ingegneria sismica
International Journal of Earthquake Engineering
Anno XLIII - Num. 2 - 2026

Speech Emotion Recognition and Service Quality Association Analysis for Cabin Service Conversation Scenarios

Jianhui Wang[1,*]

[1] Culture & Tourism School, Shaanxi Vocational and Technical College, Xi'an, Shaanxi, 710038, China

SUMMARY: *In this paper, the speech sensor is first used to obtain the original speech emotion signal of the cabin service dialog scene, and the speech collected by the speech sensor is preprocessed, and its preprocessing includes pre-emphasis, frame-splitting, windowing, and end-point detection, which can effectively eliminate the noise in the original speech signal. Subsequently, with reference to several common speech emotion features in the cabin service dialog scene, Mel's acoustic spectrogram and Mel's cepstrum coefficient are finally adopted to carry out the speech emotion feature extraction. Aiming at the unsatisfactory performance of RepVGG network, CBAM is adopted to improve the RepVGG network, and finally the speech emotion recognition model based on the improved RepVGG network is obtained. By constructing a service quality detection platform form, the speech emotion recognition technology is directly applied to the service process of the cabin service dialogue scene, in order to realize the correlation between speech emotion recognition and service quality of the cabin service dialogue scene. The speech emotion signal processing capability, speech emotion signal recognition capability, and speech emotion signal classification capability have moderate positive correlation with the platform service quality, and their Pearson's coefficient values are 0.672, 0.467, and 0.487, which sufficiently reveal the correlation between the speech emotion recognition and the service quality of the cabin service dialog scene.*

KEYWORDS: *repVGG network; CBAM; speech emotion recognition model; cabin service dialog scenarios; service quality*

1 Introduction

Cabin service mainly refers to the effectiveness and quality of the services and safeguards provided by the airline for passengers in the process of passengers leaving the aircraft after boarding to the aircraft landing, which mainly includes two service contents of technical quality class and functional class [1, 2]. Technical quality services are mainly cabin pilots, security officers, flight attendants on the relevant business processes and service content of proficiency and application of flexibility [3, 4]; functional services mainly include cabin seats and other infrastructure, cabin food and beverage [5], voice entertainment equipment, etc., including hardware facilities, emphasizing the degree of perfection of the configuration of these hardware facilities and the degree of humanization and so on. As the core content of travel services provided by airlines for passengers, cabin service is the core competitiveness of airlines, and its service quality directly determines the satisfaction of passengers with airlines [6-8].

The traditional cabin service quality assessment is mainly based on questionnaire surveys

*wangjh.110023@gmail.com
https://doi.org/10.65102/is2026529
Published: 30 April, 2026

or user complaints, which is inefficient, ineffective and highly subjective. In recent years, with the development and application of artificial intelligence, speech emotion recognition has gradually developed into a key technology for emotion recognition in customer service context [9, 10]. As passengers come from different countries and regions, emotional expression and understanding become one of the biggest obstacles in cabin service conversation scenarios [11, 12], happy, sad, angry, calm, surprised, disgusted, fear, these basic emotions leave different traces in their voices, and in order to avoid unnecessary trouble that may occur between passengers and crew members, speech emotion recognition is very necessary. Speech emotion recognition is based on machine learning and deep learning methods, which are able to train models with a large amount of speech data to extract features of speech from cabin service dialog scenes [13, 14]. These features can include voice tone, pitch, volume, frequency, etc. Then, these speech features are correlated with emotional states through the trained model [15]. Finally, new speech input is predicted to determine the emotional state of the speaker, thus enhancing the smooth flow of the conversation and further improving the service quality [16, 17].

The original emotional speech signal is set to come from the cabin service dialog scene, and the speech collected by the speech sensor is preprocessed, and its main processing procedures include pre-emphasis, frame splitting, window addition, and endpoint detection, which can reduce the noise in the original emotional speech signal and thus improve the quality of speech emotion feature extraction. The Meier acoustic spectrogram and Meier cepstrum coefficients are then used to carry out speech emotion feature extraction, and the RepVGG network is used to construct a speech emotion recognition model, and the limitations of the RepVGG network are detected, and in this regard, it is proposed to introduce CBAM to realize the optimization of the RepVGG network, in order to enhance the performance of the RepVGG network's speech emotion recognition and ultimately, design a speech emotion recognition model based on the improved RepVGG network. RepVGG network, and finally design a speech emotion recognition model based on the improved RepVGG network. Immediately after that, the speech emotion recognition technology is directly applied to the service process of the cabin service dialog scene, which can detect the emotional state of the customer service and the user in real time, thus establishing a cabin service quality detection platform, and revealing the degree of correlation between the speech emotion recognition and service quality of the cabin service dialog scene through the empirical analysis of the platform.

2 Research on Speech Emotion Recognition for Cabin Service Conversation Scenarios

2.1 Speech preprocessing for cabin service dialog scenarios

Raw speech signals are usually derived from cabin service conversation scenarios, and the speech collected by the speech sensors is preprocessed, and the main processing flow includes pre-emphasis, frame-splitting, windowing, and end-point detection, which reduces the noise in the raw speech signals and thus improves the quality of feature extraction.

2.1.1 Pre-exacerbation

The intensity interval of the speech signal is usually between 300Hz and 3500Hz, and when the speech signal reaches 8000Hz, the energy value decreases. At this time it is necessary to increase the energy of the high frequency region through the method of pre-emphasis, so that the signal spectrum changes more smoothly in energy. The calculation process is shown in

equation (1) equation (2):

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

$$\overline{s(n)} = s(n) - as(n-1) \tag{2}$$

where $H(z)$ is the first-order high-pass filter, a is the pre-emphasis coefficient, close to 1, $s(n)$ represents the speech signal, and $\overline{s(n)}$ is the speech signal after the pre-emphasis filter at the nth moment, which is commonly used to take values between 0.938 and 0.972.

2.1.2 Splitting frames and adding windows

Since the speed of the articulation organ is much slower compared to the sound vibration, the digitized speech signal will have short-term smoothness. And when the speech signal is in the range of 8~33ms, each characteristic parameter of the signal remains basically unchanged. Therefore, the short-time processing of the speech signal is carried out first, and the signal is divided into several frames, and each frame is generated by the previous frame through frame shifting, so there will be some overlap in the middle of the neighboring frames, and in this way, the processing of the continuous speech signal is converted into the processing of the continuous features. The windowing of the speech signal is required in the frame splitting. The windowing function is shown in equation (3):

$$s_w(n) = s(n)w(n) \tag{3}$$

where $s(n)$ is the input speech signal, $s_w(n)$ is the speech signal after the windowing, $w(n)$ is the windowing function, the windowing function generally has a rectangular window, Hamming window, triangular window, Hanning window, etc., of which the Hamming window and rectangular window are more common. Rectangular window will make the bandwidth of the signal increase, resulting in discontinuity between the signals. The rectangular window function is shown in equation (4):

$$w(n) = \begin{cases} 1, & 0 \le n \le N-1 \\ 0, & \text{Other} \end{cases} \tag{4}$$

The Hamming window will make the speech frames smoother from each other and can effectively express the short-time frequency characteristics. The Hamming window function is shown in equation (5):

$$w(n) = \begin{cases} 0.54 - 0.46\cos\left[\dfrac{2\pi n}{N-1}\right], & 0 \le n \le N-1 \\ 0, & \text{Other} \end{cases} \tag{5}$$

where N denotes the total number of frames after the signal has been subframed and windowed.

2.1.3 Endpoint detection

Endpoint detection algorithms (VAD) refers to the use of relevant technical methods to detect

the start and end positions of valid speech in order to remove silent segments from speech. The common endpoint detection methods are double threshold method, variance method, spectral distance method, etc. In this paper, double threshold method is used for endpoint detection. The double threshold method is proposed based on the short-time energy and short-time average zero crossing rate, where the short-time energy indicates the energy of a short segment of speech, and the short-time average zero crossing rate indicates the frequency of crossing zero per unit time in the time domain waveform.

The speech signal $s(n)$ is $s_i(m)$ after framing and windowing, where i denotes frame i, and the energy of each frame is calculated as shown in equation (6):

$$AMP_i = \sum_{m=1}^{N} s_i^2(m) \tag{6}$$

where N represents the total frame length obtained after the signal has been divided into frames and windowing operations, and AMP_i represents the energy calculated for each frame of the signal.

Because of the zero drift phenomenon may occur in the speech signal, in order to ensure the stability of the calculated zero rate, it is necessary to carry out the center truncation processing of $s(n)$, the center truncation processing as shown in equation (7):

$$s_i(m) = \begin{cases} s_i(m), & |s_i(m)| > \delta \\ 0, & |s_i(m)| < \delta \end{cases} \tag{7}$$

where δ represents a very small value.

