Journal of Network Intelligence ©2025 ISSN 2414-8105
Taiwan Ubiquitous Information Volume **10**, Number **3**, August **2025**

Attentional Convolutional Network-based Emotional Feature Classification in Vehicular Interaction

En-Lin Xie[1,*], Cai Zhao[2]

[1]Xiangsihu College, Guangxi Minzu University, Nanning 530000, P. R. China
foolishxel@163.com

[2]Lee Kong Chian Faculty of Engineering and Science,
Tunku Abdul Rahman University, Kuala Lumpur 31900, Malaysia
mq5749@163.com

[*]Corresponding author: En-Lin Xie

Received February 8, 2024, revised June 25, 2024, accepted September 29, 2024.

Abstract. *The application of machine learning sentiment analysis in in-vehicle interaction system can realize more humanized and intelligent services, enhance driving experience and safety, and is also one of the important trends in the development of future intelligent vehicles. However, the accuracy of sentiment feature classification in real in-vehicle environments still fails to meet the practical needs, mainly due to noise interference and insufficient model complexity. Therefore, this work proposes an emotion feature classification method based on improved attention convolutional network. Firstly, the speech signal was pre-processed in a discrete model using a sampling frequency of 16 KHz with frames and windows. Then, three commonly used emotion features, namely rhythmic features, spectral features and phonological features, were employed in the feature extraction stage. Next, the typical deep convolutional neural network was improved in two aspects, including (1) the inclusion of a data enhancement strategy and (2) the inclusion of a SE attention mechanism module. Finally, the improved deep convolutional neural network model is applied to extract deep information about speech emotion from the combination of acoustic features. The experimental results on the CASIA speech emotion database show that the accuracy of the deep convolutional neural network model is significantly improved by adding the data enhancement strategy and the SE attention mechanism, and reaches a maximum of 91.32 % at 0 dB.*

Keywords: Deep convolutional neural network; attention mechanism; feature classification; in-vehicle interaction; data enhancement

1. **Introduction.** With the rapid development of artificial intelligence and machine learning technology, in-vehicle interaction systems are becoming more and more intelligent. Modern in-vehicle interaction systems are usually equipped with voice recognition functions, which can realize various operations, such as navigation, telephone, music playback, etc., through voice commands [1, 2]. The development of this technology enables drivers to interact with their vehicles in a safer and more convenient way. Some in-vehicle interaction systems have been able to understand the driver's natural language commands and respond accordingly. This allows drivers to communicate with their vehicles in a more natural way without having to learn complex specific commands [3].

Modern in-vehicle interaction systems usually allow drivers to personalize their configuration according to their preferences and needs [4]. For example, seat adjustments, music playlists, navigation settings, etc. can be customized to meet different drivers' preferences.

Some in-vehicle interaction systems are equipped with driver monitoring functions, which can detect the driver's emotional state, fatigue level, and attention level through devices such as sensors and cameras [5, 6]. Based on this data, the system can adjust the driving mode, provide relevant suggestions, or remind the driver to take a break in order to provide a more personalized driving experience. Emotion analysis can make the in-vehicle interaction system more intelligent, for example, when recognizing the human's happy or excited emotions, the system can actively change the atmosphere of music and lighting to create a better driving experience for humans [7]. By analyzing the user's emotional state, the in-vehicle interaction system can provide personalized services and suggestions for the user according to different emotional states. For example, when the in-vehicle system can recognize human fatigue, it can provide relevant rest site information or suggest the driver to take a rest.

Currently, recognizing and understanding drivers' emotional states through machine learning techniques is a hot research topic in related fields. Researchers are exploring how to utilize multiple data sources such as speech, facial expression, physiological sensors, etc. to infer the driver's emotional state and provide corresponding driving support based on this information [8]. However, the classification accuracy of emotional features in real in-vehicle environments still fails to meet the practical needs, mainly due to noise interference and insufficient model complexity [9]. In dealing with these problems, targeted audio pre-processing techniques are needed to minimize the effect of noise and more complex models are considered to improve the accuracy. Therefore, the aim of this study is to improve Deep Convolutional Network (DCNN) models using data augmentation techniques and attention mechanisms, and to use them to mine rich features in in-vehicle interactive speech data to improve the accuracy of sentiment feature classification. Deep Convolutional Networks can effectively learn and extract high-level features in speech signals, while the attention mechanism allows the system to be more focused and focused on important parts of the speech data related to emotions and improve the recognition accuracy. Data enhancement techniques can enhance the clarity of speech signals by performing audio processing on noisy speech.

1.1. **Related work.** Machine learning-based classification of emotional features in in-vehicle interaction speech is a relatively new research area. It aims to classify and identify emotional features in in-vehicle interaction speech through machine learning algorithms to achieve a more intelligent and personalized in-vehicle interaction experience. However, speech emotion feature classification is a relatively mature research field.

Shami and Verhelst [10] used a variety of machine learning algorithms for speech emotion feature classification, such as Support Vector Machine (SVM), Multilayer Perceptron (MLP) and k-Nearest Neighbor (k-NN). By tuning and comparing the performance of different algorithms, the most suitable algorithm for emotion recognition is found and guidance for comparative study and method selection is provided. Zheng et al. [11] proposed a method using deep residual learning for speech emotion recognition. By introducing a deep neural network (ResNet) with residual connectivity, a deeper model was successfully constructed to better capture the emotional features in speech signals. The method was experimentally demonstrated on several speech emotion datasets, and deep residual learning was able to achieve better classification performance compared to traditional machine learning methods. Chen et al. [12] proposed a speech emotion recognition model based on the attention mechanism. By introducing the attention mechanism, the model can automatically focus on some features related to emotion recognition, thus improving the classification performance. Experimental results on multiple datasets show that the model is able to better capture the emotion information in speech signals compared to traditional feature extraction methods and classifiers.

