Speech Emotion Recognition Enhanced Traffic Efficiency Solution for Autonomous Vehicles in a 5G-Enabled Space–Air–Ground Integrated Intelligent Transportation System

Liang Tan, Keping Yu, *Member, IEEE*, Long Lin, Xiaofan Cheng, Gautam Srivastava, *Senior Member, IEEE*, Jerry Chun-Wei Lin, *Senior Member, IEEE*, and Wei Wei, *Senior Member, IEEE*

***Abstract*— Speech emotion recognition (SER) is becoming the main human–computer interaction logic for autonomous vehicles in the next generation of intelligent transportation systems (ITSs). It can improve not only the safety of autonomous vehicles but also the personalized in-vehicle experience. However, current vehicle-mounted SER systems still suffer from two major shortcomings. One is the insufficient service capacity of the vehicle communication network, which is unable to meet the SER needs of autonomous vehicles in next-generation ITSs in terms of the data transmission rate, power consumption, and latency. Second, the accuracy of SER is poor, and it cannot provide sufficient interactivity and personalization between users and vehicles. To address these issues, we propose an SER-enhanced traffic efficiency solution for autonomous vehicles in a 5G-enabled space–air–ground integrated network (SAGIN)-based ITS. First, we convert the vehicle speech information data into spectrograms and input them into an AlexNet network model to obtain the high-level features of the vehicle speech acoustic model. At the same time, we convert the vehicle speech information data into text information and input it into the Bidirectional Encoder Representations from Transformers (BERT) model to obtain the high-level features of the corresponding text model. Finally, these two sets of high-level features are cascaded together to obtain fused features, which are sent to a softmax classifier for emotion matching and classification. Experiments show that the proposed solution can improve not only the SAGIN's service capabilities, resulting in a large capacity, high bandwidth, ultralow latency, and high reliability, but also the accuracy of vehicle SER as well as the performance, practicality, and user experience of the ITS.**

***Index Terms*— Speech emotion recognition, autonomous vehicles, artificial intelligence, 5G-enabled SAGIN, ITS.**

I. Introduction

INTELLIGENT transportation systems (ITSs) represent the effective integration of advanced information technology, data communication transmission technology, electronic sensing technology, electronic control technology and computer processing technology for application to the entire transportation management system [1], [2]. Such a system has a large range and provides comprehensive transportation and management services that work in all directions, offer real-time performance, and are accurate and efficient. Through close cooperation among people, vehicles, and roads, transportation efficiency can be improved, traffic congestion can be relieved, the road network passing capacity can be increased, the occurrence of traffic accidents can be reduced, and energy consumption and environmental pollution can be lowered [3], [4]. Such systems have gradually been adopted for passenger flow guidance at airports and transit stations, intelligent urban traffic dispatch, intelligent highway dispatch, operational vehicle dispatch management, and automatic motor vehicle control.

Recently, continuous innovations in autonomous vehicles for next-generation ITSs have deeply affected the entire automotive industry [5], [6]. Therefore, many automakers and suppliers have accelerated their deployment of future vehicles, including autonomous and semiautonomous vehicles, and speech recognition is gradually becoming one of the

Manuscript received January 13, 2021; revised July 16, 2021; accepted September 15, 2021. Date of publication October 28, 2021; date of current version March 9, 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 61373162, in part by the Sichuan Provincial Science and Technology Department Project under Grant 2019YFG0183, in part by the Sichuan Provincial Key Laboratory Project under Grant KJ201402, and in part by the Japan Society for the Promotion of Science (JSPS) Grants-in-Aid for Scientific Research (KAKENHI) under Grant JP18K18044 and Grant JP21K17736. The Associate Editor for this article was N. Zhang. *(Corresponding author: Keping Yu.)*

Liang Tan is with the College of Computer Science, Sichuan Normal University, Chengdu 610101, China, and also with the Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China (e-mail: jkxy_tl@sicnu.edu.cn).

Keping Yu is with the Global Information and Telecommunication Institute, Waseda University, Tokyo 169-8050, Japan (e-mail: keping.yu@aoni.waseda.jp).

Long Lin and Xiaofan Cheng are with the College of Computer Science, Sichuan Normal University, Chengdu 610101, China (e-mail: 569074330@qq.com; sail967642@gmail.com).

Gautam Srivastava is with the Department of Computer Science, Brandon University, Brandon, MB R7A 6A9, Canada, and also with the Research Centre for Interneural Computing, China Medical University, Taichung 40402, Taiwan (e-mail: srivastavag@brandonu.ca).

Jerry Chun-Wei Lin is with the Department of Computer Science, Electrical Engineering and Mathematical Sciences, Western Norway University of Applied Sciences, 5063 Bergen, Norway (e-mail: jerrylin@ieee.org).

Wei Wei is with the School of Computer Science and Engineering, Xi'an University of Technology, Xi'an 710048, China (e-mail: weiwei@xaut.edu.cn).

Digital Object Identifier 10.1109/TITS.2021.3119921

standard features of autonomous vehicles [7]. In recent years, various devices in self-driving vehicles have been designed to be operated by driver speech recognition so that the driver does not need to operate these devices directly. However, waking up the onboard speech assistant for simple operations such as navigating, playing music, and opening and closing windows while an autonomous vehicle is in motion is no longer attractive to users.