After the center cutoff the over-zero rate is calculated for each frame and the function for calculating the over-zero rate is shown in equation (8):

$$ZCR_i = \sum_{m=1}^{N} \left| sign\left[s_i(m) \right] - sign\left[s_i(m-1) \right] \right| \tag{8}$$

where $sign\left[s_i(m) \right]$ is shown in equation (9):

$$sign\left[s_i(m) \right] = \begin{cases} 1, & |s_i(m)| \geq 0 \\ -1, & |s_i(m)| < 0 \end{cases} \tag{9}$$

2.2 Speech emotion feature extraction

The common speech emotion features in the cabin service dialog scene are Mel's acoustic spectrogram, short-time energy, Mel's frequency cepstrum coefficient, over-zero rate, etc., while Mel's acoustic spectrogram and Mel's cepstrum coefficients are used in this paper to carry out the work of speech emotion feature extraction.

2.2.1 Mel Sound Spectrogram (Mel)

Speech signal analysis mainly contains time domain analysis and frequency domain analysis. In the time domain analysis, the speech signal can be directly expressed by its waveform, and

the important characteristics of speech can be seen by observing the time domain waveform. According to the voice signal has a short time invariant characteristics, so the voice signal can be converted to the frequency domain through the Fourier transform, the frequency domain can analyze the signal amplitude with the frequency change. However, both time domain analysis and frequency domain analysis have limitations, time domain analysis can not understand the frequency of the signal, and frequency domain analysis can not observe the relationship between the signal transformed over time. Therefore, short time Fourier transform (STFT) can be used for time domain and frequency domain analysis. The acquired signal is first sub-framed, windowed, and then STFT is performed frame by frame, and finally the results are stacked in another dimension to obtain a two-dimensional acoustic spectrogram feature. After time-frequency analysis of the speech signal using STFT, in order to obtain a smaller acoustic spectrogram feature, the acoustic spectrogram must first be transformed into a Mel spectrum by a Mel scale filter bank. The calculation of the Mel spectrum is shown in equation (10):

$$mel(l) = 2595 * \log_{10}\left(1 + \frac{f}{700}\right) \tag{10}$$

where f denotes the frequency and $mel(l)$ denotes the corresponding Meier frequency.

2.2.2 Meier inverse spectral coefficient (MFCC)

MFCC is based on the human hearing mechanism and the results of listening experiments to analyze the spectrum of the speech signal to obtain better sound characteristics, the MFCC feature extraction process is shown in Figure 1:

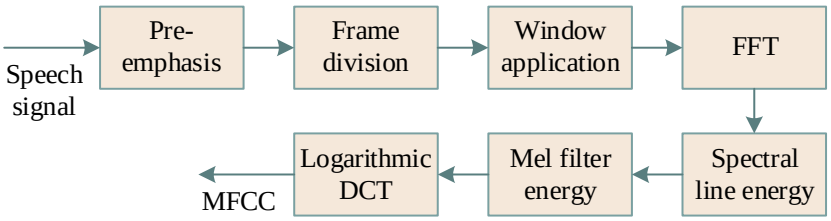


Figure 1: MFCC feature extraction process

(1) Preprocessing

The captured speech is preprocessed. Firstly, the original signal is pre-emphasized to compensate for the loss of high-frequency components in the signal, secondly, framing is performed based on the short-time characteristics of the signal, and finally, the leakage in the frequency domain is reduced by adding windows. The calculation process of pre-emphasis has been introduced in Section 2.1.1, while the calculation process of frame splitting and windowing has been introduced in Section 2.1.2.

(2) Fast Fourier Transform

The Fast Fourier Transform (FFT) is performed on the signal after preprocessing, frame by frame, to convert the signal from the time domain to the frequency domain. The FFT transform from time domain to frequency domain is shown in equation (11):

$$X(i,k) = FFT\left[X_i(m)\right] \tag{11}$$

where $X_i(m)$ represents the sequence obtained after speech preprocessing, where i

represents the ird frame of data after framing, and k represents the kth spectral line in the frequency domain of the signal.

(3) Spectral energy

After the FFT transform, the spectral energy of each frame needs to be calculated, and the function of spectral energy is shown in equation (12):

$$E(i,k)=[X(i,k)]^2 \tag{12}$$

where $E(i,k)$ denotes the calculated spectral line energy.

(4) Calculation of energy after Mel filtering

The energy of each frame $E(i,k)$ in Mel filter $H_m(k)$ is calculated using Mel filter as shown in equation (13):

$$S(i,m)=\sum_{k=0}^{N-1}E(i,k)H_m(k),\ 0\le m<M \tag{13}$$

where i denotes the ind frame of spectral energy, k denotes the kth spectral line in the frequency domain, $H_m(k)$ denotes a set of bandpass filters, and M denotes the number of filters, usually M is an integer between 20 and 30. All bandpass filters are characterized by triangular filtering, where the triangular bandpass filter is able to smooth the spectrum as well as reduce harmonics and highlight resonance peaks in speech. The Mel filter is defined as shown in equation (14):

$$H_m(k)=\begin{cases}0, & k<f(m-1)\\ \dfrac{k-f(m-1)}{f(m)-f(m-1)}, & f(m-1)\le k\le f(m)\\ \dfrac{f(m+1)-k}{f(m+1)-f(m)}, & f(m)<k\le f(m+1)\\ 0, & k>f(m+1)\end{cases} \tag{14}$$

where $f(m)$ is the center frequency of the filter, m is greater than or equal to 0 and less than or equal to M.

(5) Calculate the discrete cosine transform (DCT)

Calculate the DCT of the Mayer filter by taking the logarithm of its total energy $S(i,m)$ as shown in equation (15):

$$mfcc(i,n)=\sqrt{\frac{2}{M}}\sum_{m=0}^{M-1}\log\left[S(i,m)\right]\cos\left(\frac{\pi n(2m-1)}{2M}\right) \tag{15}$$

where n is the spectrum after DCT transformation and $mfcc(i,n)$ denotes the obtained MFCC features.

2.3 Speech emotion recognition model based on improved RepVGG network

In order to improve the recognition accuracy and inference speed of the speech emotion

recognition model in the cabin service dialog scene application, RepVGG network is selected as the baseline model. And the model is improved based on the attention mechanism, so that the network focuses on the key emotion feature information and suppresses the emotion-irrelevant information to improve the emotion recognition performance.

2.3.1 RepVGG network modeling

The goal of the research in this paper is to realize the emotion recognition of two independent speakers based on the cabin service dialogue scenario, and in the functional requirements of the actual application, it is hoped that the model can achieve a certain degree of real-time, and realize the rapid inference and recognition of the emotional state.The network structure of the RepVGG in the training phase and the inference phase is schematically illustrated, and the structure of the RepVGG network is shown in Fig. 2.The network structure of the Rep VGG model is divided into training phase and inference phase. In the training phase, the 3×3 convolutional layers of the Rep VGG Block are designed as a multi-branch structure with reference to the residual connections in the ResNet network structure, and the whole Rep VGG network is composed of continuously stacking Rep VGG Blocks. In addition, in contrast to ResNet, the module incorporates a multi-branching structure for each 3×3 convolutional layer in the training phase, and the 3×3 convolutional layers for each layer have a branching design as shown in the following equation:

$$y = x + g(x) + f(x) \tag{16}$$

In Eq. (16), x denotes the Identity branch, which performs a constant mapping of the input features; $g(x)$ denotes the convolutional branch of 1×1; and $f(x)$ denotes the main branch of the convolutional layer of 3×3, where a BN layer is added to each branch before summing.