Data enhancement plays an important role in sentiment feature classification. In sentiment feature classification, data augmentation refers to the generation of more diverse and richer datasets by transforming, expanding, and synthesizing the original data in order to improve the performance and robustness of the sentiment classification model. Mustaqeem and Kwon [13] proposed an approach combining data enhancement and CNN for speech emotion recognition. Two data enhancement techniques were used: acoustic perturbation and temporal perturbation. Acoustic perturbation consists of adding white noise, low-pass filtering and high-pass filtering to simulate various noise situations in a real environment. Temporal perturbation, on the other hand, adds variety to the data by randomly speeding up or slowing down the playback of the speech signal. Experimental results show that the use of data augmentation can significantly improve the performance of emotion recognition, especially in noisy environments. Badshah et al. [14] explored an alternative approach to using data augmentation in emotion recognition, based on CNNs and spectrograms. The researchers used a variety of data enhancement techniques including level flipping of spectrograms, random cropping and smoothing of spectrograms. They also tried different combinations of data enhancement strategies and compared their performance on emotion recognition tasks. The experimental results show that data enhancement can significantly improve the robustness of the model and the accuracy of sentiment recognition.

1.2. Motivation and contribution.

In real driving scenarios, speech signals are often interfered by various environmental noises, which may lead to degradation of acoustic feature extraction performance. Data enhancement can simulate different noise situations, such as white noise, low-pass filtering, high-pass filtering, etc., so that the model can better adapt to the noisy environment during the training phase [15, ?]. In addition, the emotional information in speech signals may have long-range inter-channel dependencies, which may not be adequately captured by traditional DCNNs [16]. The Squeeze-and-Excitation (SE) attention mechanism is capable of adaptively adjusting the relationship between feature channels by introducing a gating mechanism [17]. Therefore, this work proposes an improved DCNN based on data augmentation and SE attention mechanism for human interaction emotion classification in in-vehicle environment. The main innovations and contributions of this work include:

(1) Because of the presence of noisy audio signals, channel clutter, and poor results during the acquisition of speech in the in-vehicle environment, acoustic feature extraction encounters great difficulties. Therefore, a data enhancement method based on spatial impact response is proposed to reduce the interference encountered in the acoustic feature extraction model.

(2) There are usually some problems in traditional DCNN, such as insufficient dependence on inter-channel dependence and insufficient consideration of feature importance. In order to solve these problems, this paper introduces the SE module to improve the network performance of CNNs, and enhance the feature expression and generalisation ability.

2. Preparations for the Classification of Human Emotional Characteristics.

2.1. Speech signal preprocessing.

Discrete and dimensional models are two common models used to describe and categorize human emotional traits. Discrete models categorize emotional traits into a discrete set of emotional categories. These categories are usually the basic dimensions of emotions, such as joy, anger, sadness, and fear. A common example of a discrete model is an emotion model based on six basic emotions (Ekman's Six Basic Emotions Model), which includes joy, anger, sadness, fear, disgust, and surprise. Dimensional models represent human emotional traits as continuous values on multiple dimensions, and use these values to describe and categorize different emotional expressions. Commonly used dimensional models include Pleasure-Arousal-Dominance (PAD) model and Valence-Arousal-Dominance (VAD) model. Among them, Valence (describes the degree of positivity and negativity of the emotion, Arousal describes the intensity of the emotion, and Dominance describes the degree of control over the emotion. This model is more suitable for a more detailed and comprehensive description and categorization of emotions.

Overall, discrete and dimensional models are two important directions for categorizing human emotion features. The discrete model provides simple and intuitive emotion categories, while the dimensional model provides richer and more precise emotion descriptions. In contrast, the discrete emotion model is simple and intuitive, and the emotions used are common emotions in daily life, which are equally applicable to human emotional states. Therefore, the discrete sentiment model is adopted in this work.

In order to obtain feature parameters with high emotional representativeness for subsequent emotion classification, we first need to preprocess the original signal, which mainly includes noise reduction, pre-emphasis, and frame-wise windowing. Because the frequency range of the language is 300 hz~3.4 Khz, according to Nyquist's theorem, the sampling frequency needs to be greater than two times of the maximum frequency value. Therefore, 8 Khz sampling frequency is usually sampled. In order to have a better language effect, the sampling frequency of 16 Khz is used in this paper. The denoising method used is to design the response filter to reduce the noise involved in the speech signal.

Speech signals due to the innate conditions of human pronunciation decided that the low-frequency band energy is greater than the high-frequency band energy, and pre-emphasis is the high-frequency portion of the signal before transmission is aggravated, and then de-emphasis at the receiving end, so as to make the entire information transmission of the signal-to-noise ratio has been improved to improve the quality of the signal transmission. Pre-emphasis is the addition of a first-order high-pass filter to increase the high-frequency part, as shown below:

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

where μ represents the pre-emphasis coefficient, generally takes the value between 0.9~1, this time let μ is 0.937. If the original language signal is $s(n)$, after the pre-emphasis processing the expression is.

$$s'(n) = s(n) - \mu s(n-1) \tag{2}$$

The speech after pre-emphasis has two obvious advantages. One is that the high spectral value caused by the sound gate pulse can be weakened by adding a zero point to finally achieve the trend of speech spectral flattening, and the characteristic parameters obtained are more in line with the original channel model; the second is that it belongs to the high-pass filter, after the speech passes through it, so that the amplitude of the high-frequency part of the amplitude is improved, and the amplitude of the low-frequency part of the amplitude is suppressed to reduce the range of the spectral dynamic changes.

