Journal of Network Intelligence
Taiwan Ubiquitous Information
 ISSN 2414-8105
Volume **10**, Number **1**, February **2025**

Driver Speech Emotion Recognition Based on Blind Source Separation and Residual Neural Network

En-Lin Xie*

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

Cai Zhao

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 7, 2024, revised May 28, 2024, accepted July 25, 2024.

Abstract. *In the field of intelligent driving, drivers do not need to manually operate the steering wheel, and can complete the control of the vehicle directly through voice. Human voice contains not only surface textual content information, but also implied emotional information, so that the car understands human emotion is one of the keys to achieve intelligent driving. Therefore, this work proposes a driver speech emotion recognition method based on blind source separation and residual neural network. Firstly, a discrete model is used to construct the required driver speech emotion model and to analyse the noise environment in the car. Considering the influence of other special noises and in order to ensure the effectiveness of speech emotion recognition, a signal-to-noise ratio of -10 30 dB is chosen for the study. Then, due to some limitations of FastICA algorithm in signal separation, especially when Gaussian noise exists in the signal or the data dimension is very high, the effect of signal separation may be affected. Therefore, an improved speech enhancement algorithm combining Complete EEMD with Adaptive Noise (CEEMDAN) and FastICA is proposed to address the shortcomings of the FastICA algorithm. Finally, the ResNet 34 model with four stages and multiple residual residual blocks in each stage is used for driver speech emotion recognition.The inputs of the ResNet 34 model are the denoised and enhanced speech signal features.Experimental results on the CASIA Speech Emotion Database show that the FastICA algorithm combined with empirical mode decomposition is optimised for the separation of speech signals.The ResNet 34 model is used as an input to the ResNet 34 model, and the inputs are the denoised and enhanced speech signal features. The recognition rate of the ResNet 34 model reaches a maximum of 86.93% at 0 dB.*
Keywords: Deep learning; residual networks; blind source separation; emotion recognition; driver speech; ResNet

1. **Introduction.** Driver speech emotion recognition plays a key role in AI driving by providing real-time monitoring and recognition of the driver's emotional state for a smarter, safer and more comfortable driving experience [1, 2]. A driver's emotional state has a significant impact on driving safety and experience. Speech emotion recognition can automatically detect and identify the driver's emotional state, such as anger, stress, and fatigue, by analysing the driver's voice features [3]. This allows the intelligent driving system to make corresponding adjustments and feedbacks to the driver's emotions, such

as adjusting the interior environment and providing a safer and more comfortable driving experience [4].

Driver voice emotion recognition can be used to improve the human-computer interaction experience. Intelligent driving systems can adjust the response and tone of voice assistants according to the driver's emotions, making them more human and emotional. This can facilitate effective communication and co-operation and enhance the interaction experience between the driver and the vehicle. A driver's emotional state may affect his or her driving behaviour and reaction time [5, 6]. By monitoring the driver's emotional state in real time, the intelligent driving system can make targeted adjustments to the driving environment and vehicle operation to improve driver safety and comfort. For example, the system can remind the driver to take a break or take safe driving measures when the driver is in a poor mood or fatigued [7]. Through voice emotion recognition, the intelligent driving system can also monitor the driver's state of health. For example, it can detect the driver's level of fatigue or high-pressure emotions to remind the driver to take a break or take appropriate action to safeguard the driver's safety.

Challenging in the recognition of driver speech emotion is the processing of signals generated in noisy and constantly changing automotive environments, and how driver speech is collected in real environments directly affects the accuracy of the experimental results. Because the driver's emotional state related to vocal cords, such as the pitch and volume of the voice, can be disturbed by vehicle noise, researchers have considered a variety of devices for recording voice, including directional microphones, condenser microphones, and microphone arrays. As former vehicles are increasingly equipped with a wide variety of sensors, in order to better detect driver emotions, in the future speech will come together with facial and physiological signals to combine multiple signals will improve performance. In addition, the contextual information of the speech is also an important indicator [8, 9], and this multimodal emotion recognition will bring a large number of parameters while improving the accuracy. If the traditional machine learning approach is still used the performance will be greatly reduced, in order to be able to find the features that express the driver's emotion from large-scale data, the use of deep learning approach is the study of such problems is becoming increasingly popular. The aim of this study is to separate the driver's speech signal from the complex driving environment using blind source separation technique, and then accurately identify the emotional states, such as fatigue, nervousness, anger, and other negative emotions through residual neural networks, so as to warn potential driving risks and reduce traffic accidents

1.1. **Related Work.** Deep learning-based speech emotion recognition is a research hotspot in the field of artificial intelligence in recent years, which is mainly applied to emotional robots, online education, customer service centres, assisted driving, criminal investigation and other fields.

Convolutional neural network (CNN) [10], as a deep learning model with strong feature extraction and modelling capabilities, has received wide attention in the field of speech emotion recognition. Some preliminary researches have achieved good results and shown greater research potential. Hajarolasvadi and Demirel [11] proposed a speech emotion recognition method based on spectrogram and CNN. Firstly, Wiener filtering algorithm based on a priori signal-to-noise ratio is used to extract and normalise the spectrogram features of speech, and then CNN is selected to classify the spectrogram, and the experimental results achieve a high emotion recognition rate. Makhmudov et al. [12] propose a deep learning-based speech emotion recognition method. Firstly, the speech signal is preprocessed to extract features such as Mel Frequency Cepstrum Coefficient (MFCC).