Currently, autonomous vehicles are facing a fundamental change in the logic of human–computer interaction [8]. Today's users are eager to form a closer and more conversational interactive relationship with their devices. Accordingly, new artificial intelligence (AI) systems [9], [10] are expected to give rise to a new generation of vehicles that will provide a new type of experience for users: the vehicle itself will become an interactive interface that connects the driver, the passengers, the vehicle itself and Internet of Things vehicle-control-related mobile devices. Using speech emotion recognition (SER) technology (technology that can perceive and analyze human emotional expression), vehicles will soon be able to perceive our reactions and emotions and respond accordingly. These vehicles with "emotional awareness" will benefit the automotive industry and users in many ways. At the International Consumer Electronics Show held in January 2018, the media took note of Qualcomm showcasing the Amazon Alexa speech recognition function on its Smart Audio platform in a vehicle; this in-vehicle virtual assistant is designed to make speech a natural communication interface between the driver and the vehicle [11]. As platforms such as VehiclePlay, Android Auto and Echo Auto penetrate the passenger vehicle market, in-vehicle SER technology is expected to become mainstream [12].

The SER function of an autonomous vehicle collects speech signals from within the vehicle and uploads them to the cloud. The SER model in the cloud then recognizes the collected speech. Once the semantics of the user in the vehicle are understood, interaction with the user is carried out. For the next generation of ITSs, SER in self-driving vehicles still faces two problems. One is the insufficient service capacity of the in-vehicle communication network. As the foundation for the operation of autonomous vehicles, in-vehicle communication networks provide the basic implementation of information exchange between autonomous vehicles and other vehicles in the cloud environment. In urban areas, in-vehicle communication networks mainly achieve high-data-rate access through dedicated short-range-communication 802.11p networks and cellular networks; in remote rural areas, satellite communication systems can provide seamless connections [13], [14]. A space–air–ground integrated network (SAGIN) is a communication network that can provide in-vehicle services for the next generation of ITSs. Such a network is usually composed of several ground networks, low-Earth-orbit (LEO) satellites, unmanned aerial vehicles (UAVs), and high-altitude platforms. However, neither ground networks nor air networks can meet the needs of SER for the next generation of autonomous vehicles for intelligent transportation in terms of the data transmission rate, power consumption and delay [15]. The second problem is the poor semantic recognition performance of vehicle SER. Current in-vehicle SER methods no longer use traditional SER algorithms, such as hidden Markov models (HMMs) with binding states, adaptive technology, Gaussian mixture models (GMMs), HMMs, or modified GMMs based on these algorithms, because they face complex problems. The number of states in the HMM algorithm increases exponentially, causing the system to become complex and difficult to control, while the GMM algorithm cannot learn deep nonlinear feature transformations and has certain limitations in the extraction of abstract features. In recent years, however, the rapid development of deep learning has enabled great breakthroughs in SER [16]. Deep learning models [17]–[19] such as backpropagation (BP) neural networks, convolutional neural networks (CNNs), and long short-term memory (LSTM) networks can be used to extract more information from massive multisource vehicle corpora, and based on this valuable information, a recognition rate of close to 70% can be achieved. However, for vehicle SER, the accuracy is still low.

To address this issue, this paper proposes a fifth-generation (5G) networking technology-based SAGIN multimodal emotion recognition model that combines speech and text. Our main contributions are as follows:

1) To solve the problem of the insufficient service capacity of the in-vehicle communication network, we propose a 5G-based SAGIN solution. Through 5G technology, the capacity, bandwidth, and reliability of the in-vehicle communication network can be improved, and the communication delay can be reduced.
2) To improve the accuracy of vehicle SER, a multimodal emotion recognition model combining vehicle speech and text is proposed. We improve the in-car SER performance to 74.0% in terms of the weighted accuracy WA) and 65.4% in terms of the unweighted accuracy (UA).

The structure of this paper is as follows: Section II describes the related work. Section III describes the architecture of a 5G-enabled SAGIN for an ITS. Section IV demonstrates the SER model for autonomous vehicles in a 5G-enabled SAGIN for an ITS. The experimental design and analysis are presented in Section V. Finally, Section VI summarizes the paper.

II. Related Work

In this section, we will review relevant research on SAGINs, existing in-vehicle-based systems, and in-vehicle speech recognition.

A SAGIN uses modern information network technology to connect space, air, and ground network segments [15]. The existing SAGIN research includes network design and resource allocation as well as performance analysis and optimization. For example, Qu [20] proposed a software-defined networking (SDN)-based SAGIN architecture and introduced an SDN controller into a SAGIN to improve the flexibility and programmability of network management. Other examples include [21], [22]. In addition, some performance optimization methods have been reported. For example, Kato [23] proposed using AI technology to optimize SAGIN performance and