In the frame-plus window, the human ear is not as accurate as the human vocal organs, although the sound emitted by the human vocal organs is constantly changing from high to low. In a very short period of time window interval, the speech signal has a short-time smoothness, and the speech signal within 10~30 ms can be regarded as unchanged. Generally used window functions are Haining window, Hamming window and rectangular window. The Hamming window is chosen for the experiment because it has smoother low-pass characteristics and matches the frequency characteristics of the speech signal. The function of Hamming window is shown below:

$$w(n) = \begin{cases} 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right), & 0 \leq n \leq N-1 \\ 0, & \text{others} \end{cases} \tag{3}$$

where $\alpha = 0.46$ was chosen for this work and N is the window length.

2.2. **Feature extraction.** The features for emotion recognition are divided into acoustic and non-acoustic features. For the driver's voice and the easy extraction of speech features, all acoustic features are used in this work. Three commonly used affective features, namely rhythmic features, spectral features and phonological features, are used in the feature extraction stage to form the feature set required for subsequent emotion recognition.

2.2.1. *Rhythmic features.* Rhythmic features are also referred to as suprasegmental features because the frame durations for speech analysis are large, typically 30–100 ms. The fundamental frequency is extracted by first detecting the fundamental period, and then using an autocorrelation algorithm to obtain the vibrational pattern of the vocal folds in the language.

$$R(k) = \sum_{n=0}^{N-n-1} x(n)x(n+k) \tag{4}$$

where N is the number of samples in each frame and k is the time interval. The fundamental period and the fundamental frequency are inversely related to each other, the fundamental period is selected as the maximum value of $R(k)$, and the fundamental frequency is obtained by taking the reciprocal.

The other short-time energy can also reflect the change of emotion, which indicates the amount of energy intensity of the speech signal within a frame.

$$E = \sum_{0}^{N-1} x^2(n) \tag{5}$$

$$Z = \frac{1}{2}\sum_{n=1}^{N-1} |sgn[x(n)] - sgn[x(n-1)]| \tag{6}$$

where N is the frame length and $sgn[\cdot]$ represents the sign function.

$$sgn[x] = \begin{cases} 1 & x > 0 \\ 0 & x = 0 \\ -1 & x < 0 \end{cases} \tag{7}$$

2.2.2. *Spectral features.* In speech emotion classification, spectral features refer to features extracted from the frequency spectrum of speech signals, which are used to describe the energy distribution of speech signals in frequency domain. Spectral characteristics are usually obtained by Fourier transform of speech signals. The cepstrum feature used in this work is Mel Frequency Cepstrum Coefficient (MFCC). Assume that the input speech is signal $x(n)$. The MFCC parameters are obtained by taking the logarithm of the energy of the Mel filter and then calculating the Discrete Cosine Transform (DCT).

$$mfcc(i,n) = \sqrt{\frac{2}{M}}\sum_{m=0}^{M-1} \log[S(i,m)]\cos\left(\frac{\pi n(2m-1)}{2M}\right) \tag{8}$$

$$S(i,m)=\sum_{k=0}^{N-1}E(i,k)H_m(k),\ 0\le m\le M \tag{9}$$

where $H_m(k)$ represents the bandpass filter, M represents the number of triangular filters, m represents the m-th Mel filter, and n is the spectrum after DCT.

2.2.3. *Speech quality features.* Sound quality features refer to the features related to sound texture, timbre, tone and so on. These features are very important for identifying emotional expressions in speech. Sound quality features mainly focus on the attributes related to sound quality in time domain or frequency domain. As air is expelled through the vocal folds, the contraction between the vocal folds varies, thus creating emotion in the source signal. There are correlations between different types of vocalisations and emotion. The calculation of the Relative Average Perturbation (RAP) of the fundamental cycle is shown below:

$$RAP=\frac{1/(N-2)\sum_{i=2}^{N-1}\left|\frac{T_{i-1}+T_i+T_{i+1}}{3}-T_i\right|}{1/N\sum_{i=1}^{N}T_i} \tag{10}$$

where T_i represents the period of the signal in the i-th pronounced cycle.

3. Improved DCNN-Based Sentiment Feature Classification in Vehicular Interaction.

3.1. Analysis of the noise environment inside the car.

In-vehicle noise environment analysis in in-vehicle interaction systems is a process performed to evaluate and improve the in-vehicle interaction experience. In-vehicle noise may negatively affect speech recognition, auditory feedback, and voice interaction for drivers and passengers. Therefore, by analyzing the in-vehicle noise environment, measures can be taken to reduce the interference of noise on the in-vehicle interaction system. Spectral and time domain analysis of in-vehicle noise is required to understand the frequency distribution, intensity and time-varying characteristics of the noise. This can help determine which frequency ranges of noise have the greatest impact on the performance of the in-vehicle interaction system. For example, in a speech recognition task, low-frequency noise may interfere with the clarity and accuracy of the speech signal.

Low-frequency noise in a vehicle usually refers to sounds with frequencies between 20 Hz and 200 Hz. This noise comes primarily from vibrations and resonances in the engine, exhaust system, suspension, tires, and road surfaces. Speech captured in the vehicle environment is not pure speech. Generally speaking, people normally communicate at around 60 to 70 dB, which are measured in SPL (Sound Pressure Level).

$$L_p=20\log\left(\frac{P}{P_0}\right) \tag{11}$$

where P_0 is the reference sound pressure, the size of 20 micropascals, is the lowest sound pressure at a frequency of 1000 Hz audible to the human ear.