Then, the extracted features are modelled using CNN and the model output is classified using fully connected neural network (FCN).

Recurrent neural networks (RNNs) [13], especially long and short-term memory networks (LSTMs) [14] and gated recurrent units (GRUs) [15], have strong sequence modelling capabilities and can capture the temporal information in speech signals, and thus have a greater application value in the field of speech emotion recognition. Jiang et al. [16] proposed a speech emotion recognition using RNNs with a local attention mechanism as a method. The RNN is used to capture the temporal information of the speech signal and combined with the local attention mechanism to improve the attention of the model. This allows the model to better recognise emotionally relevant sound features and improve the accuracy of emotion recognition. Zhao et al. [17] proposed an audio sound emotion recognition method based on CNN and LSTM. CNN is used to extract features from the audio stream and LSTM network is used to capture the temporal information. An approach similar to an attention mechanism is also used to help the network pay better attention to the sound features relevant for emotion recognition. The entire model is end-to-end and allows learning emotion recognition tasks directly from raw audio data. Speech emotion recognition refers to identifying the emotional state of a speaker by analysing the emotional information in the speech signal. However, speech signals often have redundancy and noise, which affect the accuracy of emotion recognition. Therefore, speech enhancement plays an important role in speech emotion recognition, which can improve the accuracy and stability of emotion recognition. Hasan et al. [18] proposed a speech enhancement method based on spectral subtraction and energy operator, which analyses and corrects speech signals in the frequency domain on the basis of the short-time Fourier transform, and achieves effective suppression of noise. The results show that the method exhibits good performance in speech emotion recognition tasks. Lei et al. [19] proposed a speech enhancement method based on Fast Independent Component Analysis (FastICA) [20, 21] and nonlinear spectral subtraction by applying FastICA to the speech signal to by applying FastICA to the speech signal, independent components are extracted and the signal is denoised using nonlinear spectral subtraction to achieve effective elimination of noise interference.

1.2. **Motivation and contribution.** Residual neural network solves the problem of gradient vanishing in deep neural networks by introducing residual connections, which allows the model to be deeper and improves the representation ability of the model. Therefore, compared with other deep learning models, residual neural networks have greater potential for application in the field of speech emotion recognition. The FastICA algorithm has some limitations in signal separation, especially when the signal is in the presence of Gaussian noise or when the data dimensionality is very high [22], the effectiveness of the signal separation may be affected. For example, suppose there is a mixed-signal dataset containing multiple audio signals, where each signal is affected by Gaussian noise. In this case, the FastICA algorithm may be disturbed by the noise during signal separation, resulting in failure to accurately separate the original signals.

Therefore, this work proposes a driver speech emotion recognition method based on improved FastICA and residual neural networks. The main innovations and contributions of this work include:

(1) Aiming at the problem of Gaussian noise or data dimensionality that cannot be solved by traditional FastICA algorithms in complex in-vehicle environments with low signal-to-noise ratios, it is proposed that Complete EEMD with adaptive noise (CEEMDAN) [23, 24] be used for signal preprocessing. The original speech signal is decomposed

into intrinsic modal functions (IMFs) of different frequencies, and then independent component analysis is performed using FastICA to find out the most informative sound components, so as to enhance the speech signal.

(2) To address the problem of degradation of deep models during deep learning-based speech emotion recognition, this work uses ResNet 34 model instead of CNN model for speech emotion recognition. ResNet 34 [25] has four stages, each stage has multiple residual residual blocks. Each block has 3 convolutional layers. ResNet 34 is characterised by the design of the residual network which allows it to train deeper networks, solving the problem of degradation of deeper networks.

2. **Driver voice emotion modelling.** Emotion is a special attribute of higher animals, and human dialogue and communication contain emotional expression. By discovering the hidden emotional factors, we can better understand the semantics and know the mood of both parties at this time in order to communicate better. The first step in studying emotions is to choose a representation model that can define and compare different emotional states. Currently, researchers mainly classify emotion classification models into discrete and dimensional models.

(1) Discrete Models

A discrete model of emotion is one that views emotions as several discrete categories of emotion, with all emotions being modifications and combinations of these basic emotions in varying degrees. McDougall introduced the concept of basic emotions in the year of 1919, which was the beginning of the study of emotions in psychology. Different emotions are named with different labels, e.g., anger or fear correspond to two different emotion labels respectively. Different researchers have their own views on how to categorise emotions with labels and how many labels are needed. The relationship between similar facial expressions and emotions in different cultures has resulted in six "basic emotions", such as anger, disgust, happiness, sadness, shock, and fear. In addition, eight bipolar emotions, anger, happiness, sadness, disgust, fear, approval, anticipation, and alarm, which also include intensity and mixed emotions, have been generated.

(2) Dimensional Modelling

Dimensional models are categorised as one, two, three, and multidimensional emotion models. Dimensional models are used to express different emotions by expressing them at different locations in dimensional space. On the other hand, continuous expression of emotions emphasises that emotional states can be constantly changing. Some applications require the identification of emotion level as a continuous variable. The identification of mood level as a continuous variable is done on the basis of its three basic elements: (a) arousal, (b) value, and (c) dominance. Each element is a continuous variable containing values within a given range. For example, a low value indicates low attention, while a high value indicates high attention. Any discrete emotion can be interpreted in this continuous three-dimensional space.