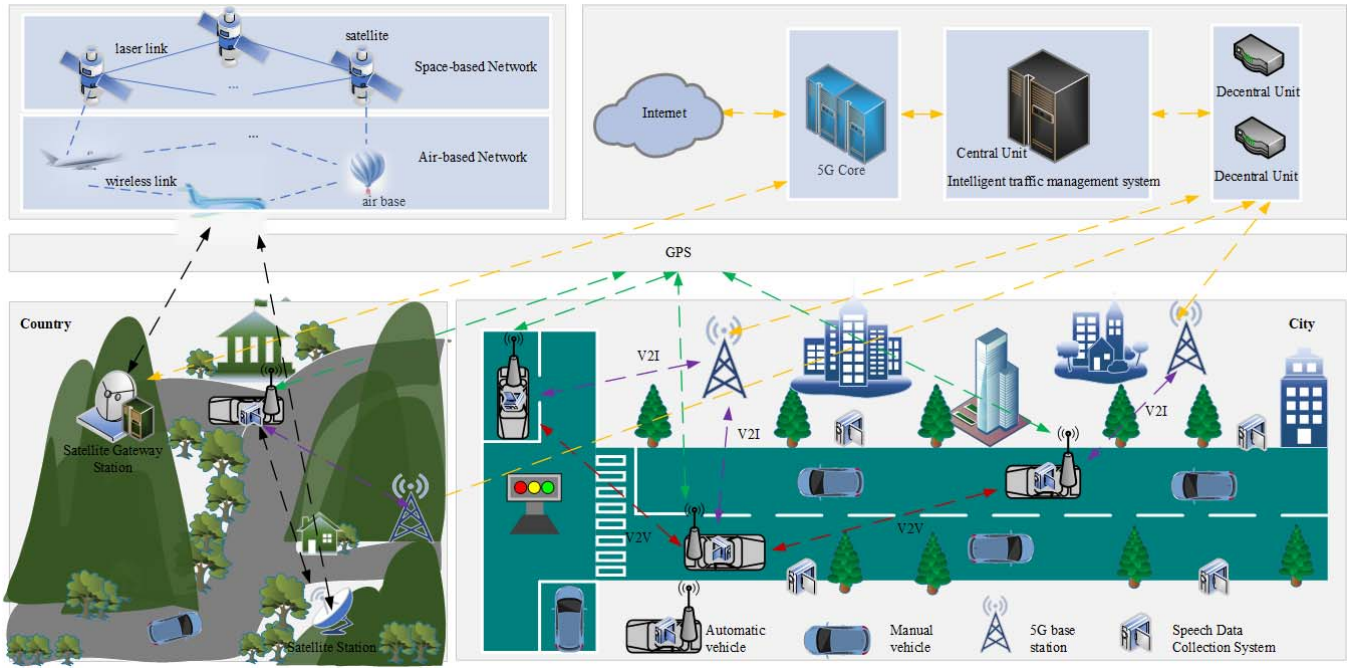


Fig. 1. Architecture of the 5G-enabled SAGIN.

then, using satellite traffic balancing as an example, designed a method based on deep learning to enhance traffic control performance. Other similar examples include [24], [25].

At the same time, to address the application requirements of intelligent traffic management [26], [27], intelligent vehicle control [28], and intelligent road network information services, several studies on SAGINs and on-board systems have emerged. For service provision, Zhang *et al.* [29] proposed an SDN architecture with a hierarchical structure to support various in-vehicle services. For network resource management, Wu [30] proposed an SDN resource management framework by designing a hybrid hierarchical space–air–ground integrated vehicular network control architecture with real-time formulation of resource management strategies to balance different situations and address the performance and overhead of system state acquisition. For authentication, Zhao [31] proposed an identity authentication and privacy protection scheme based on blockchain [32], [33] for a SAGIN, aiming to address the security challenges presented by high mobility and low latency.

In addition, with the development of emerging technologies such as AI, in-vehicle voice recognition has become a research topic of widespread concern, and related research can be mainly divided into three areas. Regarding service provision, to improve driving safety, Su [34], Idros [35] and Mahtab *et al.* [36] studied a vehicle control system based on voice recognition technology for controlling the vehicle by voice, with the aim of improving driving safety. At the same time, to address the vulnerabilities of automatic speech recognition modules, the authors of [37] designed and developed the SIEVE system to effectively distinguish among the voice of the driver, the voices of passengers, and electronic speakers. Second, regarding user behavior and experience, Zhang *et al.* [38] designed a quality of experience (QoE) system for in-vehicle voice cloud services, and Yu *et al.* [39] proposed a vehicle-mounted voice cloud server test system that records user experience indicators. Finally, regarding performance evaluation and tuning, Amman *et al.* [40] studied the influence of microphone position on vehicle speech recognition and configured corresponding strategies to improve accuracy.

As seen from the above summary, in the existing research, there are few works related to the combination of SAGINs and vehicle speech recognition; in particular, a multimodal emotion recognition model based on a 5G-enabled SAGIN that combines speech and text is lacking.

III. A 5G-Enabled SAGIN for an ITS

A SAGIN is a communication network that provides in-vehicle services for the next generation of ITSs. It is usually composed of several ground networks, LEO satellites, UAVs, and high-altitude platforms. Such a network can not only be used in dense urban areas to provide high-data-rate access but can also provide seamless connections with rural areas. As the latest generation of cellular mobile communication technology, 5G technology provides a network infrastructure that can realize the interconnection of humans, machines and things with high speed, low latency and large-scale connectivity.

To implement vehicle SER, we design a 5G-enabled SAGIN, as shown in Fig. 1. In cities, the speech collector on a self-driving vehicle will receive speech in the cabin and transmit it to the cloud via 5G. In remote villages, the speech signals will be transmitted to the cloud via satellite or low-altitude aircraft. Satellites and ground stations together form a space-based network, i.e., a "space network–ground station" network. In this kind of network, the satellites are network

switches or routers with onboard processing capabilities, and there are intersatellite links. The intersatellite links have network routing capabilities, endowing the space segment with network layer functions. The main characteristic of the space segment is that it does not need to be supported by local gateways or ground networks and can provide communication among various user terminals in different satellite coverage areas through the intersatellite links. Low-altitude aircraft and ground stations together form an air-based network, i.e., an "air network–ground station" network. In this network, each low-altitude aircraft is a network switch or router with processing capabilities, and there are interdevice links between the aircraft. These interdevice links have network routing capabilities, endowing the air segment with network layer functions. Similar to the space segment, the air segment does not need to be supported by local gateways or ground networks and can provide communication among various user terminals in different aircraft coverage areas through the interdevice links.