In this paper, the signal-to-noise ratio is used to analyse how much the driver's speech signal is affected by environmental noise. The signal-to-noise ratio expression is.

$$SNR=10\lg\frac{\sum x^2(n)}{\sum r^2(n)}=10\lg\frac{W_s}{W_r} \tag{12}$$

where $\sum x^2(n)$ is the clean driver speech signal energy; $\sum r^2(n)$ is the ambient noise signal energy inside the vehicle; and W_s and W_r represent their respective average power.

The sound power W can also represent the magnitude of the sound, which can be ignored in the surroundings. When the shape of the source is considered as a point, the sound pressure level L_p and power level L_w are transformed as follows.

$$L_p(r) = L_w - 10\lg\left(4\pi r^2\right) \tag{13}$$

where r is the straight line distance between the sound source and the measurement point. The power level L_w expression is.

$$L_w = 10\lg\left(\frac{W}{W_0}\right) \tag{14}$$

where W and W_0 represent sound power and reference sound power, respectively.

The baseline sound power is $10^{-12}W$. 60 dB sound pressure level and 50 dB ambient noise are chosen to calculate the SNR of 10 dB. Selecting 60 dB SPL and 50 dB ambient noise, we can get the SNR of 10 dB under normal driving conditions, and considering the influence of other special noises and to ensure the effectiveness of the speech emotion recognition system, this paper selects the SNR of -10∼30 dB for the study.

3.2. **Improved DCNN.**

3.2.1. *Data enhancement (DE) strategy.* Because of the noisy audio signals, channel clutter and poor results in the process of capturing speech in the in-vehicle environment, it makes the acoustic feature extraction encounter great difficulties. Therefore, this paper proposes a data enhancement method based on spatial impact response to reduce the interference encountered in the acoustic feature extraction model.

Firstly, time domain enhancement is performed, changing the audio samples by transforming the length and speed. In terms of volume, the entire signal is directly multiplied by some constant. In terms of sampling rate, the signal is resampled, e.g., a 16,000 Hz signal is resampled to 8,000 Hz to simulate the voice signal of a wired telephone. (1) In terms of speech rate, the rate of the audio is multiplied by some constant near 1, e.g., 0.9, 1.1. (2) In terms of pitch, the fundamental frequency of the audio is randomly offset. (3) In terms of encoding format, take the audio encoded in linear pulses, re-encode it to MP3 format, and then convert it to WAV format.

Then, data augmentation is used to generate the data. An in-vehicle space is described using a set of parameters, which can be described in terms of length, width and height since the in-vehicle space is mostly cubic. Determine the location of the signal source and the location of the recording device, which can be described in terms of their stereo coordinates in the in-vehicle space. The Room Impulse Response (RIR) is used to describe the reception of the signal source by the recording device in the in-vehicle space, including: the signal source travelling in a straight line and reaching the recording device; the signal source travelling in the other direction and reflecting back to the recording device. The stereo coordinates are shown in Figure 1.

The mathematical expressions for background noise and spatial reverberation are shown below:

$$x_r(t) = x(t) * h_s(t) + \sum_{i=1}^{M} n_i(t) * h_i(t) + d(t) \tag{15}$$

where $x(t)$ is the source, $h_s(t)$ is the corresponding RIR, M is the number of noise sources, $n_i(t)$ is the noise source, $h_i(t)$ is the RIR corresponding to the noise source, $d(t)$ is an additive noise source, $x_r(t)$ is the augmented signal notation, and $*$ denotes the convolution operation.

According to Equation (15), we can collect speech signal, noise signal and RIR separately to get various combinations, and adjust the signal-to-noise ratio by weighting.

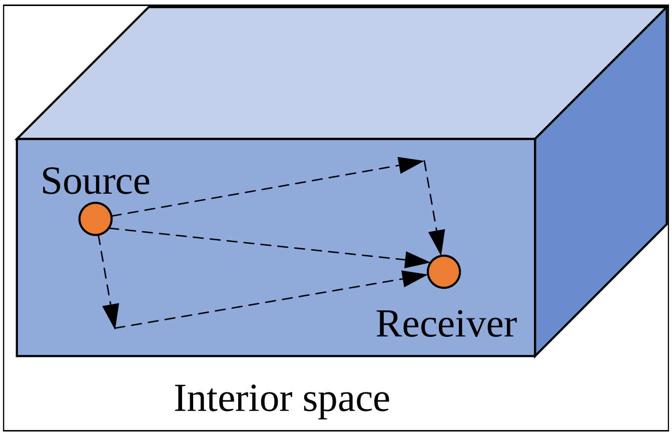

Figure 1. Stereogram

During training, we need to consider which part of the data to input into the model, for the driver speech recognition in car environment discussed in this thesis, the data need to pass through the VAD model first to get the speech segments. In order to reduce the dependence of the speech recognition model on a specific KWS/VAD model, we can then use data enhancement methods for detecting perturbations, including: randomly offsetting the output of the KWS/VAD model and thus randomly perturbing the threshold of the KWS/VAD model.

Next, frequency domain enhancement is used, where the frequency domain is enhanced by changing frequencies, adding background noise, etc. All the values in a specific frequency range in the original time-frequency spectrogram are replaced with the mean value of the original time-frequency spectrogram.