A number of studies have investigated the most frequent emotional states while driving and found that these emotional labels (generated when asking drivers how they feel) are labels that partially overlap with those most frequently described by discrete emotion models. As a result, most researchers have used discrete models, with a few opting for dimensional models. The dimensional model, although more precise and expressive in its description of emotion, is very difficult to quantify specifically across the three dimensions and relies heavily on subjective judgement by the subject, making it difficult to pinpoint a particular emotion. In contrast, the discrete emotion model is simple and intuitive, and the emotions used are common emotions in daily life, which are also applicable to the driver's emotional state. Therefore, the discrete emotion model was adopted in this work.

3. Speech enhancement based on blind source separation in in-vehicle environments.

3.1. Analysis of the noise environment inside the car.

Most of the noise inside the car is low-frequency noise (250Hz $\sim$ 500Hz), which has strong penetrating power and is not easy to deal with in acoustic sound insulation technology. Therefore, in-vehicle noise is unavoidable in the process of car travelling. In-vehicle noise is unavoidable in the process of car travelling. In-vehicle noise is transmitted into the car by the way of resonance between the sound and the car, especially the low-frequency noise. The normal sound of people's normal communication is 40-60 decibels, so the noise inside the car is within the acceptable range of 60 decibels or less.

Voice recorded in a driving environment is not pure voice. Generally speaking, people normally communicate at around 60 to 70 dB, which are measured in SPL (Sound Pressure Level).

$$L_p = 20 \log \left(\frac{P}{P_0} \right) \tag{1}$$

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.

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

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.

$$L_p(r) = L_w - 10 \log \left(4\pi r^2 \right) \tag{3}$$

where r is the straight-line distance between the sound source and the measurement point.

The power level L_w expression is.

$$L_w = 10 \log \left(\frac{W}{W_0} \right) \tag{4}$$

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$\sim$30 dB for the study.

3.2. Speech enhancement based on conventional blind source separation.

FastICA is a statistical method for blind source signal separation that recovers the independent components of the original signal from the mixed signal. Compared with other speech enhancement techniques, FastICA does not require a priori knowledge, has low complexity, and can improve speech enhancement performance even at low signal-to-noise ratios. FastICA can still work well in extremely harsh driving environments.

Suppose there are N unknown source signals $s_i(t), i = 1, 2, \ldots, N$, which are statistically independent from each other, these N source signals form a vector $s = [s_1(t), \ldots, s_N(t)]$, which is captured by M sensors through the linear instantaneous mixing system, and get M observation signals $x = [x_1(t), \ldots, x_M(t)]$, then the M observed signals are linear combinations of these N source signals.

$$x = A \cdot s \tag{5}$$

where A is the linear instantaneous mixing system $m \times n$ mixing matrix, and A is the invertible matrix.

The independent component analysis aims to obtain the estimated signal vector $y = [y_1(t), \ldots, y_N(t)]$ of the source signal by estimating the aliasing matrix A.

$$y = Wx \tag{6}$$

When the above conditions are satisfied then the relationship between the source signal s, the observed signal x, and the estimated signal y can be established.

$$y = Wx = WAs \tag{7}$$

where W is the unmixing matrix. Define the global matrix $G = WA$, ideally, W is the inverse matrix of the admixture matrix A, and the global matrix G is the unit substitution matrix. The unit substitution matrix is the matrix obtained by exchanging some rows and columns of the unit array. Therefore, the process of blind source separation can be mathematically interpreted as the unmixing process of the collapsed matrix. Under the condition that the alias matrix A and the source signal s are unknown, the estimated signal y is optimally approximated to the source signal S according to certain criteria.

The above assumption is a noise-free scenario, but the transmission channel itself and the environment will also be accompanied by a certain amount of noise, so the received observation signals are noisy signals, then Equation (5) is converted as follow:

$$x = As + n \tag{8}$$

where $n = [n_1(t), \ldots, n_N(t)]$ is the noise vector.

The selection of the objective function is often based on the statistical information of the source signal, based on information theory or based on higher order statistics. Whereas, when dealing with blind separation of underdetermined conditions, other means are usually required to obtain the lack of information about the source signal, such as through dictionary learning methods and the construction of virtual multichannels. Centring the mixed signal allows the components of the mixed signal to be transformed into quantities with a mathematical expectation of zero. Centring can be a first step towards simplifying the methods of independent component analysis and reducing the computational complexity, as shown below:

$$\tilde{x} = x - E(x) \tag{9}$$

where x is the mixed signal, $E(x)$ is the mathematical expectation, and $\tilde{x}$ is the centred mixed signal.

Whitening, i.e., eliminating the correlation between the components of the observed signal by a specific linear transformation, and normalising it so that its covariance is 1 and the covariance matrix is a unit array, is usually achieved by using eigenvalue decomposition. The whitened signal is denoted as $\bar{x}$.

$$\bar{x} = Q\tilde{x} \tag{10}$$

where Q is the whitening matrix.

