dropout technique can be used, which reduces the number of parameters in the network by randomly discarding a portion of neurons, thereby helping to prevent overfitting.

2.3.3. *Recurrent Neural Network Model.* The original design intention of Recurrent Neural Network (RNN) is to capture temporal information, which is usually applied to solve time related problems. In this network model, the implementation of memory function is mainly due to its ability to utilize information from previous states in the current input and output processes. By establishing time series connections between RNN neurons, this structure endows the network with memory capabilities when processing sequential data.

In the RNN model, each neuron receives the current input and the output of the previous time step as new inputs, and passes the results to the next time step, specifically:

$$h_t = F_w(h_{t-1}, x_t)\,, \tag{8}$$

where h_t and h_{t-1} refer to the states at time t and time $t{-}1$, respectively, and x_t represents the input at time t; F_w is the tanh function as a recursive function, which is an important mathematical tool. The details are further explained as follows:

$$h_t = \tanh(W_h h_{t-1} + W_x x_t)\,, \tag{9}$$

$$y_t = W_y h_t, \tag{10}$$

where W_h and W_x are the weights of the previous state and input, respectively; y_t is the output.

Scholars have proposed an efficient emotion recognition technique based on RNN, which utilizes advanced learning algorithms to train RNN and can process contextual information from a distance, effectively mapping all frames in the same speech to consistent emotion labels. However, a major challenge faced by deep neural networks is the instability of gradients, which can lead to exponential growth or decay of gradients, known as gradient explosion or disappearance problems. In RNN, this problem is particularly severe because gradients need to be calculated in multiple time steps and layers, making the problem of gradient vanishing more significant.

LSTM overcomes the gradient vanishing problem by introducing additional interaction mechanisms in the network. LSTM networks can selectively retain or forget information in their cellular states, maintaining long-term dependencies by establishing connections between the past and present, which enables them to exhibit good memory capabilities when processing long sequence data. However, due to the complex network structure of LSTM, which involves multiple gating mechanisms and parameters, it presents certain challenges in the training and parameter adjustment process.

The components of LSTM are shown as follow:

$$f_t = \sigma(W_f h_{t-1} + W_f x_t)\,, \tag{11}$$

$$i_t = \sigma(W_i h_{t-1} + W_i x_t)\,, \tag{12}$$

$$o_t = \sigma(W_o h_{t-1} + W_o x_t)\,, \tag{13}$$

$$\tilde{C}_t = \tanh(W_c h_{t-1} + W_c x_t)\,, \tag{14}$$

$$C_t = i_t \,\varepsilon\, \tilde{C}_t + f_t \,\varepsilon\, C_{t-1}, \tag{15}$$

$$h_t = o_t \,\varepsilon\, \tanh(C_t), \tag{16}$$

where i_t, o_t, and f_t respectively represent the input gate, output gate, and forget gate; σ represents the sigmoid activation function, weight sets W_i, W_o, W_f represent the weights of the input gate, output gate, and forget gate, W_c is the weight of the cell state, and h_t is the new state.

3. **Speech emotion recognition method based on dual fusion of time-frequency features.** In the human-computer interaction process of digital museums, noise in speech signals still affects the effectiveness of model training. Considering that speech signals exhibit significant performance in both time and frequency domains, these two features are naturally suitable for emotion recognition. Although single time-domain or frequency-domain features can achieve emotion recognition, previous studies have shown that they can achieve good accuracy. To this end, this article uses parallel CNN to extract basic features from MFCC, and enhances key information while suppressing irrelevant information through a weighting mechanism, in order to achieve comprehensive analysis of time-frequency features. In addition, this article also introduces a dual fusion module that can capture complementary information between time and frequency features and reduce information redundancy, thereby improving the recognition performance of the model.

3.1. Time frequency feature extraction.

3.1.1. *Low level time-frequency feature extraction.* This article selects MFCC (Mel Frequency Cepstral Coefficients) containing both time-domain and frequency-domain information as the input data for the model. It has been widely confirmed that CNN performs well in feature extraction in the field of image processing, and this parallel network structure can effectively integrate time-domain and frequency-domain information.