3.2.2. *Incorporating SE attention mechanism.* There are usually some problems in traditional DCNN, such as insufficient inter-channel dependency and insufficient consideration of feature importance. In order to solve these problems, this paper introduces the Squeeze-and-Excitation (SE) module [18, 19] to improve the network performance of DCNNs, and enhance the expressive and generalisation abilities of features. The traditional design of DCNNs ignores the correlation between different channels in the features as well as the contribution of each channel to the learnt features.

The SE module can improve the network's ability to perceive inter-channel dependencies by modelling the correlations between channels, leading to better extraction and expression of features. The SE module introduces an adaptive attention mechanism that learns the weights of each channel and infers the activation level of each channel based on the globally described feature vectors. This allows the network to better focus on important features and suppress irrelevant features, thus improving the accuracy and expressiveness of feature representation. The SE module enables the network to automatically learn the importance of each channel, enhancing the representation of useful features and suppressing irrelevant information, thus improving the flexibility and accuracy of feature representation. The SE module is used to normalise the raw inputs, thus enhancing the learning ability of the model. This is a typical structure consisting of a global pooling layer, activation functions and adjustable weights. Thus, the SE module helps to allow the network model to utilise some additional auxiliary networks to extract information efficiently. The SE module design is shown in Figure 2.

SE is an attention mechanism for enhancing the performance of neural networks. It is proposed to solve the problem of insufficient inter-channel dependency of traditional

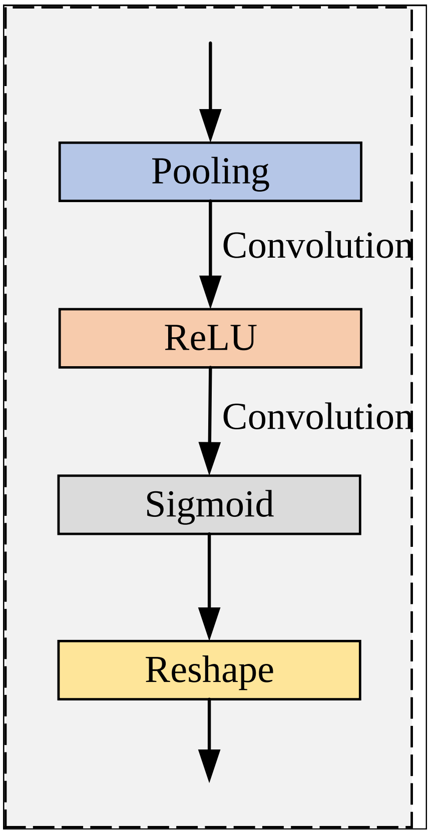

Figure 2. SE module design drawing

DCNN in extracting features. The SE module consists of two main steps [20]: Squeeze and Excitation.

(1) Squeeze. compresses the channel information of each feature map through a global average pooling (GAP) operation, which integrates the features of each channel to obtain a globally described feature vector. This step allows the network to quantify and summarise the importance of each channel [21, 22].

(2) Excitation. learns the weights of each channel by introducing a fully connected layer and an activation function. The activation level of each channel is inferred from the globally described feature vectors. These channel weights are used to re-weight the features in each channel, thus enhancing useful features and suppressing irrelevant ones.

(3) Scale: the normalised weights obtained earlier are weighted to the features of each channel.

Squeeze and Excitation are two very critical operations, and the structure of the SE module is shown schematically in Figure 3.

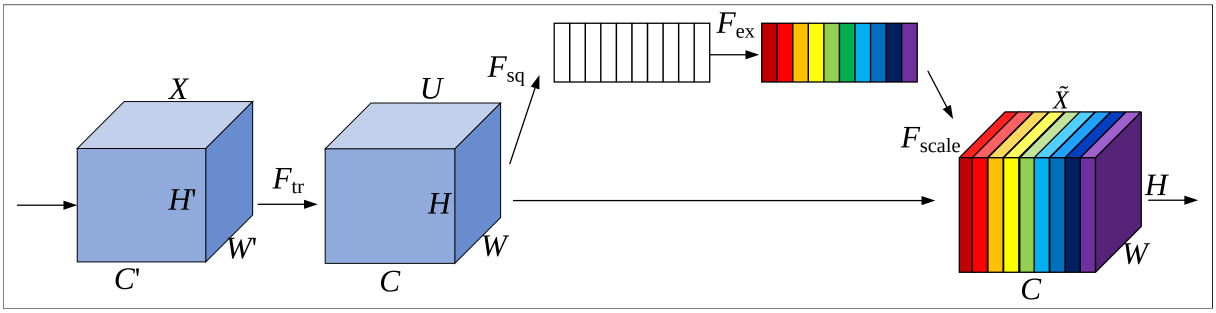

Figure 3. Structure diagram of the SE module

The input is x, whose feature channel number is C', and a feature with feature channel number C is obtained after a series of convolution operations. Firstly, F_{tr} is a standard convolution operation [23], defined as shown below:

$$F_{tr} : X \to U, X \in \mathbb{R}^{H' \times W' \times C'}, U \in \mathbb{R}^{H \times W \times C} \tag{16}$$

$$u_c = v_c * X = \sum_{s=1}^{C'} v_c^s * x^s \tag{17}$$

where V_c represents the cth convolution kernel and X^s represents the sth input. The U obtained from F_{tr} is called the tensor, while u_c represents the cth 2D matrix in U, and the subscript c represents the channel.

Next, there is the Squeeze operation, which is calculated as follows.

$$z_c = F_{sq}(u_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} u_c(i, j) \tag{18}$$

The input of $H \times W \times C$ is converted into the output of $1 \times 1 \times C$, which is equivalent to the process of calculating the numerical distribution of the C feature maps of the layer, or global information [24], corresponding to F_{sq} in Figure 3.