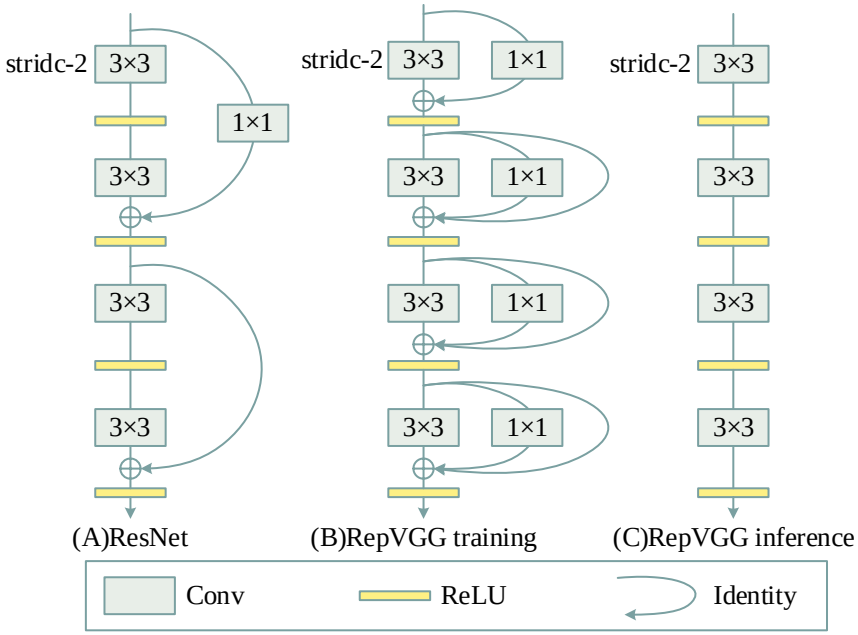
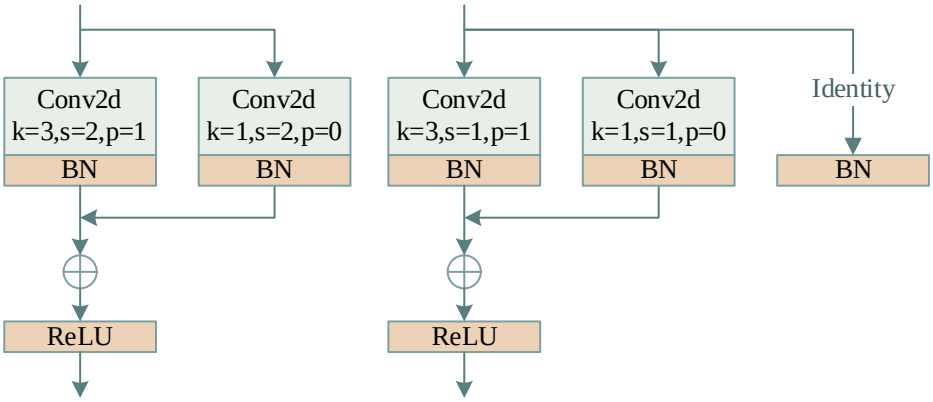


Figure 2: RepVGG network structure

The branching structure of Rep VGG Block in the training phase is shown in Figure 3. In the training phase, the parallel multi-branch network structure can improve the model's

characterization ability, better help the network training, and improve the model accuracy. In the inference stage, Rep VGG decouples and removes the multi-branches through structural parameter reconstruction, and the multi-branched network structure is algebraically transformed to be equivalent to a one-way 3×3-convolutional network structure for inference, which greatly accelerates the inference speed of the model and reduces the inference time.



(a) Downsampling Branch Structure

(b) Parallel Three-Way Branch Structure

Figure 3: The RepVGG-Block branch structure diagram during the training phase

The structural reparameterization process performed by RepVGG in the inference stage is shown in Fig. 4. Since the convolutional and BN layers in all three branches are linear operations, it is possible to fuse the convolutional and BN layers and equate them to a new convolutional operator with bias. In this case, the BN layer for input feature x is calculated as follows:

$$BN\left(x\right)=\gamma\cdot\frac{x-\mu}{\sqrt{\sigma^2+\varepsilon}}+\beta \tag{17}$$

In the above equation (17), x denotes the input features.The BN layer has four parameters: μ, σ, γ, and β denote the mean, standard deviation, scale factor, and bias parameter, respectively, where ε is an infinitesimal value in order to prevent the denominator from being 0. The BN layer has four parameters: 2, 3, 4, and 5 denote the mean, standard deviation, scale factor, and bias parameter, respectively. Since the BN layer fits a bias parameter to each channel, the convolution computation is brought into the BN layer for fusion, the convolution layer does not use bias, and the fused equation is as follows:

$$BN\left(Conv\left(x\right)\right)=\gamma\cdot\frac{w\cdot x-\mu}{\sqrt{\sigma^2+\varepsilon}}+\beta \tag{18}$$

In the above equation (18), w denotes the weight parameter in the convolution calculation and the bias of the convolution is omitted. Further simplification of the equation leads to the following equation:

$$BN\left(Conv\left(x\right)\right)=\left(\frac{\gamma\cdot w}{\sqrt{\sigma^2+\varepsilon}}\right)\cdot x-\frac{\gamma\cdot\mu}{\sqrt{\sigma^2+\varepsilon}}+\beta=w_{bn}\cdot x+b \tag{19}$$

In a three-branch structure, a convolutional branch of 1×1 can be represented as a convolution of 3×3 using zero padding. Similarly, for the constant mapping in the Identity branch can be viewed as a convolution of 1×1 with the unit matrix as the convolution kernel, and then zero-padded into a convolution of 3×3. In this way all three branches are equivalently treated as convolutional layers of 3×3 with bias, and the equivalent formula for each branch structure is as follows:

$$BN_{id}\left(Conv_{id}\left(x\right)\right)=w_0\cdot x+b_0 \tag{20}$$

$$BN_{1\times1}\left(Conv_{1\times1}\left(x\right)\right)=w_1\cdot x+b_1 \tag{21}$$

$$BN_{3\times3}\left(Conv_{3\times3}\left(x\right)\right)=w_3\cdot x+b_3 \tag{22}$$

And then get:

$$Conv_{3\times3}\left(x\right)=\left(w_0+w_1+w_3\right)\cdot x+\left(b_0+b_1+b_3\right) \tag{23}$$

After the above two steps of structure reparameterization, the multi-branch structure is successfully reconstructed into a single path structure in the inference phase. Meanwhile, according to the number of stacks of Rep VGG Blocks, Rep VGG networks are categorized into different structure types.

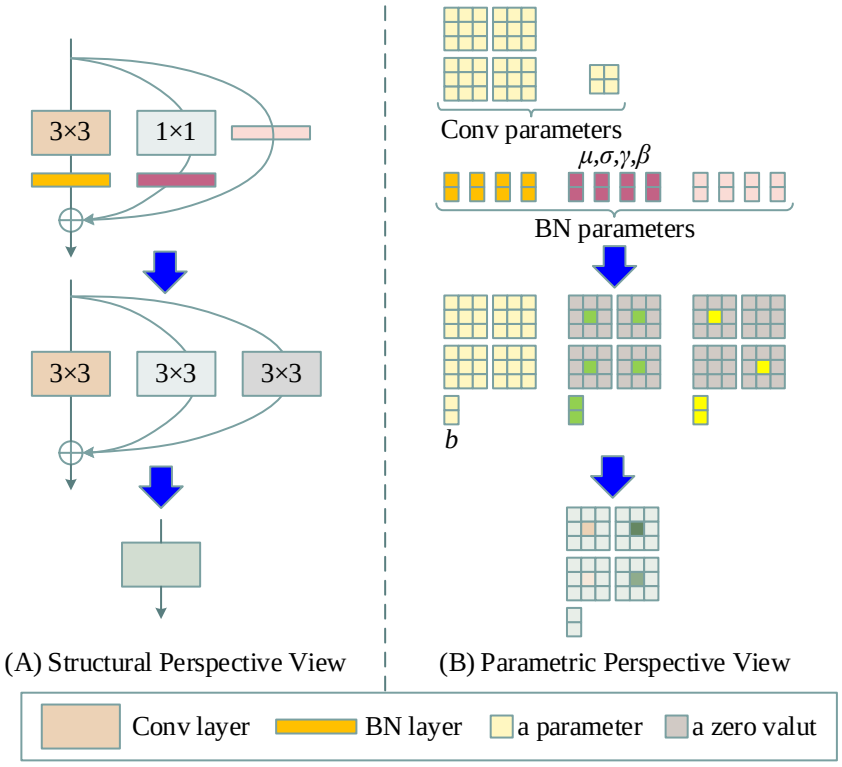


Figure 4: RepVGG structure reparameterization process

2.3.2 Model improvement based on attention mechanism

In speech emotion recognition tasks, feature information in different channel dimensions and spatial dimensions may have different importance for emotion expression. The baseline model structure is improved by utilizing the Convolutional Attention Module (CBAM).The CBAM attention module consists of two sub-modules, the Channel Attention Module (CAM) and the Spatial Attention Module (SAM). The structure of the channel attention module is shown in Fig. 5.The channel attention module is mainly used to enhance the attention to the channel dimension information in the input features. Firstly, the global maximum feature vector and average feature vector of each channel are obtained through maximum pooling and average pooling operations, respectively. Then input into a shared fully connected layer to compute the attention weights. Finally the weights of the features are restricted to be between 0 and 1 by a Sigmoid function. After channel attention weighting, the channel weights of the input features can be adaptively adjusted so that the model can focus more on the important channel information in the current task and suppress irrelevant channel information.