By whitening, $\bar{x}$ is made to have unit variance and its covariance matrix is $R_x = I$. Eigenvalue decomposition of R_x.

$$R_x = V\Lambda V^T = V\Lambda^{1/2}\Lambda^{1/2}V^T \tag{11}$$

Then the whitening matrix Q is defined as.

$$Q = \Lambda^{-1/2}V^T \tag{12}$$

where V is the orthogonal matrix of eigenvectors and Λ is the diagonal array of eigenvalues.

When the source signal S is a combination of several independent signals, the above steps need to be continued sequentially. Subtract each independent source signal $s_i(t)$ from the observed signal X until the source signals are completely separated. The FastICA algorithm has some limitations in signal separation, especially when the signal is in the presence of Gaussian noise or when the data dimensions are very high, the effectiveness of the signal separation may be affected. In order to solve the problems that arise, this paper proposes an improved speech enhancement algorithm that combines CEEMDAN and FastICA to address some of the shortcomings of the FastICA algorithm.

3.3. **Improved speech enhancement based on CEEMDAN-FastICA.** By combining CEEMDAN and FastICA, the time-frequency localisation of modal decomposition and the independent component separation ability of FastICA can be fully utilised to enhance signal separation. CEEMDAN can be used as a pre-processing means to convert the signal to the time-frequency domain, and then FastICA algorithm can be applied to separate the independent components. This combination method can improve the accuracy, robustness and reliability of signal separation, which is especially suitable for speech signal enhancement problems in complex in-vehicle environments.

CEEMDAN is an improved EMD method that can handle nonlinear and nonsmooth signals more effectively, making the decomposition results more accurate and reliable. Compared with EMD and EEMD, CEEMDAN introduces an adaptive noise model, which can better remove the noise in the signal and improve the accuracy and stability of the signal. CEEMDAN adopts an integrated learning approach to decompose the signal by combining several independent EMD algorithms, which improves the stability and robustness of the algorithm.

Therefore, in this paper, a combination of CEEMDAN and FastICA can be used to solve the mixed signal separation problem in the speech enhancement task. CEEMDAN is a data-based adaptive signal decomposition method, which can decompose the mixed signal into multiple intrinsic modal functions (IMFs), and each IMF corresponds to a signal component at different frequencies. FastICA is an independent component analysis method which can estimate and separate the independent components of a mixed signal. By using the IMFs obtained from CEEMDAN as inputs and combining the separation capability of FastICA, the different signal components in the mixed signal can be separated more effectively, thus effectively improving the effect of speech enhancement.

Functions or physical quantities such as instantaneous frequencies that exhibit local features show their great advantages for the analysis of non-stationary signals. Therefore, before discussing the CEEMDAN-based separation algorithm, the instantaneous frequencies are established first. For an arbitrary time series $x(t)$, its Hilbert transform is:

$$y(t) = \frac{1}{\pi}\int_{-\infty}^{\infty}\frac{x(\tau)}{t-\tau}d\tau \tag{13}$$

Construct the parser function as follows:

$$z(t) = x(t) + iy(t) = a(t)e^{j\varphi(t)} \tag{14}$$

where $a(t) = \sqrt{x^2(t) + y^2(t)}$ is the magnitude function, and $\varphi(t) = \arctan(y(t)/x(t))$ is the phase function.

The phase function $\varphi(t)$ is derived from time as the instantaneous frequency.

$$\omega(t) = \frac{d\varphi(t)}{d(t)} \tag{15}$$

In order to overcome the drawback of introducing pairs of additional noises in EMD and generating noise residuals in the reconstructed signal, CEEMDAN adds a specific noise at each stage of the decomposition, generating a residual component that is used only for obtaining the intrinsic modal function at each stage, providing an accurate reconstruction of the original signal and a better separation of the signal components at a lower computational cost. A set of noise ω_i is first added to the original signal x.

$$M_i = x + \omega_i \tag{16}$$

EMD is done separately for M_i and averaged over each of the obtained imf to obtain the final first-order eigenmode function $imf_1[n]$.

$$r_1[n] = x[n] - imf_1[n] \tag{17}$$

Define the operator $E_j(\cdot)$ to denote the jth mode obtained by performing EMD on the contents of the brackets. Perform a first-order modal computation on a set of additional noise $\omega_i[n]$ and add it to the first-order residual component $r_1[n]$.

$$r_1[n] = r_1[n] + \varepsilon_1 E_1(\omega_i[n]) \tag{18}$$

EMD is performed on the above equation, and the first-order modes obtained are ensemble averaged as the second-order eigenmode function of the original signal x.

$$imf_2[n] = \frac{1}{I}\sum_{i=1}^{I} E_1\left(r_1[n] + \varepsilon_1 E_1(w_i[n])\right) \tag{19}$$

By analogy, the CEEMDAN decomposition of the original signal is obtained.

$$imf_{k+1}[n] = \frac{1}{I}\sum_{i=1}^{I} E_1\left(r_k[n] + \varepsilon_k E_k(\omega_i[n])\right) \tag{20}$$

CEEMDAN reduces the number of iterations while accurately eliminating the effect of additional noise on the residual noise of the reconstructed signal, and solves the problem of choosing different additional noise sizes to produce different decomposition modes.

Using CEEMDAN's ability to adaptively decompose the signals themselves, the resulting eigenmode functions are used as virtual extensions of the multichannel signals, thus transforming the underdetermined puzzle into positive definite conditions. The source signals are then separated by borrowing the FastICA separation algorithm. For the three source signals s_1, s_2 and s_3, taking the two observed signals x_1 and x_2 obtained after aliasing as an example, the steps of the improved CEEMDAN-FastICA-based speech enhancement method are as follows.