Then comes the Excitation operation, which is calculated as follows.

$$s = F_{ex}(z, W) = \sigma(g(z, W)) = \sigma\left(W_2 \delta(W_1 z)\right) \tag{19}$$

Firstly, multiplying W_1 by z (the result obtained by squeeze earlier is z), is a fully connected layer operation. The dimension of W_1 is $C/r \times C$, where r is a scaling parameter, which in this paper takes the value of 16. The purpose of using the scaling parameter is to reduce the number of channels and thus reduce the computational effort. Since the dimension of the scaling parameter z is $1 \times 1 \times C$, the result of $W_1 z$ is $1 \times 1 \times C/r$.

Then, after a ReLU layer, the output has the same dimension and is multiplied with W_2. Note that the multiplication with W_2 is also a fully connected layer process. The dimension of W_2 is $C \times C/r$, so the dimension of the output is $1 \times 1 \times C$. Finally, a sigmoid function yields s, which corresponds to F_{ex} in Figure 3.

Once s is obtained, the original tensor U can be manipulated as follows:

$$\widetilde{x}_c = F_{scale}(u_c, s_c) = s_c \cdot u_c \tag{20}$$

where u_c is a two-dimensional matrix and s_c is a weight, which is equivalent to multiplying each value in the matrix u_c by s_c, corresponding to F_{scale} in Figure 3.

SE embeds the CNN network using global pooling as a Squeeze operation. The Bottleneck structure effectively reduces the number of parameters and computational complexity in the model, and ultimately outputs the same weights as the input features. The feature dimension is reduced to $1/r$ of the input. This is followed by a ReLU activation function, and then returned to the previous dimensions through a fully connected layer [25].

3.3. **DE-SE-DCNN Based Speech Emotion Recognition.** In this paper, the proposed improved DCNN is referred to as DE-SE-DCNN, which mainly consists of the two improvements mentioned above; (1) the inclusion of the Data Enhancement (DE) strategy and (2) the inclusion of the SE Attention Mechanism module.

DE-SE-DCNN was applied to extract the deep information of speech emotion from the acoustic feature combinations. The selected acoustic feature combinations are shown in Table 1, including rhythmic features, spectral features (MFCC and Delta_MFCC) and tonal features. The total dimension of the acoustic feature combinations is 39. The sample speech is cut into equal length 3-second speech segments in the preprocessing stage to ensure that the input acoustic feature combinations are of the same size, and the shortfall is filled with zeros. The number of samples per frame was 512, the size of the frame was 256 × 256, and the formant was hamming.

Table 1. Combination of acoustic features

Feature name	**Dimension**
Short-term energy	1
Short-time past zero rate	1
Gene frequency	1
MFCC	12
Delt MFCC	12
RAP	12

Firstly, the time-series features are input into the DE-SE-DCNN network as training parameters for the experimental subjects. Here the temporal features are in the form of a set of two-dimensional vectors, with the horizontal coordinates representing the time series and the vertical coordinates being the dimensions. The experimental model has a total of 3 convolutional layers. The output of the convolution operation is then normalised using the BN function and activated with the ReLU function, followed by connecting a pooling layer and a Dropout layer. The first fully connected layer output is processed by the BN function and the ReLU function, and the second fully connected layer is followed by classification using the softmax function.

Since the order of magnitude varies greatly between the original features, this requires centring and normalisation. Letting each feature x transformed to obey a normal distribution eliminates errors caused by different orders of magnitude of features.

$$x_i' = \frac{x_i - mean(x_i)}{var(x_i)} \tag{21}$$

When the acoustic features are input, the pooling operation after the convolution of the last two layers is performed only in the time dimension in order to ensure the resolution in the feature dimension. The loss function can reflect how well the model is trained, and the smaller the loss function, the closer the predicted value is to the true value. The cross-entropy loss function is used to speed up the weight update in the gradient descent algorithm.

$$CE = -\frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{n} p\left(x_{ij}\right)\log\left(q\left(x_{ij}\right)\right) \tag{22}$$

where m is the number of samples in a batch, n is the number of output categories, and $q(x_{ij})$ is the probability of observation of sample i on category j.

When the sample i includes the category j, $p(x_{ij}) = 1$, otherwise it is 0. The training process is optimised with Adam's algorithm, the Epoch is set to 400, the initial learning rate is 0.001, and the learning rate is adjusted by using the cosine decay function [26].

4. Experimental results and analyses.

4.1. Experimental environment and dataset.

In the experimental phase 1200 speech samples from the CASIA linguistic affective database were used, consisting of six emotions: angry, happy, fear, sad, surprise and neutral. CASIA was recorded in a pristine, noise-free studio and contains a total of 9600 different articulatory utterances for two males and two females. This corpus can be analysed to compare different emotional states expressed by acoustic and rhythmic features.

The recordings were made with a sampling rate of 16k Hz, a quantisation resolution of 16bit and a storage format of pcm. The training and test sets were divided in a 4:1 ratio, with 960 pure language samples for training and the remaining 240 for testing. In order to explore the results in different noise environments, different background noises are added to the test set at signal-to-noise ratios of 30 dB, 20 dB, 10 dB, 0 dB, and -10 dB. The training feature parameters are extracted 20-dimensional speech emotion features that are dimensionally reduced using the PCA algorithm.