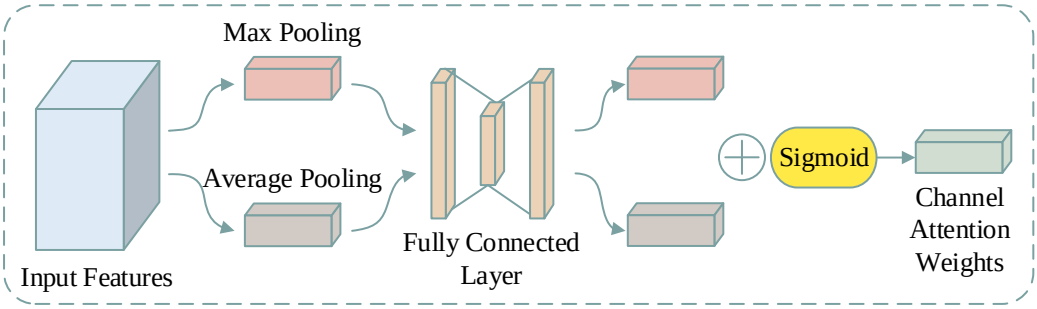


Figure 5: Structure diagram of channel attention module

The spatial attention module is used to calculate the importance of feature information in the spatial dimension in order to better capture the spatial structure in the spectrogram, and the structure of the spatial attention module is shown in Fig. 6. Firstly, the CAM output feature map is used as the input to this module and maximum pooling and average pooling are performed in the spatial dimension and the two pooling results are spliced. Then the attention weights of each pixel are obtained by a convolutional layer and a Sigmoid activation function. Finally the feature map is weighted using the spatial attention weights to better capture the spatial structure in the speech spectrogram features.

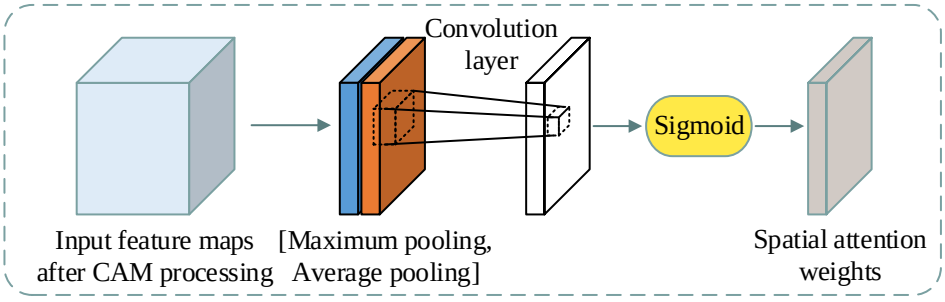


Figure 6: Structure diagram of spatial attention module

A complete convolutional attention module is obtained by combining the channel attention module and the spatial attention module, and the structure of the convolutional attention module module is shown in Fig. 7. First, the input features are passed through CAM to get the channel attention weights, and multiplied with the original feature map to get the channel attention

weighted feature map. Then, the channel attention-weighted feature map is input to SAM to get the spatial attention weights, and the channel attention-weighted feature map is multiplied with the spatial attention weights to finally output the CBAM-weighted feature map. By flexibly embedding CBAM into various CNN architectures, it can effectively help the network to pay attention to more important feature expressions.

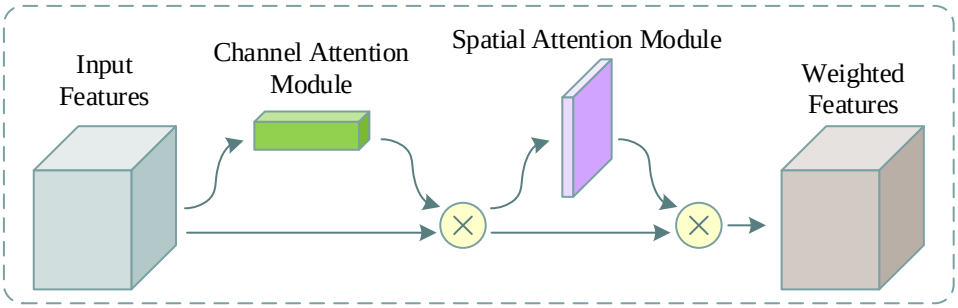


Figure 7: Structure diagram of convolutional block attention modu le

The emotion information embedded in the input spectral features is distributed in each channel and space, the learning of the emotion feature information will be carried out by the convolutional network in the training stage, but it does not pay attention to the connection and importance between the feature information in each channel and space, the initial layer network structure of RepVGG which is improved based on the attention mechanism is shown in Fig. 8. In the initial layer network structure of RepVGG in the training phase, the CBAM attention module is designed after the convolutional layer. Since the main role of the initial layer is feature extraction by downsampling, the features of this layer contain a large number of redundant features that are irrelevant to the sentiment information. Therefore, adding the attention mechanism after the initial layer enables the model to adaptively adjust the weights of the input features, suppress the redundant features and better capture the emotion key information in the spectrogram features, improving the performance and accuracy of the classification of speech emotion.

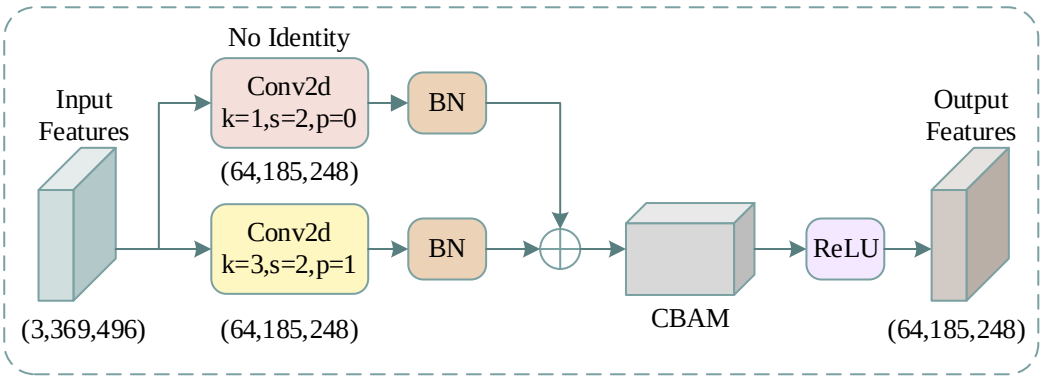


Figure 8: Initial layer structure diagram of RepVGG based on attention improvement

The existing RepVGG-A1 network structure is improved based on the CBAM attention mechanism, and the specific network parameters such as the size of the output dimension of each layer of the network, the size of the convolutional kernel in each branch, the pooling layer, and the fully connected layer are designed as shown in Table 1.

Table 1: Improved RepVGG-A1l network parameter design

Stage	Output dimension	Network parameters
Stage0	$(64,185,248)$	$\begin{bmatrix}3\times3,64\\1\times1,64\end{bmatrix}\times1$ $CBAM,64\times1$
Stage1	$(64,93,124)$	$\begin{bmatrix}3\times3,64\\1\times1,64\end{bmatrix}\times2$ $Identity\times1$
Stage2	$(128,47,62)$	$\begin{bmatrix}3\times3,128\\1\times1,128\end{bmatrix}\times4$ $Identity\times3$
Stage3	$(256,24,31)$	$\begin{bmatrix}3\times3,256\\1\times1,256\end{bmatrix}\times14$ $Identity\times13$
Stage4	$(1280,12,16)$	$\begin{bmatrix}3\times3,1280\\1\times1,1280\end{bmatrix}\times1$
Avg Pooland Flatten	(1280)	
Fully connected layer	$(num\_classes)$	$Linear(1280,num\_classes)$

3 Quality of Service Testing Platform

By directly applying the speech emotion recognition technology to the service process of the cabin service dialogue scene, the emotional state of the customer service and the user can be detected in real time, so as to realize the correlation between the speech emotion recognition and the service quality of the cabin service dialogue scene, and thus improve the overall service quality of the cabin service. In addition, a service quality detection platform is constructed based on the above speech emotion recognition model to improve the overall image of the cabin service as well as the user reputation.

3.1 Demand Analysis of Service Quality Testing Platform

3.1.1 Functional requirements

The main purpose of functional requirements is to clarify and define the functions that a software system should have in order to meet the needs and expectations of users. Through the clear definition of functional requirements, it can ensure that the software development team and users agree on the system functions, reduce the demand changes and corrections in the later development, and improve the development efficiency and software quality. Through the above analysis of the requirements of the service quality inspection platform, the specific functional modules can be divided into four parts: speech emotion recognition, auditing recognition results, audio clip editing, and querying and display of quality inspection results.