(1) The observed signals x_1 and x_2 are decomposed into multiplexed eigenmode functions $imf_1 = [imf_1, imf_2, \ldots, imf_n]^T$ and $imf_2 = [imf_{2,1}, imf_{2,2}, \ldots, imf_{2,n}]^T$ using CEEMDAN.

(2) Calculate the similarity coefficients between each intrinsic modal function imf_1 and imf_2, and the observed signals, select the component with the largest correlation coefficient, eliminate the redundant components, and assume that imf_1 is the component with the largest correlation coefficient and (x_1, x_2) to form a three-channel mixed signal.

(3) Separation of the newly generated observation signals is achieved using the FastICA algorithm.

In ICA, computational complexity makes the calculation of negentropy often not straightforward. Therefore, an approximate estimation of the negentropy is often made by approximating the probability density function with a non-polynomial function.

$$J(w) \approx c\left[E\left\{G\left(w^T\bar{x}\right)\right\} - E\left\{G(v)\right\}\right]^2 \tag{21}$$

where $G(\cdot)$ is a non-quadratic function, c is a constant, w is a separation vector, v is a Gaussian random variable with zero mean and unit variance, and $\bar{x}$ is a mixed signal pre-processed with centring and whitening.

The approximate estimation of negative entropy requires a normalisation constraint on the random variables, i.e., the variance of the separated signals is required to be 1. This constraint flows from the preprocessed signal vector $\bar{x}$ to the separation vector w, which normalises the separation vector, i.e., $\|w\|^2 = 1$. The cost function of the FastICA algorithm can be fully expressed as follows.

$$\begin{cases} J(w) \approx c\left[E\left\{G\left(w^T\bar{x}\right)\right\} - E\left\{G(v)\right\}\right]^2 \\ \|w\|^2 = 1 \end{cases} \tag{22}$$

FastICA uses Newton's iteration method for optimisation operations. Newton's iterative method can solve for approximate roots or even heavy roots of an equation in the real complex number domain. The function $f(x)$ is expanded in a Taylor series in some neighbourhood of the point x_0 as.

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + \frac{f''(x_0)(x - x_0)^2}{2!} + \cdots + \frac{f^{(n)}(x_0)(x - x_0)^n}{n!} + R_n(x) \tag{23}$$

The iterative relation equation for Newton's iterative method is derived as:

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} \tag{24}$$

According to the Kuhn-Tucker condition, the condition for the cost function equation (20) to reach an extreme point is.

$$E\left\{xg\left(w^Tx\right)\right\} - \beta w = 0 \tag{25}$$

where β is a constant and g is the first order derivative of the function G.

$$F(w) = E\left\{xg\left(w^Tx\right)\right\} - \beta w \tag{26}$$

Solve for the extreme points of the cost function, i.e., solve for $F(w) = 0$. Solve the above equation according to Equation (22).

$$w_{k+1} = w_k - \frac{F(w_k)}{\partial F(w_k)} \tag{27}$$

where k is the number of iterations, $\partial F(w_k)$ is the first order derivative of $F(w_k)$ and w_k is the separation vector. Taking the first order derivative of $F(w_k)$ and normalising it, the iterative counting method of FastICA can be obtained as follows:

$$w^* = E\left\{xg\left(w_k^T x\right)\right\} - E\left\{g'\left(w_k^T x\right)\right\} w_k \tag{28}$$

$$w_{k+1} = \frac{w^*}{\|w^*\|} \tag{29}$$

4. **Speech emotion recognition based on residual neural networks.** In the process of deep learning-based speech emotion recognition, it has been found that when the number of model layers increases to a certain level, the effectiveness of the model will decrease instead of increasing. In other words, degradation of the deep model occurs. To solve this problem, the internal structure of the model needs to be made capable of constant mapping to ensure that during the process of stacking the network, the network will not be degraded at least by continuing to stack. Therefore, in this work, the ResNet 34 model is used instead of the CNN model for speech emotion recognition. The inputs to the model are the spectral features of the speech signal enhanced by the data described above.

ResNet 34 has 4 stages, each with multiple residual residual blocks. Each block has 3 convolutional layers with ReLU activation function and batch normalisation in between. The residual block in the first stage has 2 convolutional layers and the residual block in the rest of the stages has 3 convolutional layers. The size of each convolutional layer of ResNet 34 is 3×3 and the total number of layers is 34, so it is called ResNet 34. ResNet 34 is one of the most commonly used models in image-net training. ResNet 34 is used for image classification, is more effective, and has the advantages of high accuracy and fast training process, making it one of the most applicable models.

ResNet 34 is a classic model in the ResNet family [26, 27] that uses 34 convolutional layers to extract features. ResNet 34 features a residual network design that allows it to train deeper networks, solving the problem of degradation of deeper networks. ResNet 18 has a relatively simple network structure with shallow depth, which is a good choice for scenarios with limited computational resources, but ResNet 34 performs better in complex in-vehicle environments. ResNet 34 has connectors that ensure that the deeper networks perform at least as well as the shallower networks. A good choice, but in complex in-vehicle environments, ResNet 34 performs better. ResNet 34 deepens the network while ensuring that the performance of the deeper network is at least equal to that of the shallower network. The structure of the ResNet 34 layer network is shown in Figure 1.