The specific parameters of convolutional filters are described in the format of [filter number, height × width]. Firstly, two specially designed convolutional filters are introduced to extract low-level time-domain and frequency-domain features, respectively. The MFCC features are processed through a [16, 1×3] time-domain convolution filter to obtain a low-level time-domain feature matrix; after passing through a frequency domain convolution filter of [16, 3×1], the low-level frequency domain feature matrix is obtained. Taking the MFCC features extracted from the IEMOCAP dataset as an example, their original sizes were $(1, 26, 57)$. After parallel CNN extraction, the size of the low-level time-domain and frequency-domain feature matrices became $(16, 26, 57)$.

3.1.2. *Time frequency feature weighting.* CNN utilizes convolutional layers to represent spatial connections by combining receptive fields and feature information. This approach not only captures the spatial properties of features, but also involves the interactions between channels. However, most existing models mainly focus on expanding the receptive field to enhance the expressiveness of spatial features, often ignoring the intrinsic connections between channels. To compensate for this deficiency, it is recommended to conduct in-depth modeling of the interrelationships between each channel. By exploring the logical connections within the channel, it helps the model to more effectively characterize key features and highlight important information of the image in the relevant field. Introducing attention mechanism can promote the model's understanding of the logical relationships between channels during training, thereby achieving the capture of more comprehensive feature information and enhancing the overall performance of the model. Introducing attention mechanisms into network models can better identify potential logical connections between different channels, enabling the model to capture complete feature information more effectively during training. In view of this, this chapter utilizes a feature weighting module (DW) with specific dimensions to enhance relevant information and reduce the influence of irrelevant information. It introduces the plug and play channel attention method SE Net and improves it to weight features from all dimensions simultaneously.

SE Net module is an innovative architecture focused on establishing potential connections between channels. Unlike traditional network modules, SE Net emphasizes the correlation within channels and adopts a novel channel compression and activation mechanism designed to enhance the model's understanding and processing capabilities of channel relationships. The core of this strategy is to learn the potential logical relationships between channels by recalibrating the importance of each channel, rather than redefining feature channels through network and spatial dimension transformations. Based on these learned connections, re evaluate the importance of each channel, focus on key information and ignore secondary information.

The main goal of SE Net is to explore potential correlations between channels and use feature maps as the basis for all operations. With its simple structure, low complexity, small memory footprint, and ease of use, it has significant advantages compared to other modules of the same period and can be easily integrated as a plugin module into various network architectures. This module focuses on analyzing and calculating the dependencies and their importance between channels, in order to improve the overall performance of the network and enhance sensitivity to these dependencies. This module evaluates and ranks the importance of the current channel, and uses weight values to set the priority of each channel in the next stage. In addition, it learns the interaction information between channels, so as to adaptively adjust the weights of each channel, enhance the attention to key information channels, and reduce the attention to non key information channels. This design makes the SE Net module more efficient and accurate in handling logical associations at the channel level.

3.2. **Model Architecture.**

3.2.1. *Dual feature fusion module.* In order to effectively integrate complementary information between time and frequency features while eliminating feature redundancy, we propose a feature dual fusion module. This module consists of two parts: time-frequency adaptive fusion block and global fusion block. The time-frequency adaptive fusion block aims to adaptively capture and fuse complementary information between time-domain and frequency-domain features, dynamically adjusting the weights of the two features through attention mechanisms. The global fusion block aims to further integrate information between different channels and enhance the global representation ability of features. The design concept of the dual feature fusion module is to first capture the differences and complementarity between time-domain and frequency-domain features in the channel dimension, and then fuse the information between different channels in the spatial dimension, ultimately obtaining richer and more discriminative feature representations.
Time Frequency Adaptive Fusion. Using matrix operations and connection methods similar to attention mechanisms, the extracted time-frequency features are adaptively fused. Firstly, connect the time and frequency features for simple fusion:

$$\hat{X}_{TF} = \mathrm{concat}(X'_T,\, X'_F) \tag{17}$$

Then, a convolutional layer with a kernel size of 1 × 1 is applied to capture cross channel interactions and generate two weight maps to reallocate time and spectral features:

$$\alpha_T, \alpha_F = \mathrm{split}\Big(\mathrm{softmax}\big(\mathrm{Conv2D}(\hat{X}_{TF})\big)\Big) \tag{18}$$

The segmentation operation divides the feature map along the channel dimension, generating two weight maps α_T and α_F, which represent the spatial importance corresponding to the time and frequency features and are shared among all channels. Finally, the extracted feature maps are weighted and fused based on the weighted map, so that the depth time features and depth frequency features are adaptively fused in space. Then apply 2 × 2 max pooling to obtain time-frequency fusion feature X_{TF}:

$$X_{TF} = \mathrm{maxpool}(\alpha_T\, \varepsilon\, X'_T \;+\; \alpha_F\, \varepsilon\, X'_F) \tag{19}$$