A 5G system adopts a three-level network architecture consisting of distributed units (DUs), central units (CUs), and a 5G core (5GC) network. One or more DUs and one CU together constitute a gNB (a 5G base station). There are many functional segmentation schemes between CUs and DUs that can adapt to different communication scenarios and different communication requirements. For space-based and air-based networks, depending on the relative positions of the satellite and the CU and DU in the 5G base station, there are three feasible types of nonterrestrial 5G network structures based on satellite communication. The first is the transparent satellite forwarding structure, in which the user terminal and the 5G base station are located on the ground and the satellite is used only as a relay. The implementation of this structure is relatively simple. However, it needs a 5G NR Uu interface to be satellite friendly, that is, to be able to adapt to the satellite's long delay and position fluctuations; therefore, this architecture needs to be compatible with ordinary ground 5G New Radio (NR). The second network architecture relies on an in-satellite complete 5G base station to endow a communication satellite with all the functions of a 5G base station, allowing the satellite to perform wide-area signal forwarding and low-latency processing between the 5G base station DU and CU. In this structure, it is also necessary to ensure that the 5G NR connection between the user equipment (UE) and the satellite is friendly to the satellite channel, and since the satellite needs to perform the complete functions of a base station, it is also necessary to ensure that the communication satellite has sufficient signal processing capabilities. The main difficulty in this architecture is determining the functions and service processing capabilities of the on-board base station. The third type of structure is known as the in-satellite DU mode, with a network architecture between those of the first and second types, as shown in Fig. 2. In this structure, only the DU part of the 5G base station is carried on the communication satellite, while the CU is still located on the ground. An NR air interface is adopted between the satellite DU, the satellite station, and the UE, and an F1 interface is adopted between the satellite DU, the satellite gateway station, and the CU.

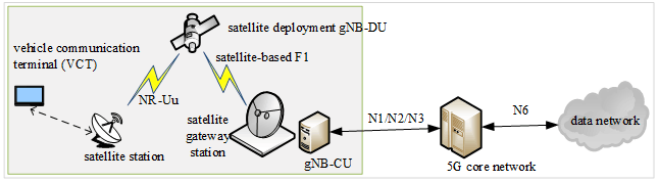


Fig. 2. Architecture of 5G-enabled satellite deployment.

Compared with the second architecture, this architecture can reduce the processing capacity requirements of the onboard base station and is also suitable for positioning satellite gateways and CUs as service convergence points, but it requires satellite channel adaptation for both the F1 and NR interfaces. In our network design, we adopt the third structure, as shown in Fig. 2. In this article, the air-based network architecture is designed using the same approach as that of the space-based network.

Next, we will discuss the communication performance. For remote areas, we design a 5G SAGIN to provide in-vehicle SER. Since the communication link between the SAGIN and the cloud is long, to improve the communication performance for vehicle SER, we apply the following measures. First, for both the space-based and air-based segments, the in-satellite or in-low-altitude-aircraft DU mode is adopted. Accordingly, an NR air interface is used between each satellite DU and UE, and an F1 interface is used between each satellite DU and ground station CU. The NR air interfaces and F1 interfaces are satellite links with 5G characteristics, and their communication performance is much higher than that of traditional satellite links. Second, for both the space-based and air-based networks, we add edge devices to the ground stations to process on-board SER; that is, we sink the work that is originally to be processed in the cloud to the edge devices instead to shorten the communication links and improve communication efficiency.

IV. SER for Autonomous Vehicles in a 5G-Enabled Space–Air–Ground Integrated ITS

For a user in a self-driving car, we can use AI to recognize the user's emotions and predict actions, give reminders or intervene in the user's behavior. However, most existing AI methods use single-modal emotion recognition for in-vehicle users based on a mode such as voice, text, or facial images, and their accuracy is often low. To improve the recognition rate for emotions in speech, we choose to merge the information contained in both the voice signal and the spoken text to predict the emotion class. This method analyzes both audio and spoken content to enable more comprehensive use of the information in the data than is possible with models that focus only on voice data. The recognition of speech emotions based on both voice data and spoken content is expected to be widely used in future in-vehicle systems; for example, in an autonomous vehicle, the vehicle's condition and the driver's fatigue level could be judged by collecting audio signals from both the vehicle and the driver as well as the driver's spoken content, with the aim of reducing the

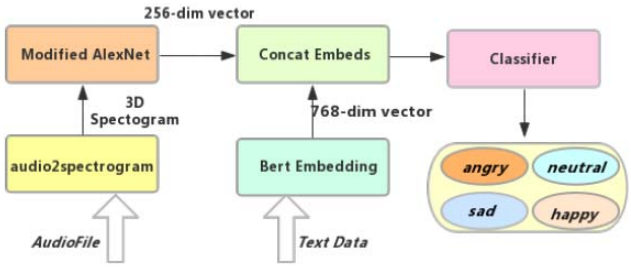


Fig. 3. SER framework.

occurrence of traffic accidents. Therefore, we design a voice- and text-based multimodal emotion recognition method for use in autonomous vehicles. The overall framework is shown in Fig. 3.