3.1.2 Non-functional requirements

Non-functional requirements refer to the conditions that need to be met according to the reality and system characteristics in addition to the functions that must be realized by the platform, and these non-functional requirements are also an indispensable part of the platform, which will have a great impact on the platform's performance, efficiency and other aspects. In this paper, the non-functional requirements that need to be considered for the service quality testing platform mainly include the following points:

(1) Performance requirements. Including the platform's overall corresponding time, throughput, concurrency, etc., such as the time required to complete the emotion recognition of a 5-minute business audio, or how many reviewers the platform can support to complete the review task of the audio online at the same time.

(2) Reliability requirements. The platform needs to maintain stability and reliability in the process of operation, can not be frequent crashes, need to have a certain fault tolerance mechanism.

(3) Security requirements. The platform needs to have access control, encrypted transmission, data protection and other functions to protect the system from malicious attacks or illegal access.

(4) Availability requirements. The platform needs to ensure the availability and accessibility, and when the platform fails, it needs to have data recovery and other mechanisms.

(5) Maintainability requirements. The entire platform needs to be easy to modify, expand and maintain, and can meet the subsequent addition of new features.

3.2 Design of Quality of Service Testing Platform

3.2.1 Platform system architecture

The overall system architecture of the platform is shown in Figure 9. The QoS platform is developed using the browser/server (B/S) architecture model, and the underlying database uses the relational database MySQL and the non-relational database Redis, which is mainly used for storing data related to the system, while Redis is responsible for storing the messages of asynchronous tasks in the system. The overall back-end framework is Django, using the REST design pattern for the unified development of the back-end interface, in addition, a large number of time-consuming tasks in the system are processed using Celery, a task queue based on distributed messaging, which can handle a large number of tasks in an asynchronous manner to prevent the system from blocking. The front-end uses the progressive JavaScript framework Vue3, which facilitates user interaction with the system.

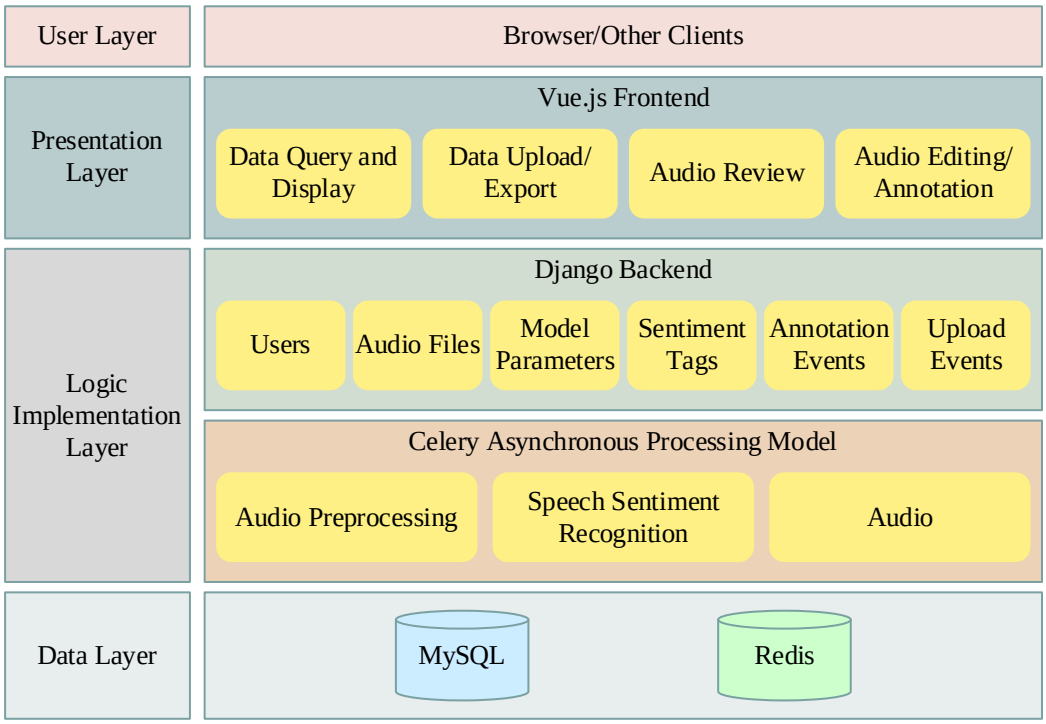


Figure 9: Overall system architecture of the platform

3.2.2 Detailed design of the platform

Detailed design of the platform is an important stage in the software development process, which is to further analyze and refine the functions and requirements of the system, clarify the interrelationships between the various modules of the system as well as the interfaces and interactions between the modules, and ultimately transform the high-level design of the system into specific implementation plans and technical specifications, providing detailed guidance and basis for the development and implementation of the system to ensure that the system is able to The development and realization of the system is completed according to the expected requirements and standards.

4 Analysis of empirical studies

4.1 Speech emotion recognition model validation analysis

4.1.1 Data sets

The dataset for the study of speech emotion recognition in cabin service dialog scenarios is a homemade dataset, respectively, which can be set as dataset A and dataset B. The distribution of the duration of each speech in dataset A is shown in Fig. 10, and this dataset contains a total of 5,526 speech samples, including 1,106 angry, 1,077 sad, 1,716 neutral, and 1,627 happy samples, which are mostly in the duration of 1.053 to 5.172 seconds, which provides data basis for the research work to be carried out.

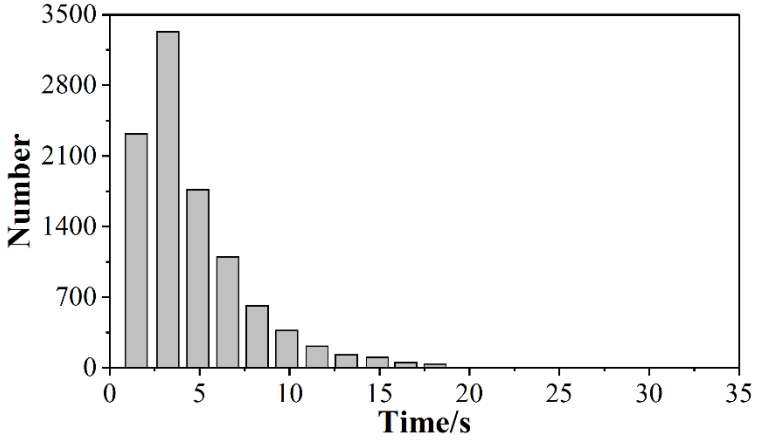


Figure 10: The duration distribution of each speech in dataset A

The distribution of the duration of each voice in dataset B is shown in Figure 11. It covers eight emotion categories: anger, sadness, happiness, surprise, neutrality, disgust, fear, and calmness. Out of a total of 1,439 speech samples, there are 193 samples for each of the other emotion categories, except for the neutral category, which has only 88 samples. This dataset has a relatively balanced distribution of emotions and does not require additional processing, and its audio samples are mainly distributed between 3.023 and 4.051 seconds in duration, which facilitates the analysis of the study below.

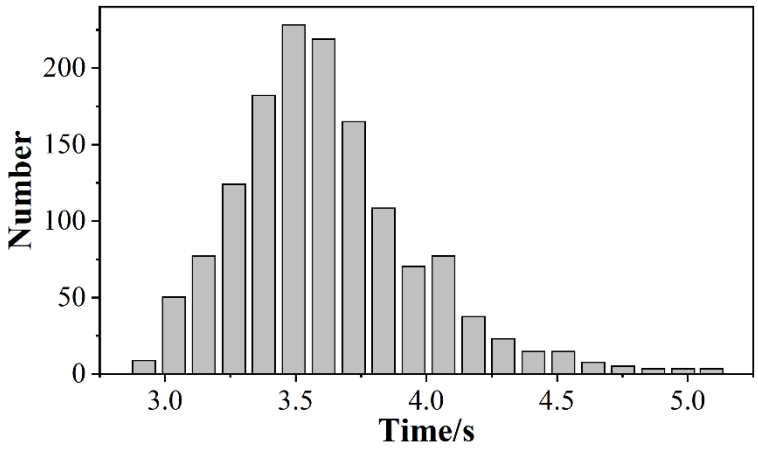


Figure 11: The duration distribution of each speech in dataset B

4.1.2 Assessment of indicators

Speech emotion recognition can basically be regarded as a multi-classification problem. In dealing with cabin dialog service scenarios, commonly used evaluation criteria include precision rate, recall rate and F1 score. However, when dealing with speech emotion recognition datasets using discrete sentiment labels, the uneven distribution of categories may result in the fact that a single use of precision rate as an evaluation criterion is not sufficient to reflect the comprehensiveness of the model. This happens because the model tends to be more biased towards those categories with a higher number of samples in them, while ignoring categories with fewer samples. Therefore, in order to more fairly assess the contribution of each category to the overall sample accuracy, weighted accuracy (WA), unweighted accuracy (UA), and F1 scores were used as model performance assessment metrics in this study.