Global Fusion. After adaptive weighted fusion of time-frequency features, the model also applies a global fusion strategy to enhance communication between cross dimensional features. The output of the first fully connected layer can be represented as:

$$Z = \sigma(X_{TF}\, U) \tag{20}$$

where $\sigma(\cdot)$ represents the activation function, GeLU is used as the activation function in this scheme. Z represents the output of the first fully connected layer, where $d = \lfloor H/2 \rfloor \lfloor W/2 \rfloor$ and U represent the network parameters of the first fully connected layer. The operation of the gating unit can be represented as:

$$\tilde{Z} = g(Z) \tag{21}$$

where $\tilde{Z}$ represents the output of the gating unit, and $g(\cdot)$ represents the operation of the gating unit. The gating unit is a key design of the gMLP block, which enables interaction of features from multiple angular dimensions. The operation of fully connected second layer can be expressed as follows:

$$Y = \tilde{Z} V \tag{22}$$

where V represents the network parameters of the second fully connected layer, and Y represents the output of the second fully connected layer. The second fully connected layer maps the feature dimension from $2d$ back to d. It is worth noting that when $g(\cdot)$ is an identity mapping, the gMLP block will degenerate into a conventional feedforward network. Introducing residual connections in each gMLP block helps with convergence during network training. In order to achieve interaction between multiple angular features, the gating unit needs to have a contraction operation in the angular dimension, and a relatively simple implementation is to use linear mapping. The linear mapping of features in the angular dimension can be expressed as:

$$f_{w,b}(Z) = WZ + b \tag{23}$$

where W is a parameter that can be trained in the gating unit, and each parameter represents the correlation between two angular features; b is the bias for each angle feature and is also a trainable parameter. The operation of a linear gating unit can be expressed as a function:

$$g(Z) = Z \odot f_{w,b}(Z) \tag{24}$$

where ε represents the multiplication operation at the element level. To reduce computational complexity, we divide Z into two parts (Z_1, Z_2) along the delay dimension. The output of the final gate control unit can be expressed as:

$$g(Z) = Z_1 \odot f_{w,b}(Z_2) \tag{25}$$

3.2.2. *Overall model.* This model uses MFCC as input features and utilizes convolutional neural networks to extract time-frequency feature representations from speech signals. However, not all information in time-frequency features contributes to emotion recognition tasks, and some features may have adverse effects on model judgment. To this end, we introduced a channel attention mechanism that assigns different weights to each channel, allowing the model to autonomously focus on features containing important emotional information while reducing the attention to irrelevant features. On the other hand, there is complementarity between temporal and frequency features, and fully utilizing their information can help improve feature representation ability and avoid feature redundancy. Therefore, this article proposes a dual feature fusion module aimed at capturing complementary features in both time and frequency dimensions, and achieving feature simplification and optimization through feature fusion, providing high-quality feature representation for subsequent emotion recognition tasks. As shown in Figure 1, this is a speech emotion recognition model framework based on TF-DF.

4. **Experiment.**

4.1. **Datasets.** In museum human-computer interaction, visitors' emotions mainly include calmness, happiness, sadness, anger, fear, disgust, surprise, and neutrality. The performance of speech emotion recognition largely depends on high-quality speech datasets. However, due to the subjectivity and abstraction of sentiment data, constructing a speech sentiment dataset containing accurate sentiment labels faces significant challenges. Individuals may have different perspectives and judgments regarding the emotional states expressed in the same speech sample. Due to the high cost and difficulty in ensuring data