As shown in Fig. 3, first, acoustic modeling and text modeling are performed based on the collected speech data and corresponding text information. For speech modeling, the speech signal data are first converted into a spectrogram, which is then used as the input data for acoustic modeling; finally, the high-level features of the acoustic model are obtained through an AlexNet network model. For text modeling, the corresponding text information is first sent to the BERT model, and the high-level features of the corresponding text model are output. Subsequently, the high-level features output by the acoustic model and by the corresponding text model are cascaded to obtain fused features. Finally, the fused features are sent to a softmax classifier for classification.

A. Preprocessing of the Data

Our method is based on the public Interactive Emotional Dyadic Motion Capture (IEMOCAP) data set. The purpose of data preprocessing is to prepare a large number of training samples for model training. The whole process can be divided into two parts: the data preprocessing for the acoustic model and the data preprocessing for the corresponding text model.

In the data preprocessing for the acoustic model, the speech signal is converted into a corresponding spectrogram. First, the one-dimensional speech signal is pre-emphasized, framed, and windowed. Pre-emphasis refers to compensating the high-frequency part of the speech signal that is suppressed by the pronunciation system to eliminate the effects of the vocal cords and lips during the vocalization process. The high-frequency formant can be highlighted through multiplication by a factor in the frequency domain. This factor is positively correlated with the frequency, so the amplitude at higher frequencies will be enhanced. Specifically, the speech signal is passed through a high-pass filter of the form $H(z) = 1 - \mu z^{-1}$. The specific implementation is as shown in Eq. 1:

$$s_n' = s_n - \mu * s_{n-1}, \tag{1}$$

where s_n and s_{n-1} are the speech signals at the current time and the previous time, respectively, and the value of μ ranges from 0.90 to 1.00 and is generally 0.97. Framing refers to dividing the speech signal after pre-emphasis processing into speech segments of the same length, generally between 20 and 40 ms. To prevent the dynamic information of the speech signal from being lost during framing, an overlapping region is included between adjacent speech frames. Each speech frame contains 1024 sample points, and the length of a single frame is approximately 22 ms. Windowing refers to using a Hamming window to smooth the signal after frame processing. Compared with the use of a rectangular window function, the use of a Hamming window weakens the sidelobe size and spectrum leakage after the fast Fourier transform (FFT) operation and enhances the continuity at both ends of the signal.

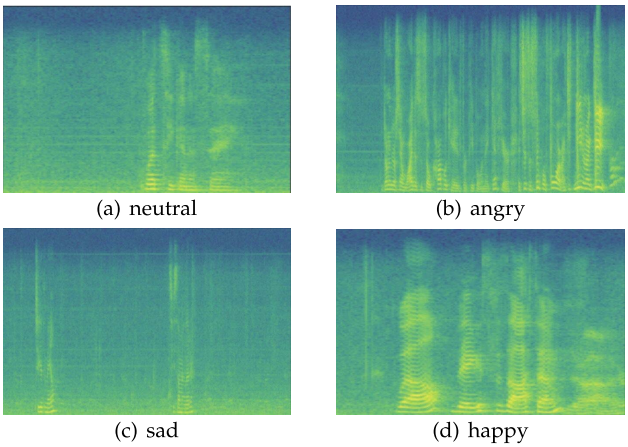


Fig. 4. Speech spectrograms for emotion recognition in autonomous vehicles.

After pre-emphasis, framing, and windowing, each frame of the one-dimensional speech data is subjected to a short-time Fourier transform (STFT), in which a window function is introduced into the Fourier transform. The local features of the signal can then be analyzed. The specific implementation is as shown in Eq. 2:

$$S_i(k) = \sum_{n=1}^{N} s_i(n) e^{(-j2\pi kn)/N} \quad 1 \le k \le K, \tag{2}$$

where N is the number of sample points, K is the length of the discrete Fourier transform, $s_i(n)$ denotes the number of sample points contained in each frame, i is an index identifying the frame, and n is an index for the sample points in one frame.

The third step is to obtain the power spectrum modulo the square of the spectrum from the second step and to convert the time-domain signal into an energy distribution in the frequency domain. The specific implementation is shown in Eq. 3:

$$P_i(k) = \frac{1}{N} |S_i(k)^2|, \tag{3}$$

where N is the number of sample points and $S_i(k)$ is the STFT of frame i of the signal.

Finally, the spectrogram is obtained, as shown in Fig. 4.

Next, the data preprocessing for the corresponding text model is presented.