Firstly, a convolutional layer, followed by a pooling layer. Then there is a series of residual units, which are again downsampled by one average pooling. Finally, the final output is obtained by a fully connected layer. The structure of this network looks very simple and is basically obtained by stacking the residual structures. The configuration parameters of the ResNet 34-layer network are shown in Table 1.

As the neural network becomes deeper, the range of gradient variations becomes larger, which can lead to a more significant occurrence of the gradient vanishing or gradient explosion problem. This is usually solved by data normalisation, weight initialisation, and batch normalisation (BN). After solving the gradient vanishing or explosion problem, we still have the problem of poor results with deep layers. This problem can be solved by residual structure, the deeper the layer, the smaller the error rate.

This model convolves the input on the main line, after two 3*3 blocks of convolved residuals. There is a connecting line on the right side to connect directly from the input to the output. The feature matrix obtained on the branch is then added to the input's feature matrix in a summing operation, which is then passed through the ReLU function.

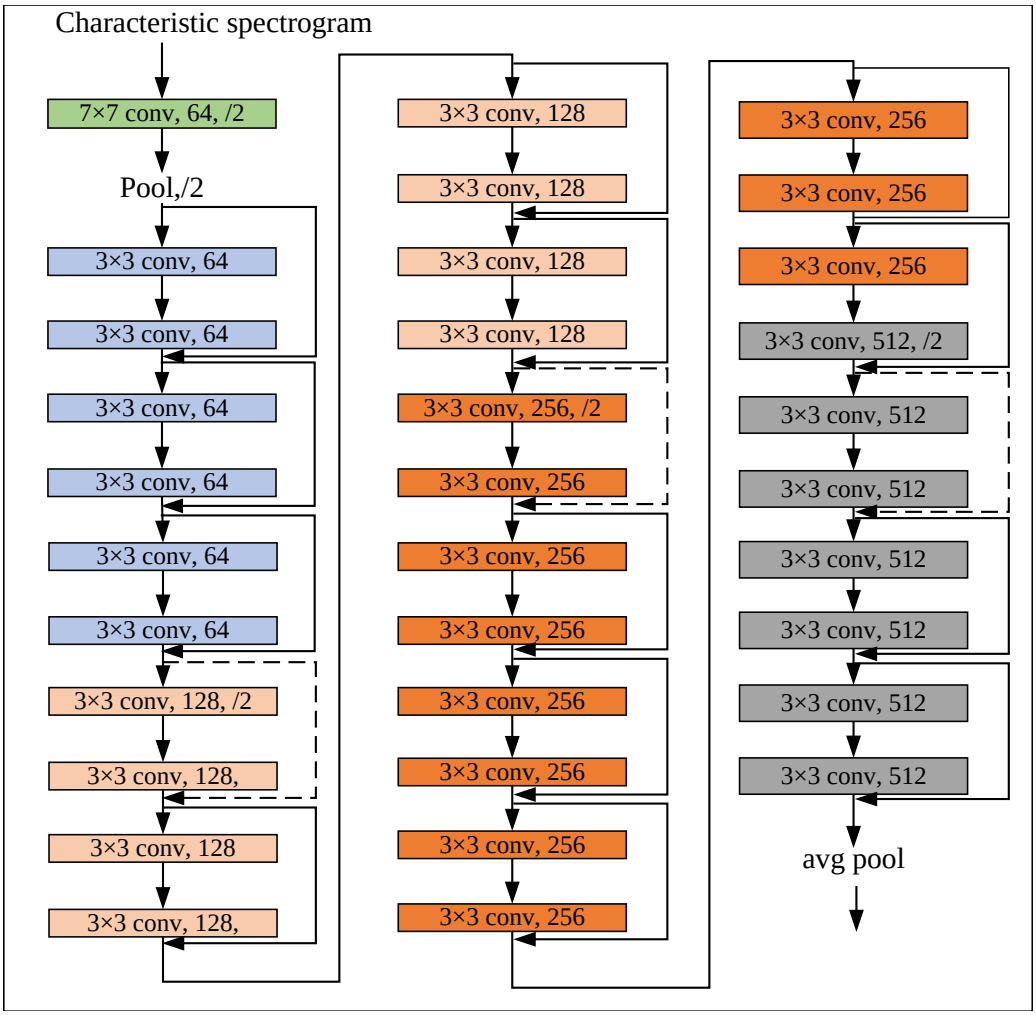

Figure 1. ResNet 34-layer network structure

Table 1. Configuration Parameters for ResNet Layer 34 Networks

Layer	**Output shape**	**Param**
Sequential_13	(None, 56, 56, 64)	9716
Sequential_14	(None, 56, 56, 64)	235572
Sequential_18	(None, 28, 28, 128)	1168884
Sequential 23	(None, 14, 14, 256)	7160308
Sequential 30	(None, 7, 7, 512)	13648372

Since it is a summing operation, it should be noted that the main branch feature matrix and the output feature matrix of the shortcut must be the same.