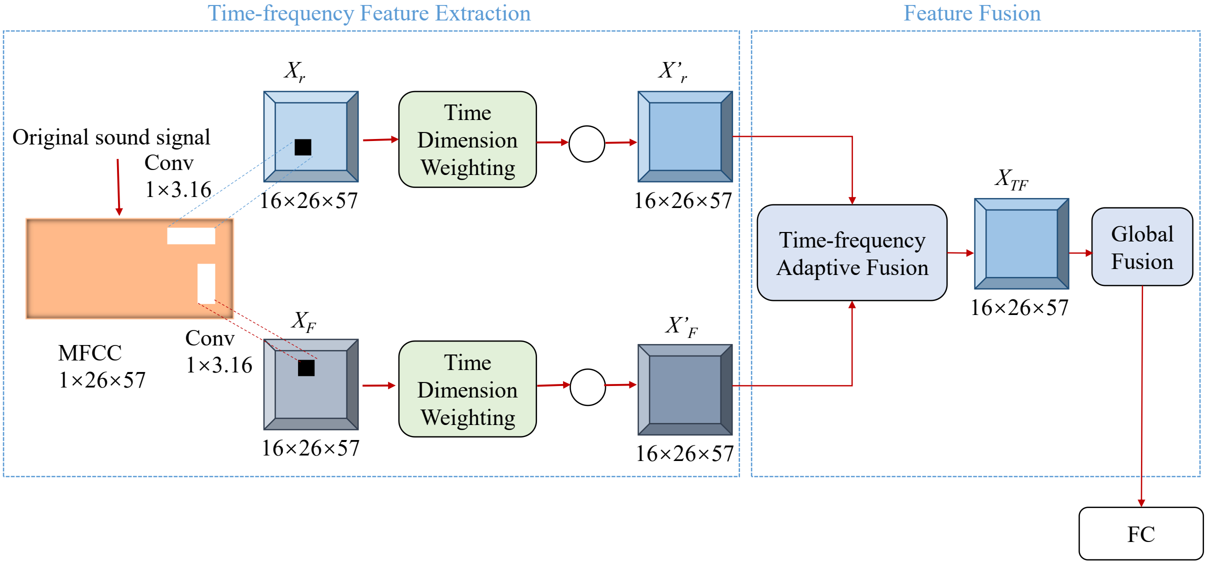


Figure 1. The accuracy of each fold of the TF-DF model on various datasets

quality of self built voice emotion databases, most researchers choose to use existing public datasets. Most widely used speech emotion databases adopt discrete emotion description models, covering the six basic emotions defined by Ekman, such as IEMOCAP [22], EMODB, and RAVDESS [23]; In addition, the VAM database adopts a dimensional sentiment description model, where sentiment labels are represented in numerical form ranging from -1 to 1. This study used two publicly available English datasets, IEMOCAP and RAVDESS.

4.2. **Evaluating indicator.** Speech emotion recognition can basically be regarded as a multi classification problem. In scenarios involving multiple classifications, commonly used evaluation criteria include precision, recall, and F1 score. However, when dealing with speech emotion recognition datasets using discrete emotion labels, uneven distribution of categories may result in using accuracy as the sole evaluation criterion being insufficient to reflect the comprehensiveness of the model. The reason for this situation is that models tend to lean towards categories with a larger sample size, while ignoring categories with fewer samples. Therefore, in order to more fairly evaluate the contribution of each category to the overall sample accuracy, this study used Weighted Accuracy (WA), Unweighted Accuracy (UA), and F1 score as performance evaluation metrics for the model. In the confusion matrix, **TP**: true positive, actual positive, predicted positive; **FN**: False negative, actually positive, predicted as negative; **FP**: False positive, actually negative, predicted as positive; **TN**: True negative, actual negative, predicted negative, P represents accuracy. The above three indicators are calculated based on the values of TP, FP, TN, and FN, and the specific formulas are as follows:

$$\mathrm{WA} = \frac{\sum_{i=1}^{N} \mathrm{TP}_i}{\sum_{i=1}^{N}(\mathrm{TP}_i + \mathrm{FP}_i)} \tag{26}$$

$$P = \frac{\mathrm{TP}_i}{\mathrm{TP}_i + \mathrm{FP}_i} \tag{27}$$

$$\mathrm{UA} = \frac{1}{N}\sum_{i=1}^{N} P_i \tag{28}$$

$$F1 = \frac{2 \times P \times R}{P + R} \tag{29}$$

4.3. **Experimental results and analysis.** In FCM type algorithms, a large number of experiments have shown that setting the membership weight index to 2 can achieve better results. Based on experience, the membership weight index of all algorithms in this section is set to 2, and the remaining parameters that need to be adjusted are selected using a grid search strategy. The ANFCM algorithm has three parameters that need to be adjusted. The first parameter is the regularization parameter α, which is used to adjust the degree of influence of global information and sample point neighbor information on clustering. The values are [0.001,0.01,0.1,1,10,100,1000,10000]; The second parameter is the count of nearest neighbor sample points k_x, and the third parameter is the count of nearest neighbor clustering prototypes k_v. The settings for both parameters are the same, [2, 3, 5, 7, 9, 15, 19, 25]. The regularization parameter λ of the AFKM algorithm is set to [0.001, 0.01, 0.1, 1, 2, 5, 10, 100].