4.1.3 Analysis of results

In order to deeply analyze the performance of the RepVGG-CBAM-based speech emotion recognition model, this study demonstrates the emotion recognition confusion matrices of the model on datasets A and B. The emotion recognition confusion matrix for dataset A is shown in Figure 12, and the emotion recognition confusion matrix for dataset B is shown in Figure 13. In dataset A, the accuracy of recognizing happy emotions is the lowest, a phenomenon that is consistent with most studies. This is mainly due to the fusion of excited and happy emotions and the diversity of ways to express happy emotions. The behavior of different individuals when expressing happiness varies significantly, and some may show extreme excitement while others may remain relatively calm, all of which add to the complexity of recognizing happy emotions. Comparatively speaking, sad emotions are better recognized, thanks to the consistency and obviousness of their expression, which leads to a relatively high recognition rate. In the evaluation of the RepVGG-CBAM model, the recognition accuracies of the four core emotions: calm, sad, angry, and happy were 73.29%, 84.29%, 71.45%, and 66.79%, in that order. The recognition rates of neutral, calm and angry on dataset B were all over 90%. This indicates that the speech emotion recognition model based on RepVGG-CBAM has excellent performance performance.

Figure 12: Emotion recognition confusion matrix of Dataset A

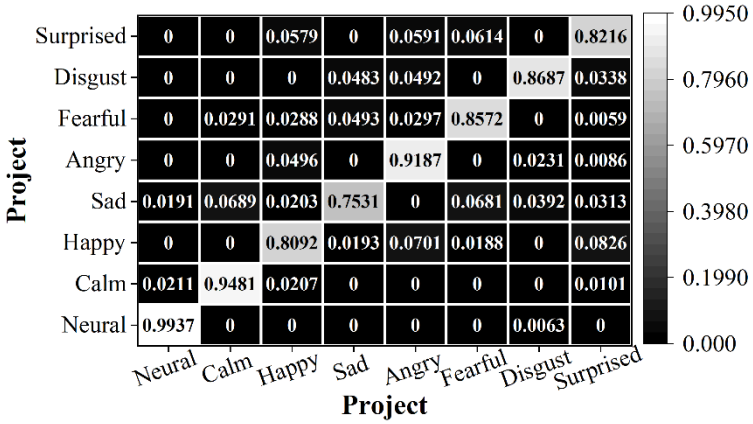


Figure 13: Emotion recognition confusion matrix of Dataset B

In order to prove the effect of each module in the model, a series of ablation experiments will be carried out to test it, and the results of the analysis of the ablation experiments on dataset A are shown in Fig. 14, and the results of the analysis of the ablation experiments on dataset B

are shown in Fig. 15. In both dataset A and dataset B, after the introduction of CBAM, Mel features, and MFCC features on the baseline RepVGG model, the model's speech emotion recognition performance indicators WA, UA, and F1 are all improved, which proves that the adoption of Mel and MFCC features to characterize the speech emotion of the cabin service dialogue scene is conducive to improving the recognition of the RepVGG-CBAM algorithm performance, and also proves that CBAM improves the baseline RepVGG model, which in turn helps the development and innovation of speech emotion recognition technology for cabin service dialog scenes.

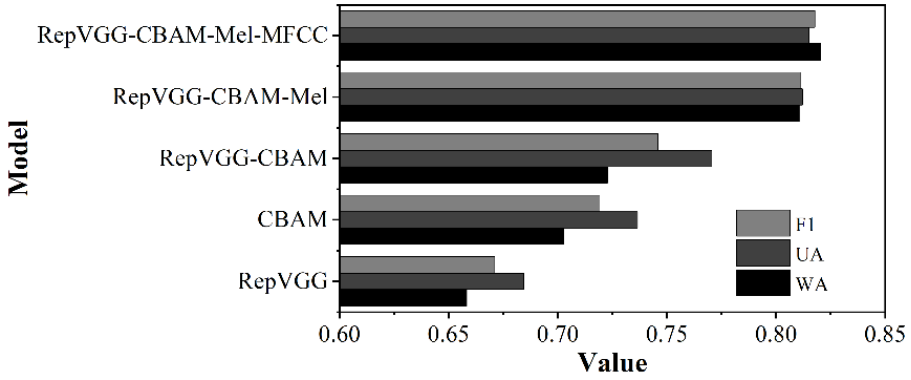


Figure 14: Analysis results of the ablation experiment on dataset A

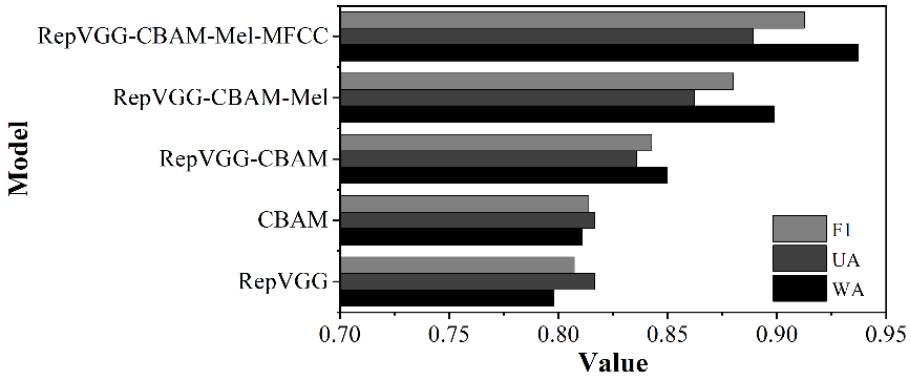


Figure 15: Analysis results of the ablation experiment on dataset B

4.2 Empirical analysis of service quality testing platforms

After verifying the speech emotion recognition model, laying the theory for the construction of the service quality detection platform, the next comprehensive analysis of the actual application of the service quality detection platform. The research data for the empirical analysis of the service quality detection platform is obtained through a questionnaire survey, in which the questionnaire contains a test scale for the ability to process voice emotion signals, a test scale for the ability to recognize voice emotion signals, a test scale for the ability to classify voice emotion signals, and a test scale for the platform's service quality, which is a five-point scaled Likert scale. On the basis of obtaining the quantitative values, the correlation analysis of the quantitative values was carried out to reveal the correlation between speech emotion recognition and service quality in the cabin service dialog scene.

4.2.1 Speech emotion signal processing capability statistics

Randomly selected after 100 users as the service quality testing platform experience object, after the experience of their questionnaire survey, to obtain the initial data of voice emotion

signal processing ability, voice emotion signal processing ability statistics shown in Table 2. Combined with the data in the table, it can be seen that the 100 users of the service quality detection platform of the voice emotion signal processing ability of the quantitative value of 3.5 or more, after calculation of the service quality detection platform of the voice emotion signal processing ability of the quantitative mean value of 3.7596, which indicates that the users of the service quality detection platform of the voice emotion signal processing ability of the recognition. For example, in the cabin service dialog scene, it is necessary to make a call first and enter the corresponding number, which is cumbersome, and in the peak period of consultation, there is also a waiting situation, so it is a waste of time and complicated. However, through the voice emotion signal processing capability of the service quality inspection platform, it is possible to dial a number or automatically transfer the contact after saying the contact, thus realizing the intelligent development of cabin service.

Table 2: Statistics on the processing capability of voice emotional signals

User	Value	User	Value	User	Value	User	Value	User	Value
1	3.873	21	3.807	41	3.755	61	3.69	81	3.701
2	3.959	22	3.675	42	3.864	62	3.659	82	3.926
3	3.867	23	3.535	43	3.5	63	3.742	83	3.911
4	3.507	24	3.786	44	3.774	64	3.618	84	3.856
5	3.711	25	3.962	45	3.708	65	3.764	85	3.632
6	3.795	26	3.537	46	3.649	66	3.772	86	3.678
7	3.655	27	3.948	47	3.501	67	3.762	87	3.787
8	3.649	28	3.782	48	3.66	68	3.797	88	3.879
9	3.89	29	3.61	49	3.847	69	3.67	89	3.821
10	3.566	30	3.818	50	3.957	70	3.997	90	3.756
11	3.833	31	3.714	51	3.644	71	3.949	91	3.904
12	3.968	32	3.625	52	3.977	72	3.768	92	3.685
13	3.728	33	3.83	53	3.905	73	3.972	93	3.961
14	3.539	34	3.999	54	3.534	74	3.871	94	3.649
15	3.89	35	3.63	55	3.781	75	3.965	95	3.853
16	3.801	36	3.695	56	3.651	76	3.777	96	3.787
17	3.611	37	3.521	57	3.694	77	3.558	97	3.82
18	3.96	38	3.563	58	3.525	78	3.902	98	3.971
19	3.739	39	3.774	59	3.638	79	3.78	99	3.726
20	3.775	40	3.713	60	3.691	80	3.581	100	3.97
Mean	3.7596								

4.2.2 Statistics on Recognition of Speech Emotion Signals

Using the same method described above, the quantitative test analysis of the speech emotion signal recognition ability of the service quality detection platform is carried out, and the statistics of the speech emotion signal recognition ability are shown in Table 3. In the known service quality detection platform voice emotion signal recognition ability quantitative value distribution interval is 3.507~3.985, calculated to get the mean value of 3.7248, indicating that users are satisfied with the service quality detection platform voice emotion signal processing ability. For example, the continuous increase in the number of cabin seats requires an increase in the application of intelligent speech recognition technology to improve the management and service quality of the service quality detection platform, while also solving diversified and complex operations, and contributing to the high-quality development of cabin services.