4.2. **Effect of DE module on accuracy.** In a pure and noise-free environment, the combination of dimensionally reduced acoustic features is used as input features for DE-SE-DCNN for speech emotion recognition, and the resulting confusion matrix is shown in Figure 4. It can be seen that the overall accuracy is 91.32% when the input is acoustic features. The proposed model has better recognition results for Suprise and Happy emotion classification, while Fear and sad emotions are most easily confused.

The effect of changing the strategy on the trained-out model is observed by comparing the accuracy data of different data enhancement strategies under the same model. The results show that data enhancement has a role in improving the accuracy of the model. Compared with the SE-DCNN model without data enhancement, the accuracy of the DE-SE-DCNN model increases from 78.61% to 83.58%. The main reason is that the spatial shock response function makes the noise reduction process more delicate, enhances the readability of the data, and allows the model to be trained better, resulting in a better-quality neural network model.

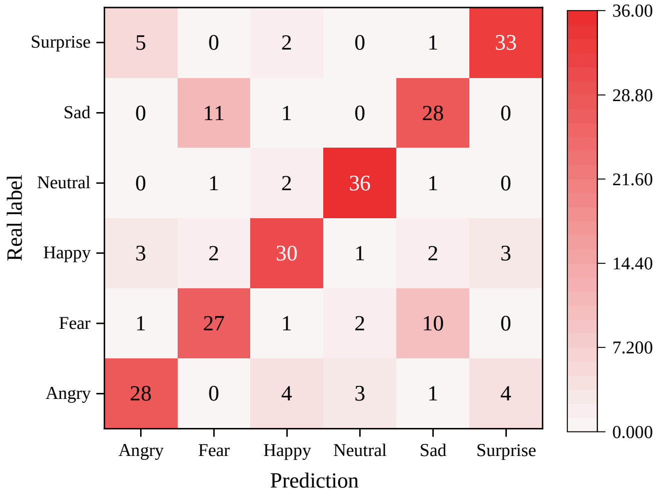

Figure 4. Confusion matrix for acoustic feature combinations

4.3. **Performance comparison of different recognition models.** In order to explore the performance of different models in noise environment (Gaussian white noise), the accuracy of emotion classification of four models (LSTM, DCNN, DCNN+Bi-LSTM, DE-SE-DCNN) under different noise levels is compared, and the results are shown in Table 2.

It can be seen that the noise level keeps increasing from no noise to 30 dB. As the noise increases, the accuracy of all models generally decreases. Under most of the noise levels, the accuracy of DE-SE-DCNN model is the highest and that of LSTM model is the lowest. The recognition rate of DE-SE-DCNN model reaches the highest 91.32% at 0 dB. When the noise reaches 30 dB, the accuracy of all models decreases very seriously, indicating that the strong noise environment is a great challenge for emotion classification. The

Table 2. Recognition rate of different models under white noise/%

Noises	∞	30 dB	20 dB	10 dB	0 dB	-10 dB
LSTM	74.45	76.55	75.96	78.89	83.38	52.25
DCNN	78.67	80.05	80.58	82.21	88.23	55.88
DCNN+Bi-LSTM	82.31	81.23	85.59	87.93	90.63	60.43
DE-SE-DCNN	83.58	84.66	87.19	89.67	91.32	62.67

overall performance of the two fusion models, DCNN+Bi-LSTM and DE-SE-DCNN, is better than that of the single LSTM and DCNN models, which indicates that the model fusion is an effective way to improve the robustness.

5. **Conclusions.** This work proposes an improved DCNN (DE-SE-DCNN) based on data enhancement and SE attention mechanism to realize human interaction emotion classification in in-vehicle environment. Firstly, because the process of capturing speech in in-vehicle environments is characterised by noisy audio signals, channel heterogeneity, and poor results, which makes the acoustic feature extraction work encounter great difficulties. Therefore, a data enhancement method based on spatial impact response is proposed to reduce the interference encountered in the acoustic feature extraction model. Secondly, there are usually some problems in traditional DCNN, such as insufficient dependence on inter-channel dependence and insufficient consideration of feature importance. In order to solve these problems, this paper introduces the SE module to improve the network performance of DCNN and enhance the feature representation and generalisation ability. Compared with the model without data enhancement, the accuracy of DE-SE-DCNN model is improved from 78.61% to 83.58%. The DE-SE-DCNN model has the best improvement in emotion recognition rate when the signal-to-noise ratio is higher than 0 dB.

Acknowledgment. This work is supported by the Project of China-ASEAN Institute of Digital Humanities Exchange in 2023 (No. CADP-YJY202309).

REFERENCES

[1] Y. Du, C. Liu, D. Wu, and S. Li, "Application of vehicle mounted accelerometers to measure pavement roughness," *International Journal of Distributed Sensor Networks*, vol. 12, no. 6, 8413146, 2016.

[2] Z. Gao, and Y. Yu, "Interacting multiple model for improving the precision of vehicle-mounted global position system," *Computers & Electrical Engineering*, vol. 51, pp. 370-375, 2016.

[3] J. Liu, X. Fu, A. Hainen, C. Yang, L. Villavicencio, and W. J. Horrey, "Evaluating the impacts of vehicle-mounted Variable Message Signs on passing vehicles: implications for protecting roadside incident and service personnel," *Journal of Intelligent Transportation Systems*, pp. 1-21, 2023.

[4] P. K. Murali, M. Kaboli, and R. Dahiya, "Intelligent In-Vehicle Interaction Technologies," *Advanced Intelligent Systems*, vol. 4, no. 2, pp. 2100122, 2022.

[5] L. A. Schmitt, "Advanced vehicle command and control system (AVCCS)," *Journal of Transportation Engineering*, vol. 116, no. 4, pp. 407-416, 1990.