This study used two datasets, IEMOCAP and RAVDESS, to test the model's generalization ability and robustness. The experiment used a 10 fold cross validation method, randomly dividing the data into a training set, a validation set, and a test set, with proportions of 80%, 10%, and 10%, respectively.

Figure 2 shows the accuracy performance of cross validation folds of TF-DF features on different datasets. Detailed analysis reveals that in the IEMOCAP dataset, the accuracy of the validation sets at the 4th, 6th, and 7th folds has exceeded the level of the final test set. Meanwhile, on the RAVDESS dataset, the accuracy of the eighth fold validation set reached over 95%, demonstrating excellent performance.

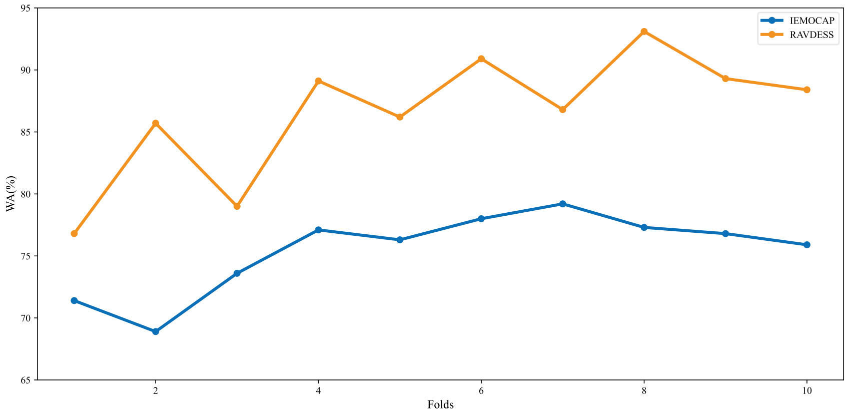
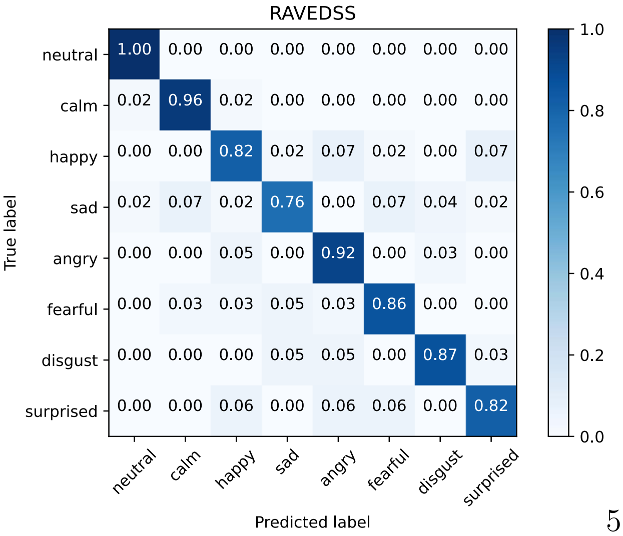


Figure 2. The accuracy of each fold of the TF-DF model on various datasets

In order to further analyze the performance of the TF-DF model, this study presents the sentiment recognition confusion matrix of the model on datasets IEMOCAP and RAVDESS. The relevant confusion matrices are detailed in Figure 3 and Figure 4.

Based on the identification confusion matrix analysis of the TF-DF model on the publicly available datasets IEMOCAP and RAVDESS, the model outperforms the model introduced in Chapter 3 in terms of emotion label recognition performance. Specifically, in the IEMOCAP dataset, the accuracy of identifying happy emotions is the lowest, which is consistent with most studies. This is mainly due to the fusion of excitement and happiness emotions, as well as the diversity of ways to express happiness emotions. There are significant differences in the behavior of different individuals when expressing happiness. Some people may appear extremely excited, while others may remain relatively calm, all of which increase the complexity of identifying happy emotions. Relatively speaking, the