5. **Experimental results and analyses.**

5.1. **Experimental environment and dataset.** In the experimental phase, 1200 speech samples from the CASIA linguistic emotion database were used, including six emotions: angry, happy, fear, sad, surprise, and neutral. CASIA was recorded in a pristine, noise-free studio, containing a total of 9600 different articulatory utterances from two males and two females. CASIA was recorded with a sampling rate of 16k Hz, a quantisation resolution of 16 bit, and a storage format of pcm. The training and test sets were divided 4:1, with 960 pure language samples for training and 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.

5.2. **Effectiveness analysis of speech enhancement.** After randomly selecting a pure speech in the CASIA linguistic sentiment database and superimposing 0dB music background noise, the FastICA algorithm and the CEEMDAN-FastICA algorithm were used for denoising and enhancement, respectively. The intuitive speech waveforms show that the pure original speech is severely impaired by the time domain waveform interference of the music noise superimposed on it. A comparison of the denoising enhancement effects of the two algorithms is shown in Figure 2.

It can be found that CEEMDAN-FastICA improves the output signal-to-noise ratio for all four noise environments under all conditions except for music noise at a signal-to-noise ratio of -10 dB, and in particular, it has the most prominent effect on engine noise and vehicle tyre noise. Also at 0 dB SNR, the CEEMDAN-FastICA algorithm has the highest overall improvement rate for all four noises. Overall, it shows that the FastICA algorithm combined with CEEMDAN is optimised for speech signal separation.

5.3. **Performance comparison of different recognition models.** In order to explore the performance of different models in white noise environments, the recognition rates of different speech emotion recognition models in white noise environments are compared after denoising and enhancement using the CEEMDAN-FastICA algorithm, as shown in Table 2.

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

White noise	∞	**30 dB**	**20 dB**	**10 dB**	**0 dB**	**-10 dB**
CNN	73.37	74.75	75.28	76.91	82.93	50.58
LSTM	74.45	76.55	75.96	78.89	83.38	52.25
CNN+LSTM	77.01	75.93	80.29	82.63	85.33	55.13
ResNet34	78.28	79.36	81.89	84.37	86.93	53.37

It can be seen that when the SNR is higher than 0 dB, all the four deep learning models are able to improve the recognition rate of speech emotion in noisy environments, with the ResNet34 model showing the best improvement. the recognition rate of the ResNet34 model reaches a maximum of 86.93

6. **Conclusion.** In this work, a driver speech emotion recognition method based on improved FastICA and residual neural network is proposed. Aiming at the problem of Gaussian noise or data dimensionality that cannot be solved by the traditional FastICA algorithm in low signal-to-noise ratio complex in-vehicle environments, it is proposed that CEEMDAN be used for signal preprocessing. The original speech signal is decomposed into intrinsic modal functions (IMFs) of different frequencies, and then independent component analysis is performed using FastICA to find out the most informative sound components so as to enhance the speech signal.

To address the problem of degradation of deep models during deep learning-based speech emotion recognition, this work uses ResNet 34 model instead of CNN model for speech emotion recognition. ResNet 34 has 4 stages, each stage has multiple residual residual blocks. Each block has 3 convolutional layers. ResNet 34 is characterised by the design of the residual network which makes it possible to train deeper networks, solving

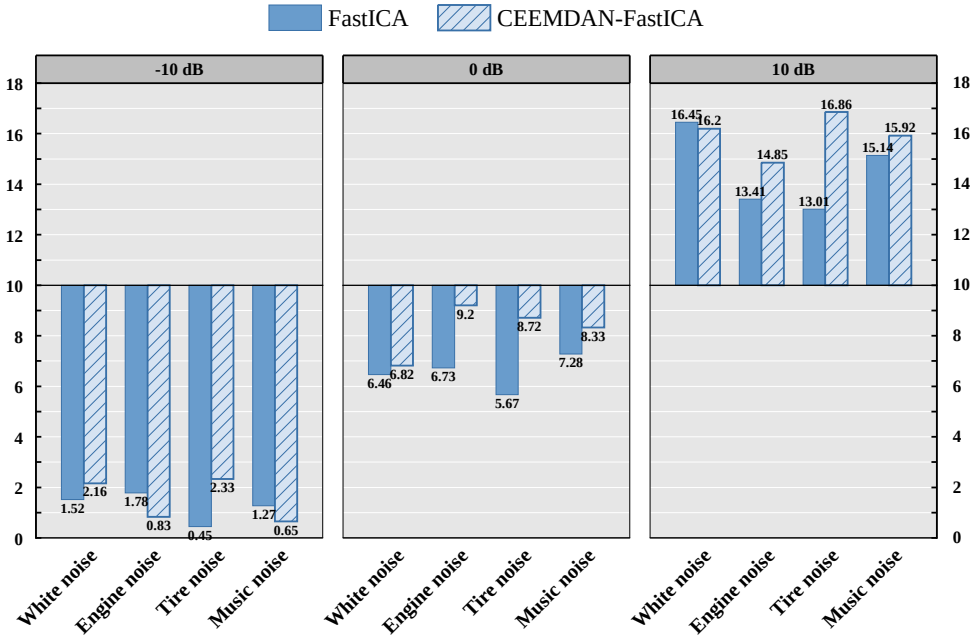

Figure 2. Comparison of Denoising Enhancement Effects

the problem of deep network degradation. Overall, it shows that the FastICA algorithm combined with CEEMDAN is optimised for speech signal separation.