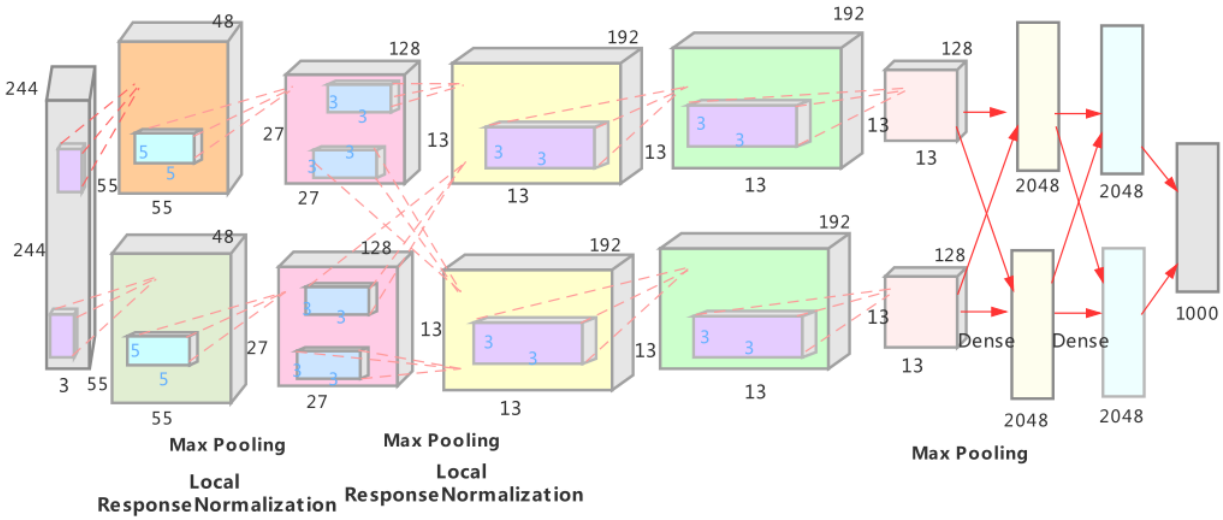


Fig. 5. Structure of the AlexNet network.

First, a word segmentation tool is applied.

Second, the word segmentation results are converted into an array. The bag-of-words features are obtained as a Numerical Python (NumPy) array. The count of each code is determined from the density expression of the bag-of-words model. However, one disadvantage of the bag-of-words model is that it records only the number of instances of each segmented word and ignores the word sequence in the input text. Therefore, we need to optimize the bag-of-words model.

Accordingly, the next step is to perform further optimization. In the term frequency–inverse document frequency (TF-IDF) method, the higher the frequency of a word is, the greater its weight. Hence, this method can be used for classification.

For a particular word i in document d_j, its importance $tf_{i,j}$ can be expressed as shown in Eq. 4:

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}. \tag{4}$$

Here, $n_{i,j}$ is the number of occurrences of the word in file d_j, and the denominator is the sum of the number of occurrences of all words in file d_j.

The IDF for a particular word is calculated as shown in Eq. 5:

$$idf_i = log\frac{|D|}{\left|\left\{j : t_i \epsilon d_j\right\}\right|}. \tag{5}$$

Here, $|D|$ is the total number of files in the corpus. $\left|\left\{j : t_i \epsilon d_j\right\}\right|$ is the number of documents containing the term t_i (the number of documents in $\left\{n_{i,j} \neq 0\right\}$); however, if the term does not appear in the corpus, this will cause the dividend to be zero, so in general, $1 + \left|\left\{j : t_i \epsilon d_j\right\}\right|$ is used instead.

Both a higher frequency of a word within a document and a lower frequency of occurrence in different documents will produce a higher TF-IDF weight. Therefore, the TF-IDF metric, defined as shown in Eq. 6, can be used to filter out common words while retaining important words:

$$TF - IDF = TF * IDF \tag{6}$$

After the TF-IDF is calculated and standardized, the corresponding word feature vector of each text can be used to represent the characteristics of that text for classification or cluster analysis.

B. Selection of the Network Model

We choose AlexNet to model the acoustic data. AlexNet has many advantages in overcoming the problem that traditional machine learning methods cannot effectively extract features to achieve good classification results. First, the rectified linear unit (ReLU) activation function is used for CNN activation in AlexNet. Its effect is much better than that of the traditional sigmoid activation function. It successfully solves the problem of gradient disappearance as the network depth increases and thus can accelerate the training speed. Second, the use of dropout in AlexNet can be avoided. The overfitting phenomenon allows the network to accurately and effectively extract features. Finally, the number of parameters in AlexNet is small, making it easy to train such that it achieves good results on existing hardware devices. The structure of this network is shown in Fig. 5.

In AlexNet, the ReLU function is introduced to solve the problem of slow training convergence caused by sigmoid gradient saturation.

The formula for the ReLU function is given in Eq. 7:

$$f\left(y_i\right) = \begin{cases} y_i, & y_i > 0, \\ 0, & y_i \leq 0. \end{cases} \tag{7}$$

The ReLU function is also plotted in Fig. 6.

The ReLU activation function is a piecewise linear function. If the input is less than or equal to 0, the output is 0; if the input is greater than 0, the output is identical to the input. Compared with the sigmoid function, the ReLU function has the following advantages: The sigmoid function

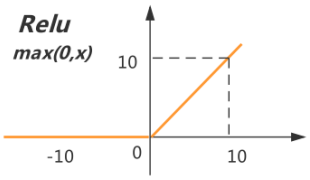


Fig. 6. ReLU function.

is computationally expensive. Its forward propagation involves an exponential operation and a reciprocal operation, whereas the ReLU function has a linear output; in BP, the sigmoid function involves an exponential operation, while the ReLU function produces an output with a derivative that is always 1. Regarding the gradient saturation problem, ReLU activation will cause the outputs of some neurons to be equal to 0, leading to network sparsity and reducing the interdependence of the parameters, which in turn reduces the occurrence of overfitting.

The pooling operation used in AlexNet has some overlap, meaning that the step length is smaller than the window length. Specifically, the pooling window used in AlexNet is a 3×3 square, and the step size is 2. This overlapping pooling operation can also suppress overfitting to a certain extent.