5

Figure 3. Identification confusion matrix of TF-DF model on datasets IEMOCAP

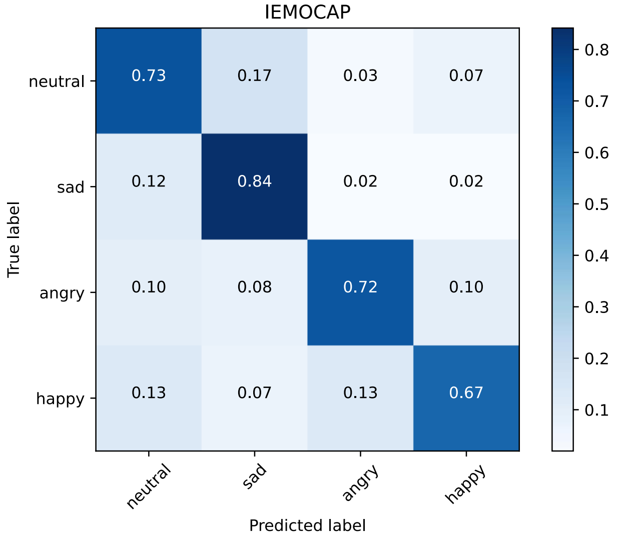


Figure 4. Identification confusion matrix of TF-DF model on datasets RAVDESS

recognition effect of sad emotions is relatively excellent, thanks to its consistent and obvious expression, which makes its recognition rate relatively high. In the evaluation of the TF-DE model, the recognition accuracy of the four core emotions: calm, sadness, anger, and happiness are 73%, 85%, 72%, and 67%, respectively. On the RAVDESS dataset, the recognition rates for neutral, calm, and angry all exceeded 90%. This indicates that the new model architecture introduced in this chapter has achieved significant improvements in overall accuracy. However, sensitivity to certain emotions still needs to be strengthened and further optimization and improvement are needed. Applying the model proposed in this article to the museum's human-computer interaction system can more accurately identify changes in tourists' emotions and provide them with a better gaming experience.

5. **Conclusions.** This article proposes a speech emotion recognition model framework based on TF-DF, which can improve the performance of museum human-computer interaction systems and enhance the visitor experience. Considering that there may be

irrelevant features in time-frequency features that do not contribute or interfere with emotion recognition, a channel attention mechanism is adopted to automatically assign different weights to each feature channel, automatically focusing on important features containing emotional knowledge and ignoring irrelevant features. On the other hand, there is complementarity between temporal and frequency features, and fully utilizing both types of feature information can help enhance feature representation capabilities and avoid redundant features. For this purpose, a dual feature fusion module was designed to capture complementary features in both time and frequency dimensions, and optimize features through time-frequency feature fusion to provide high- quality feature representation for subsequent emotion recognition tasks. The model achieved competitive performance on the IEMOCAP and RAVDESS datasets.

The experiment in this study was conducted under ideal conditions, and the dataset used did not contain noise interference or involve noisy samples in practical applications. In order to promote the application of speech emotion recognition technology in specific environments, future work will mainly focus on the problem of emotion recognition in noisy environments.

Acknowledgment. This work is supported by the Research Project on Innovative Development of Social Sciences in Anhui Province in 2023 (No. 2023KY504), the Anhui Province Higher Education Science Research Project (Philosophy and Social Sciences) (No. 2023AH051804), the Natural science research project of Anhui Xinhua University (No. 2024zr001), the Anhui Province Teaching Demonstration Course: Operating System Principles (No. 2020SJJXSFK1294), Anhui Xinhua University Internal Training Doctoral Research Start-up Fund (No. bs2025kyqd086), Anhui Province Quality Engineering Project :Exhibition Design (2024aijy453), and the Anhui Province Quality Engineering Project: Operating System Principles (No. 2019kfkc158).

REFERENCES

[1] P. F. Marty, “Museum websites and museum visitors: digital museum resources and their use,” Museum Management and Curatorship, vol. 23, no. 1, pp. 81-99, 2008.

[2] G. Varvin, H. Fauskerud, I. Klingvall, L. Stafne-Pfisterer, I. S. Hansen, and M. R. Johansen, “The journey as concept for digital museum design,” Digital Creativity, vol. 25, no. 3, pp. 275-282, 2014.

[3] F. Zhang, T.-Y. Wu, J.-S. Pan, G. Ding, and Z. Li, “Human motion recognition based on SVM in VR art media interaction environment,” Human-centric Computing and Information Sciences, vol. 9, 40, 2019.