[6] C. Spelta, V. Manzoni, A. Corti, A. Goggi, and S. M. Savaresi, "Smartphone-based vehicle-to-driver/environment interaction system for motorcycles," *IEEE Embedded Systems Letters*, vol. 2, no. 2, pp. 39-42, 2010.

[7] S. A. Valitzski, G. J. D'ANGELO, G. R. Gallagher, D. A. Osborn, K. V. Miller, and R. J. Warren, "Deer responses to sounds from a vehicle-mounted sound-production system," *The Journal of Wildlife Management*, vol. 73, no. 7, pp. 1072-1076, 2009.

[8] G. Yan, L. Zhang, C. Zheng, M. Zhang, K. Zheng, F. Song, W. Ye, Y. Zhang, Y. Wang, and F. K. Tittel, "Mobile Vehicle Measurement of Urban Atmospheric CH_4/C_2H_6 Using a Midinfrared Dual-Gas Sensor System Based on Interband Cascade Laser Absorption Spectroscopy," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1-11, 2022.

[9] C. Zhang, "Exploring user cognition difference and pleasure balance guidance method for product perceptible features in vehicle-mounted system," *International Journal of System Assurance Engineering and Management*, vol. 13, no. Suppl 3, pp. 1019-1030, 2022.

[10] M. Shami, and W. Verhelst, "An evaluation of the robustness of existing supervised machine learning approaches to the classification of emotions in speech," *Speech Communication*, vol. 49, no. 3, pp. 201-212, 2007.

[11] K. Zheng, Z. Xia, Y. Zhang, X. Xu, and Y. Fu, "Speech Emotion Recognition based on Multi-Level Residual Convolutional Neural Networks," *Engineering Letters*, vol. 28, no. 2, pp. 117-131, 2020.

[12] S. Chen, M. Zhang, X. Yang, Z. Zhao, T. Zou, and X. Sun, "The impact of attention mechanisms on speech emotion recognition," *Sensors*, vol. 21, no. 22, 7530, 2021.

[13] Mustaqeem, and S. Kwon, "A CNN-assisted enhanced audio signal processing for speech emotion recognition," *Sensors*, vol. 20, no. 1, 183, 2019.

[14] A. M. Badshah, N. Rahim, N. Ullah, J. Ahmad, K. Muhammad, M. Y. Lee, S. Kwon, and S. W. Baik, "Deep features-based speech emotion recognition for smart affective services," *Multimedia Tools and Applications*, vol. 78, pp. 5571-5589, 2019.

[15] T.-Y. Wu, H. Li, S. Kumari, and C.-M. Chen, "A Spectral Convolutional Neural Network Model Based on Adaptive Fick's Law for Hyperspectral Image Classification," *Computers, Materials & Continua*, vol. 79, no. 1, pp. 19-46, 2024.

[16] 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.

[17] T.-Y. Wu, X. Guo, Y.-C. Chen, S. Kumari, and C.-M. Chen, "SGXAP: SGX-Based Authentication Protocol in IoV-Enabled Fog Computing," *Symmetry*, vol. 14, no. 7, 1393, 2022.

[18] J. Deng, Y. Ma, D.-a. Li, J. Zhao, Y. Liu, and H. Zhang, "Classification of breast density categories based on SE-Attention neural networks," *Computer Methods and Programs in Biomedicine*, vol. 193, 105489, 2020.

[19] H. Lv, J. Chen, T. Pan, T. Zhang, Y. Feng, and S. Liu, "Attention mechanism in intelligent fault diagnosis of machinery: A review of technique and application," *Measurement*, pp. 111594, 2022.

[20] Z. Niu, G. Zhong, and H. Yu, "A review on the attention mechanism of deep learning," *Neurocomputing*, vol. 452, pp. 48-62, 2021.

[21] X. Zhao, K. Li, Y. Li, J. Ma, and L. Zhang, "Identification method of vegetable diseases based on transfer learning and attention mechanism," *Computers and Electronics in Agriculture*, vol. 193, 106703, 2022.

[22] G. Huang, J. Zhu, J. Li, Z. Wang, L. Cheng, L. Liu, H. Li, and J. Zhou, "Channel-attention U-Net: Channel attention mechanism for semantic segmentation of esophagus and esophageal cancer," *IEEE Access*, vol. 8, pp. 122798-122810, 2020.

[23] C. Yan, Y. Tu, X. Wang, Y. Zhang, X. Hao, Y. Zhang, and Q. Dai, "STAT: Spatial-temporal attention mechanism for video captioning," *IEEE Transactions on Multimedia*, vol. 22, no. 1, pp. 229-241, 2019.

[24] X. Qin, L. Yulian, W. Dongyue, and L. Bin, "Hyperspectral image classification based on SE-Res2Net and multi-scale spatial spectral fusion attention mechanism," *Journal of Computer-Aided Design & Computer Graphics*, vol. 33, no. 11, pp. 1726-1734, 2021.

[25] W. Li, K. Liu, L. Zhang, and F. Cheng, "Object detection based on an adaptive attention mechanism," *Scientific Reports*, vol. 10, no. 1, 11307, 2020.

[26] H. Gao, L. Cao, D. Yu, X. Xiong, and M. Cao, "Semantic segmentation of marine remote sensing based on a cross direction attention mechanism," *IEEE Access*, vol. 8, pp. 142483-142494, 2020.

[27] Z. Ma, and X. Li, "An improved supervised and attention mechanism-based U-Net algorithm for retinal vessel segmentation," *Computers in Biology and Medicine*, vol. 168, 107770, 2024.