Because the data distribution is nonlinear but the calculations in a neural network are linear, the activation function is introduced to nonlinearly map the output of the network to strengthen its learning ability. Traditional activation functions, such as the sigmoid and tanh functions, map the output to $(0, 1)$ or $(-1, 1)$ so that the gradient will disappear when the calculations are propagated back through the network. In contrast, the ReLU activation function has the characteristic that the derivative of its output is 1, which can be successfully applied for the training of deep networks. There is no limit on the maximum output value of the ReLU function. Therefore, the results obtained with ReLU activation must be normalized; that is, local response normalization (LRN) must be performed. The method used for LRN is given in Eq. 8:

$$b^i_{(x,y)} = \frac{a^i_{(x,y)}}{\left(k + \alpha \sum_{j=max(0,i-2/2)}^{min(N-1,i+n/2)} \left(a^j_{(x,y)}\right)^2\right)^{\beta}}. \tag{8}$$

Here, $a^i_{(x,y)}$ represents the ReLU output at position (x, y) of the i-th kernel, n represents the number of neighbors of $a^i_{(x,y)}$, and N represents the total number of kernels. $b^i_{(x,y)}$ represents the LRN result. Thus, the output ReLU result is locally normalized with respect to the neighbors within a certain range around it.

The main reason for introducing dropout into a network is to prevent overfitting. To obtain the characteristics of an ensemble, with dropout, the training and prediction processes of the neural network must be modified. The formulas for a neural network without dropout are as follows:

$$z^{l+1}_i = w^{l+1}_i y^l + b^{l+1}_i, \tag{9}$$

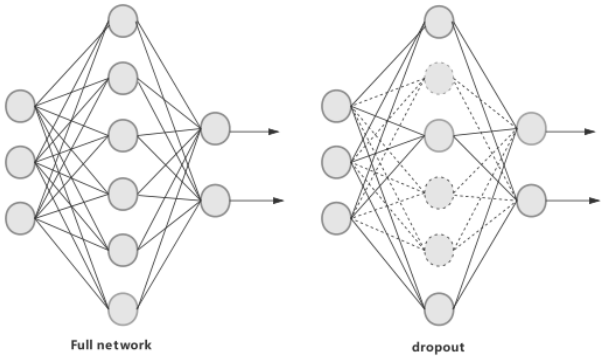


Fig. 7. Full network structure with dropout.

$$y^{l+1}_i = f\left(z^{l+1}_i\right). \tag{10}$$

Here, y^l is the input, w^{l+1}_i is the weight, b^{l+1}_i is the bias, and f is the activation function.

The formulas for a neural network with dropout are given in Eqs. 11 to 15:

$$r^j_j \sim Bernoulli\,(p), \tag{11}$$
$$\widetilde{y}^l = r^l * y^l, \tag{12}$$
$$\widetilde{y}^l = r^l * y^l, \tag{13}$$
$$z^{l+1}_i = w^{l+1}_i \widetilde{y}^l + b^{l+1}_i, \tag{14}$$
$$y^{l+1}_i = f\left(z^{l+1}_i\right). \tag{15}$$

Here, r^l obeys a Bernoulli distribution, y^l is the input, $\widetilde{y}^l$ is the input of the neural network, w^{l+1}_i is the weight, b^{l+1}_i is the bias, and f is the activation function.

Dropout means that during forward propagation, some neurons are discarded in each layer; this can make the model more general, similar to finding the optimal solution. When one step of iteration falls into a local optimum, in the next batch, the iterative process can continue, making it possible to find the globally optimal solution again. Thus, multiple search processes can help to better avoid the model falling into a locally optimal solution. During the training process, in each cycle, some neurons are randomly selected to be temporarily hidden (set to 0) while the neural network training and optimization processes are performed. Then, in the next cycle, a different set of neurons are hidden. This process continues until the end of training, as shown in Fig. 7.

Dropout is officially incorporated into AlexNet, and it is one of the important components of a neural network used to strengthen the learning ability of the network. With dropout, the network structure is different in each training cycle. One way to effectively reduce overfitting is to combine multiple models; adding dropout to a network can allow this model combination effect to be achieved with only twice the training time, making it an efficient means of improving the network performance.

We choose the Bidirectional Encoder Representations from Transformers (BERT) network to model the corresponding text of the input speech signal. The advantages of the BERT

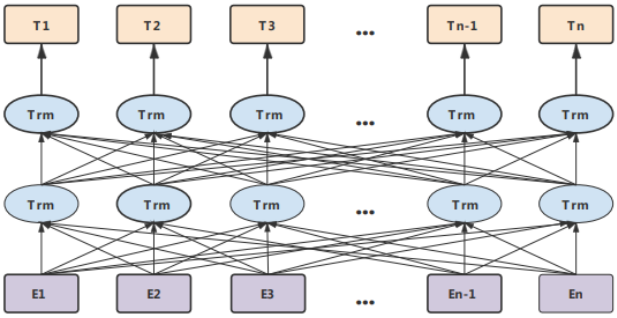


Fig. 8. BERT model structure.

model are as follows: First, BERT is based on state-of-the-art work on pretraining and context-sensitive language representation. Second, in contrast to previous models, BERT is a deep, two-way, unsupervised language representation model that uses only an unlabeled text corpus for pretraining. Third, 11 natural language processing (NLP) tasks can be solved through training and fine-tuning BERT. Compared with recurrent neural networks (RNNs), it is more efficient and can capture longer-distance dependencies. Finally, in contrast to a pretrained model alone, it truly captures bidirectional context information.

As an alternative to Word2Vec, the essence of BERT is to learn good feature representations of words by means of a self-supervised learning method based on a large-scale corpus (where the term "self-supervised learning" refers to supervised learning based on data that have not been manually labeled). In many works on NLP, the word features used for specific tasks are feature representations obtained through BERT. Therefore, as a powerful pretrained model, after some fine-tuning, BERT can be used as the feature extraction part of an NLP network.