[4] T. M. Wani, T. S. Gunawan, S. A. A. Qadri, M. Kartiwi, and E. Ambikairajah, “A comprehensive review of speech emotion recognition systems,” IEEE Access, vol. 9, pp. 47795-47814, 2021.

[5] M. Swain, A. Routray, and P. Kabisatpathy, “Databases, features and classifiers for speech emotion recognition: a review,” International Journal of Speech Technology, vol. 21, pp. 93-120, 2018.

[6] C. E. Williams, and K. N. Stevens, “Emotions and speech: Some acoustical correlates,” The Journal of The Acoustical Society of America, vol. 52, no. 4B, pp. 1238-1250, 1972.

[7] P. Asha, L. Natrayan, B. Geetha, J. R. Beulah, R. Sumathy, G. Varalakshmi, and S. Neelakandan, “IoT enabled environmental toxicology for air pollution monitoring using AI techniques,” Environmental Research, vol. 205, 112574, 2022.

[8] Z. Yao, Z. Wang, W. Liu, Y. Liu, and J. Pan, “Speech emotion recognition using fusion of three multi-task learning-based classifiers: HSF-DNN, MS-CNN and LLD-RNN,” Speech Communication, vol. 120, pp. 11-19, 2020.

[9] W. Fan, X. Xu, B. Cai, and X. Xing, “Isnet: Individual standardization network for speech emotion recognition,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 30, pp. 1803-1814, 2022.

[10] K. S. Rao, S. G. Koolagudi, K. S. Rao, and S. G. Koolagudi, “Robust emotion recognition using sentence, word and syllable level prosodic features,” Robust Emotion Recognition using Spectral and Prosodic Features, pp. 47-69, 2013.

[11] C. Wang, Y. Ren, N. Zhang, F. Cui, and S. Luo, "Speech emotion recognition based on multi-feature and multi-lingual fusion," Multimedia Tools and Applications, vol. 81, no. 4, pp. 4897-4907, 2022.

[12] L. Sun, J. Chen, K. Xie, and T. Gu, "Deep and shallow features fusion based on deep convolutional neural network for speech emotion recognition," International Journal of Speech Technology, vol. 21, pp. 931-940, 2018.

[13] W. Zehra, A. R. Javed, Z. Jalil, H. U. Khan, and T. R. Gadekallu, "Cross corpus multi- lingual speech emotion recognition using ensemble learning," Complex & Intelligent Systems, vol. 7, no. 4, pp. 1845-1854, 2021.

[14] Q. Mao, M. Dong, Z. Huang, and Y. Zhan, "Learning salient features for speech emotion recognition using convolutional neural networks," IEEE Transactions on Multimedia, vol. 16, no. 8, pp. 2203-2213, 2014.

[15] D. Issa, M. F. Demirci, and A. Yazici, "Speech emotion recognition with deep convolutional neural networks," Biomedical Signal Processing and Control, vol. 59, 101894, 2020.

[16] M. Wöllmer, F. Eyben, A. Graves, B. Schuller, and G. Rigoll, "Bidirectional LSTM networks for context-sensitive keyword detection in a cognitive virtual agent framework," Cognitive Computation, vol. 2, pp. 180-190, 2010.

[17] J.-H. Hsu, M.-H. Su, C.-H. Wu, and Y.-H. Chen, "Speech emotion recognition considering nonverbal vocalization in affective conversations," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 29, pp. 1675-1686, 2021.

[18] F. Andayani, L. B. Theng, M. T. Tsun, and C. Chua, "Hybrid LSTM-transformer model for emotion recognition from speech audio files," IEEE Access, vol. 10, pp. 36018-36027, 2022.

[19] S. PS, and G. Mahalakshmi, "Emotion models: a review," International Journal of Control Theory and Applications, vol. 10, no. 8, pp. 651-657, 2017.

[20] P. Ekman, W. V. Friesen, M. O'sullivan, A. Chan, I. Diacoyanni-Tarlatzis, K. Heider, R. Krause, W. A. LeCompte, T. Pitcairn, and P. E. Ricci-Bitti, "Universals and cultural differences in the judgments of facial expressions of emotion," Journal of Personality and Social Psychology, vol. 53, no. 4, 712, 1987.

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

[22] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," Language Resources and Evaluation, vol. 42, pp. 335-359, 2008.

[23] S. R. Livingstone, and F. A. Russo, "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English," PloS One, vol. 13, no. 5, e0196391, 2018.