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

REFERENCES

[1] H. Xiao, W. Li, G. Zeng, Y. Wu, J. Xue, J. Zhang, C. Li, and G. Guo, "On-road driver emotion recognition using facial expression," *Applied Sciences*, vol. 12, no. 2, p. 807, 2022.

[2] C. D. Katsis, N. Katertsidis, G. Ganiatsas, and D. I. Fotiadis, "Toward emotion recognition in car-racing drivers: A biosignal processing approach," *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, vol. 38, no. 3, pp. 502-512, 2008.

[3] D. Zhou, Y. Cheng, L. Wen, H. Luo, and Y. Liu, "Drivers' Comprehensive Emotion Recognition Based on HAM," *Sensors*, vol. 23, no. 19, p. 8293, 2023.

[4] G. Du, Z. Wang, B. Gao, S. Mumtaz, K. M. Abualnaja, and C. Du, "A convolution bidirectional long short-term memory neural network for driver emotion recognition," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 7, pp. 4570-4578, 2020.

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

[6] W. Li, G. Zeng, J. Zhang, Y. Xu, Y. Xing, R. Zhou, G. Guo, Y. Shen, D. Cao, and F.-Y. Wang, "Cogemonet: A cognitive-feature-augmented driver emotion recognition model for smart cockpit," *IEEE Transactions on Computational Social Systems*, vol. 9, no. 3, pp. 667-678, 2021.

[7] D. Ververidis, and C. Kotropoulos, "Emotional speech recognition: Resources, features, and methods," *Speech Communication*, vol. 48, no. 9, pp. 1162-1181, 2006.

[8] R. Fernandez, and R. W. Picard, "Modeling drivers' speech under stress," *Speech Communication*, vol. 40, no. 1-2, pp. 145-159, 2003.

[9] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572-587, 2011.

[10] H.-C. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. M. Summers, "Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1285-1298, 2016.

[11] N. Hajarolasvadi, and H. Demirel, "3D CNN-based speech emotion recognition using k-means clustering and spectrograms," *Entropy*, vol. 21, no. 5, p. 479, 2019.
[12] F. Makhmudov, A. Kutlimuratov, F. Akhmedov, M. S. Abdallah, and Y.-I. Cho, "Modeling Speech Emotion Recognition via Attention-Oriented Parallel CNN Encoders," *Electronics*, vol. 11, no. 23, p. 4047, 2022.
[13] R. Kanakala, and K. Reddy, "Modelling a deep network using CNN and RNN for accident classification," *Measurement: Sensors*, vol. 15, p. 100794, 2023.
[14] D. Zhao, Y. Liu, H. Yin, and Z. Wang, "An attentive and adaptive 3D CNN for automatic pulmonary nodule detection in CT image," *Expert Systems with Applications*, vol. 211, p. 118672, 2023.
[15] G. Z. De Castro, R. R. Guerra, and F. G. Guimarães, "Automatic translation of sign language with multi-stream 3D CNN and generation of artificial depth maps," *Expert Systems with Applications*, vol. 215, p. 119394, 2023.
[16] P. Jiang, X. Xu, H. Tao, L. Zhao, and C. Zou, "Convolutional-recurrent neural networks with multiple attention mechanisms for speech emotion recognition," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 14, no. 4, pp. 1564-1573, 2021.
[17] J. Zhao, X. Mao, and L. Chen, "Speech emotion recognition using deep 1D & 2D CNN LSTM networks," *Biomedical Signal Processing and Control*, vol. 47, pp. 312-323, 2019.
[18] M. K. Hasan, S. Salahuddin, and M. R. Khan, "A modified a priori SNR for speech enhancement using spectral subtraction rules," *IEEE Signal Processing Letters*, vol. 11, no. 4, pp. 450-453, 2004.
[19] P. Lei, M. Chen, and J. Wang, "Speech enhancement for in-vehicle voice control systems using wavelet analysis and blind source separation," *IET Intelligent Transport Systems*, vol. 13, no. 4, pp. 693-702, 2019.
[20] E. Oja, and Z. Yuan, "The FastICA algorithm revisited: Convergence analysis," *IEEE Transactions on Neural Networks*, vol. 17, no. 6, pp. 1370-1381, 2006.
[21] C. W. Hesse, and C. J. James, "The FastICA algorithm with spatial constraints," *IEEE Signal Processing Letters*, vol. 12, no. 11, pp. 792-795, 2005.
[22] X. Zhao, and H. Yang, "A new method to calculate the utility harmonic impedance based on FastICA," *IEEE Transactions on Power Delivery*, vol. 31, no. 1, pp. 381-388, 2015.
[23] J. Cao, Z. Li, and J. Li, "Financial time series forecasting model based on CEEMDAN and LSTM," *Physica A: Statistical Mechanics and Its Applications*, vol. 519, pp. 127-139, 2019.
[24] F. Zhou, Z. Huang, and C. Zhang, "Carbon price forecasting based on CEEMDAN and LSTM," *Applied Energy*, vol. 311, p. 118601, 2022.
[25] 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.
[26] Y. Ma, Y. Peng, and T.-Y. Wu, "Transfer learning model for false positive reduction in lymph node detection via sparse coding and deep learning," *Journal of Intelligent & Fuzzy Systems*, vol. 43, no. 2, pp. 2121-2133, 2022.
[27] 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, p. 40, 2019.