Table 3: Statistics on the ability to recognize voice emotional signals

User	Value	User	Value	User	Value	User	Value	User	Value
1	3.606	21	3.942	41	3.811	61	3.868	81	3.52
2	3.536	22	3.507	42	3.882	62	3.543	82	3.713
3	3.701	23	3.841	43	3.606	63	3.829	83	3.818
4	3.844	24	3.828	44	3.521	64	3.921	84	3.565
5	3.564	25	3.612	45	3.764	65	3.655	85	3.577
6	3.753	26	3.815	46	3.855	66	3.869	86	3.927
7	3.648	27	3.846	47	3.715	67	3.517	87	3.584
8	3.727	28	3.721	48	3.597	68	3.581	88	3.762
9	3.539	29	3.51	49	3.756	69	3.58	89	3.8
10	3.637	30	3.601	50	3.909	70	3.65	90	3.64
11	3.582	31	3.985	51	3.788	71	3.691	91	3.594
12	3.835	32	3.791	52	3.7	72	3.697	92	3.605
13	3.817	33	3.881	53	3.76	73	3.67	93	3.865
14	3.533	34	3.797	54	3.941	74	3.826	94	3.651
15	3.906	35	3.867	55	3.879	75	3.575	95	3.533
16	3.689	36	3.894	56	3.553	76	3.836	96	3.786
17	3.897	37	3.97	57	3.92	77	3.619	97	3.581
18	3.696	38	3.717	58	3.702	78	3.752	98	3.751
19	3.615	39	3.763	59	3.753	79	3.672	99	3.731
20	3.788	40	3.518	60	3.685	80	3.729	100	3.981
Mean	3.7248								

4.2.3 Statistics on the ability to categorize voice emotion signals

Also with the help of the scale test, the quantitative value of the voice emotion signal classification ability of the service quality detection platform is obtained and statistically analyzed, and the statistics of the voice emotion signal classification ability is shown in Table 4. After the calculation of the quantitative value of the voice emotion signal classification ability of the service quality detection platform, the mean value of the quantitative value is 4.3466, and its numerical distribution range is 4.115~4.599, indicating that the 100 experienced users hold a very satisfactory attitude towards the voice emotion signal classification ability of the service quality detection platform. For example, when a customer asks a question by voice, the keywords of the desired service can be output by voice, and the service quality detection platform improves the service ability and work efficiency of the agents through voice emotion signal classification, saves communication time, and can provide users with efficient and convenient services.

Table 4: Statistics on the classification ability of voice emotional signals

User	Value	User	Value	User	Value	User	Value	User	Value
1	4.53	21	4.361	41	4.29	61	4.51	81	4.141
2	4.308	22	4.19	42	4.252	62	4.239	82	4.15
3	4.141	23	4.246	43	4.403	63	4.579	83	4.388
4	4.327	24	4.122	44	4.216	64	4.224	84	4.526
5	4.301	25	4.323	45	4.29	65	4.255	85	4.358
6	4.599	26	4.356	46	4.276	66	4.481	86	4.435
7	4.442	27	4.39	47	4.301	67	4.138	87	4.515
8	4.266	28	4.585	48	4.432	68	4.128	88	4.317
9	4.51	29	4.323	49	4.158	69	4.462	89	4.389
10	4.233	30	4.387	50	4.305	70	4.311	90	4.46
11	4.136	31	4.403	51	4.294	71	4.152	91	4.241
12	4.456	32	4.394	52	4.227	72	4.292	92	4.154
13	4.115	33	4.191	53	4.317	73	4.423	93	4.124
14	4.449	34	4.199	54	4.172	74	4.447	94	4.586
15	4.384	35	4.384	55	4.505	75	4.136	95	4.554
16	4.197	36	4.299	56	4.574	76	4.535	96	4.579
17	4.522	37	4.367	57	4.45	77	4.341	97	4.462
18	4.464	38	4.224	58	4.397	78	4.175	98	4.22
19	4.37	39	4.537	59	4.564	79	4.409	99	4.514
20	4.36	40	4.413	60	4.487	80	4.327	100	4.267
Mean	4.3466								

4.2.4 Statistics on the quality of the platform's services

Using the platform service quality test scale in the questionnaire, the quantitative value of the platform service quality of the 100 experience users is obtained, and the platform service quality statistics are shown in Table 5. After statistical analysis, it can be obtained that the interval of platform service quality values is 3.9~4.393, and furthermore, the mean value of platform service quality is 4.1541, which demonstrates the recognition of platform service quality by 100 experience users. For example, on the basis of speech recognition technology, a service quality detection platform is established, which makes the service platform capable of recognizing interactive natural language, which can meet the needs of the cabin service dialogue, which enables the customers to directly select the required services, which can shorten the call time, which improves the quality of the customer's experience, and which greatly reduces the service time of the manual customers and reduces the intensity of the work, and therefore, improves the The quality and value of cabin service.

Table 5: Platform service quality statistics

User	Value	User	Value	User	Value	User	Value	User	Value
1	3.966	21	3.987	41	4.1	61	4.072	81	4.348
2	4.334	22	4.213	42	4.207	62	3.972	82	4.128
3	4.135	23	4.144	43	4.224	63	4.301	83	4.04
4	4.359	24	4.185	44	4.241	64	4.226	84	4.114
5	4.274	25	4.043	45	4.12	65	3.972	85	3.997
6	4.249	26	4.193	46	3.93	66	4.37	86	3.97
7	4.057	27	4.29	47	4.071	67	4.045	87	4.317
8	3.974	28	4.022	48	4.393	68	4.343	88	4.223
9	3.971	29	4.12	49	4.334	69	4.204	89	4.309
10	4.244	30	4.336	50	3.982	70	4.125	90	3.981
11	3.951	31	4.258	51	4.096	71	4.258	91	4.277
12	4.276	32	4.242	52	3.969	72	3.9	92	4.129
13	4.213	33	4.286	53	4.345	73	4.321	93	3.998
14	4.266	34	4.028	54	4.072	74	3.913	94	3.902
15	3.948	35	3.979	55	4.08	75	4.021	95	4.34
16	4.185	36	4.276	56	4.294	76	4.107	96	4.237
17	4.081	37	4.078	57	4.118	77	4.128	97	4.287
18	4.17	38	4.081	58	4.277	78	4.106	98	4.004
19	4.298	39	4.01	59	4.389	79	4.245	99	4.32
20	4.15	40	4.264	60	4.134	80	4.347	100	3.995
Mean	4.1541								

4.2.5 Correlation analysis

Through the statistical analysis above, the mean values of voice emotion signal processing ability, voice emotion signal recognition ability, voice emotion signal classification ability, and platform service quality are obtained, based on which the quantitative mean correlation analysis is carried out by using SPSS software, and the results of the correlation analysis are shown in Fig. 16, in which X1, X2, X3, and Y denote the voice emotion signal processing ability, voice emotion signal recognition ability, respectively, speech emotion signal classification ability, and platform service quality. Combined with the numerical performance in the figure, it can be seen that the Pearson's coefficient values of speech emotion signal processing ability, speech emotion signal recognition ability, speech emotion signal classification ability and platform service quality are 0.672, 0.467, 0.487, which indicates that the speech emotion signal processing ability, the speech emotion signal recognition ability, the speech emotion signal classification ability and the platform service quality have moderate positive correlation, and it is more intuitive to reveal that the passenger cabin service service quality is more important than the platform service quality, which is more important than the platform service quality, which is more important. intuitively reveals the correlation between speech emotion recognition and service quality of the cabin service dialog scene. For example, the quality of cabin service can be maximized by processing speech emotion signals, speech emotion signal recognition, and speech emotion signal classification through the service quality detection platform.

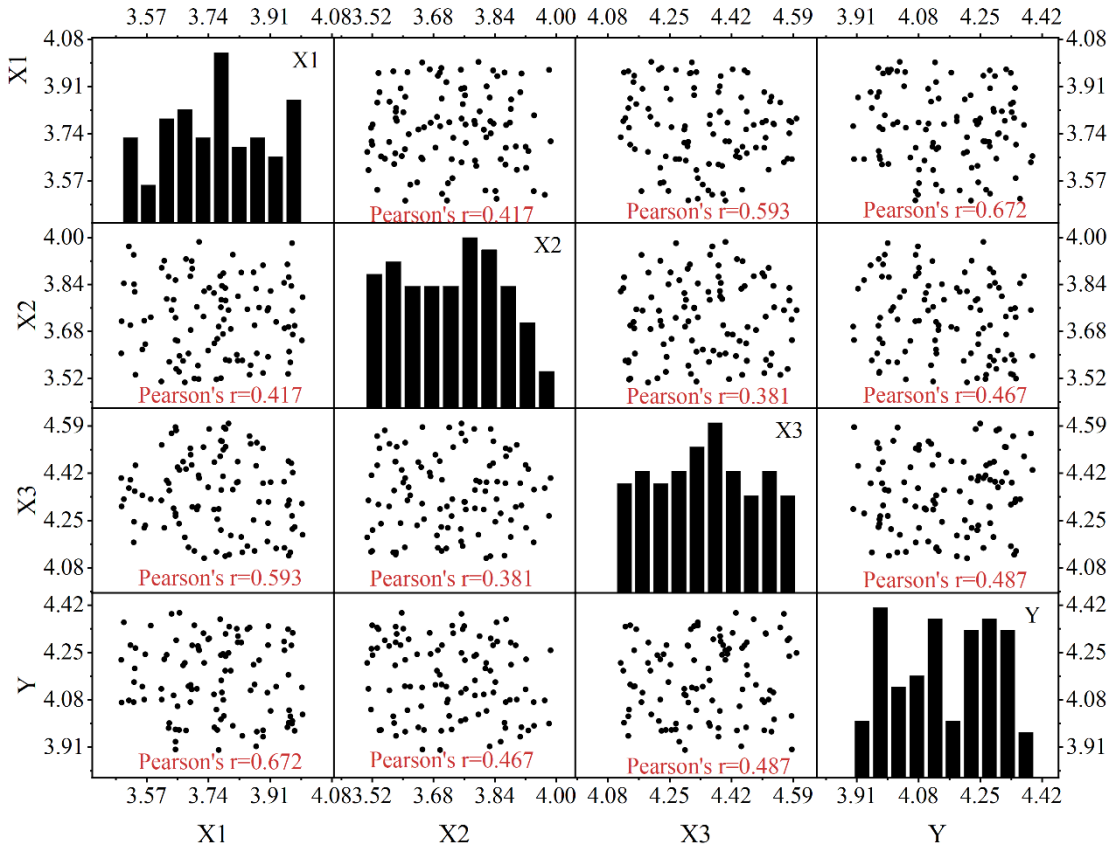


Figure 16: Results of correlation analysis

5 Conclusion

With the continuous popularization of information technology, speech recognition technology has unique advantages in cabin service conversation scenes, which in turn helps to improve the quality of cabin service. This paper combines Mel, MFCC, RePVGG, and CBAM to construct a speech emotion recognition model and design a service quality detection platform at the same time, in order to demonstrate the correlation between speech emotion recognition and service quality for cabin service dialog scenes.

(1) Through the analysis of ablation experiments, after adding CBAM, Mel features, and MFCC features to the baseline RepVGG model, the values of WA, UA, and F1 are improved, which not only indicates that Mel and MFCC features have excellent speech emotion extraction ability, but also illustrates the improvement of CBAM to the baseline RepVGG model, which fully verifies that this paper's speech emotion recognition model of this paper.

(2) Through the correlation analysis, it can be seen that the Pearson coefficient values of speech emotion signal processing ability, speech emotion signal recognition ability, speech emotion signal classification ability and platform service quality are 0.672, 0.467, 0.487, i.e., the correlation between speech emotion recognition and service quality of the cabin service dialog scene is verified.

Funding

This work was supported by Research Topic of Vocational Education Teaching Reform in 2024 by Shaanxi Vocational and Technical Education Association: Research on the Construction of

New Form Textbooks for Civil Aviation Professional English Based on the ESP Concept (Project No. 2024SZX481).

About the Author

Jianhui Wang was born in Xianyang, Shaanxi Province, P.R. China, in 1974. She received the Master's degree from Lanzhou University, P.R. China. Currently, she works in School of Culture and Tourism, Shaanxi Vocational and Technical College. Her research interests focus on Applied Linguistics, with an emphasis on ESP (English for Specific Purposes) in aviation, hotel services, and tourism.

References

[1] KAYALAR, Y., & KAYALAR, A. K. (2021). Cabin Services in Civil Aviation Sector in Turkey. Research & Reviews in Social, Human and Administrative Sciences, 241-251.

[2] Kim, Y. J. (2018). Service Failure Recovery Strategies Through Human Service Capability: A Case Study of Airline Cabin Service. Journal of Korea Society of Industrial Information Systems, 23(5), 145-157.

[3] Campese, C., Silva, T. N. R. D., Silva, L. L. G. D., Figueiredo, J. P., & Menegon, N. L. (2016). Assistive technology and passengers with special assistance needs in air transport: contributions to cabin design. Production, 26(2), 303-312.

[4] Pujangkoro, S., Wahyuni, D., & Panama, J. (2019, May). Evaluating Working Time and Work Capacity of Aircraft Cabin Line Maintenance Services. In IOP Conference Series: Materials Science and Engineering (Vol. 505, No. 1, p. 012021). IOP Publishing.

[5] Ernits, R. M., Pupkes, B., Keiser, D., Reiß, M., & Freitag, M. (2022). Inflight catering services–A comparison of central and decentral galleys inside the aircraft cabin, a concept-based approach. Transportation Research Procedia, 65, 34-43.

[6] Park, Y. S., & Park, I. S. (2017). Effects of low cost airline cabin service quality, customer satisfaction, and loyalty to Airline. Journal of the Korean Society for Aviation and Aeronautics, 25(4), 101-110.

[7] Taehui, K. I. M. (2024). The Effect of Cabin Crew Service Quality on Customer Loyalty. The Journal of Industrial Distribution & Business, 15(9), 11-19.

[8] Wang, S. M., & Park, H. Y. (2016). The effect of the cabin service quality on customer loyalty and airline image. Journal of the Korean Society for Aviation and Aeronautics, 24(2), 47-58.

[9] Rjsé, V., Jylkäs, T., & Miettinen, S. (2023). AI Enabled Airline Cabin Services: AI Augmented Services for Emotional Values. Service Design for High-Touch Solutions and Service Quality. Design Management Journal, 18(1), 100-115.

[10] Płaza, M., Kazała, R., Koruba, Z., Kozłowski, M., Lucińska, M., Sitek, K., & Spyrka, J. (2022). Emotion recognition method for call/contact centre systems. Applied Sciences,

12(21), 10951.

[11] Pupphavesa, N., & Yanasugondha, V. (2017). Problems of Thai Airways senior cabin crew toward English language communication with guests in royal first class and new business class on international flights. Unpublished independent study paper). Thammasat University, Bangkok, Language Institute.

[12] Perrodin, D. D., Liangruenrom, N., & Taworntawat, C. (2024). Navigating Intercultural Communication Challenges: Addressing Language Barriers and Foreign Language Anxiety Among Thai Low-Cost Airline Ground Staff. Trends in Psychology, 1-14.

[13] Yurtay, Y., Demirci, H., Tiryaki, H., & Altun, T. (2024). Emotion recognition on call center voice data. Applied Sciences, 14(20), 9458.

[14] Dimitrova-Grekow, T., Klis, A., & Igras-Cybulska, M. (2019). Speech emotion recognition based on voice fundamental frequency. Archives of acoustics, 277-286.

[15] Tsiourti, C., Weiss, A., Wac, K., & Vincze, M. (2019). Multimodal integration of emotional signals from voice, body, and context: Effects of (in) congruence on emotion recognition and attitudes towards robots. International Journal of Social Robotics, 11(4), 555-573.

[16] Wani, T. M., Gunawan, T. S., Qadri, S. A. A., Kartiwi, M., & Ambikairajah, E. (2021). A comprehensive review of speech emotion recognition systems. IEEE access, 9, 47795-47814.

[17] Huang, Z., & Epps, J. (2018). Prediction of emotion change from speech. Frontiers in ICT, 5, 11.