The network architecture of BERT is a multilayer Transformer structure. In this architecture, an attention mechanism is used to convert the distance between two words in any positions to 1, thereby effectively solving the thorny long-term dependence problem in NLP. The BERT network architecture is shown in Fig. 8.

A Transformer has an encoder–decoder structure that is formed by stacking several encoders and decoders. The Transformer structure is shown in Fig. 9.

The left part of Fig. 9 shows the encoder, which is composed of a multihead attention mechanism and a fully connected layer, which is used to convert the input corpus into feature vectors. The right part shows the decoder. Its input consists of the output of the encoder and the predicted result. The decoder is composed of a masked multihead attention mechanism, a multihead attention mechanism, and a fully connected layer to output the conditional probability of the result. $E_1, E_2, \ldots, E_N$ represents the input values, and vectorized representations of words can be obtained through a bidirectional Transformer encoder. The Transformer model is a Seq2Seq model based on self-attention. Seq2Seq also has an encoder–decoder structure. The function of the encoder is to encode an input sequence of variable length. The decoder,

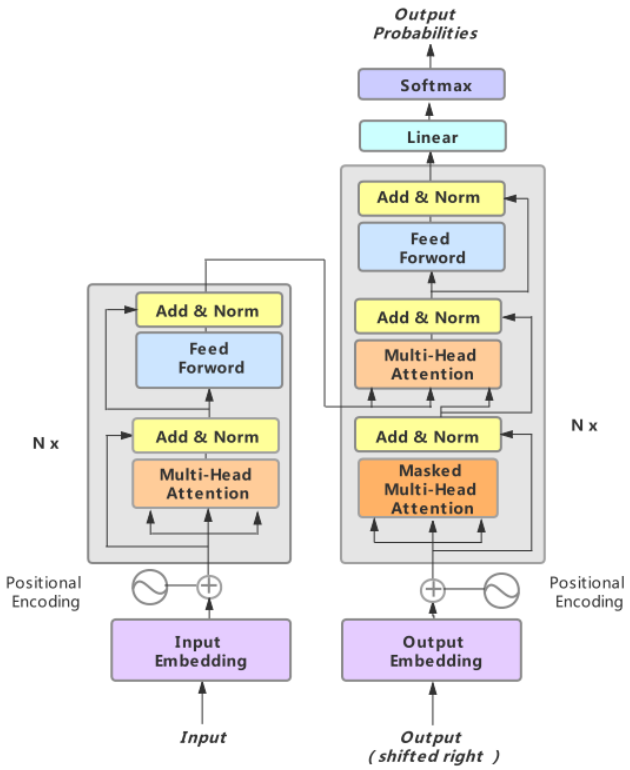


Fig. 9. Transformer model structure.

in turn. generates a sequence of variable length based on the semantic vector trained by the network.

C. Training of a Multimodal Emotion Recognition Model Based on Speech and Text

Once the network model has been selected, network training must be performed. The entire training algorithm can be divided into five steps, as given below, and the specific algorithm flow is shown in Algorithm 1.

Step 1: Calculate the output $\hat{\mathbf{y}}_k$ value for the current sample after the ReLU activation function through forward propagation.

Step 2: Starting from the output neuron, calculate the gradient of the next hidden layer neuron and the weight and bias after the hidden layer neuron. Continue to backpropagate the gradient of the weight and bias before the hidden layer.

Step 3: For each parameter, obtain the partial derivative via the chain rule. The corresponding partial derivative can be obtained from the corresponding gradients g_j and e_h, as calculated above.

Step 4: Update the weights, thresholds and biases by using the learning rate in combination with the partial derivatives to obtain $\Delta\omega_{hj}$, $\Delta\upsilon_{ih}$, $\Delta\theta_j$ and $\Delta\gamma_h$.

Step 5: Terminate the calculation when the parameters that minimize the loss function are found.

V. Experimental Design and Analysis

A. Experimental Environment and Settings

In this experiment, PyTorch's deep learning framework was used to perform SER. The model was trained using Nvidia's

Algorithm 1 Training Algorithm

input: Data $D = \{(\mathbf{x}_k, \mathbf{y}_k)\}_{k=1}^{m}$; Learning rate η

output: Multimodal emotion recognition model determined by connection weights or thresholds

function $F(D, \eta)$

Randomly initialize all connection weights and thresholds in the network within the range $(0, 1)$

repeat

for all $(\mathbf{x}_k, \mathbf{y}_k) \in D$ **do**

Calculate the current sample output $\hat{\mathbf{y}}_k = f(\beta_j - \theta_j)$

Calculate the gradient of the output neuron

$$\begin{aligned} g_j &= -\frac{\partial E_k}{\partial \hat{y}^k} \cdot \frac{\partial \hat{y}^k}{\partial \beta_j} \\ &= -(\hat{y}_j^k - y_j^k) f'(\beta_j - \theta_j) \\ &= \hat{y}_j^k (1 - \hat{y}_j^k)(\hat{y}_j^k - y_j^k) \end{aligned}$$

Calculate the hidden neuron gradient terms

$$\begin{aligned} e_h &= -\frac{\partial E_k}{\partial b_h} \cdot \frac{\partial b_h}{\alpha_h} \\ &= -\sum_{j=1}^{\ell} \frac{\partial E_k}{\partial \beta_j} \cdot \frac{\partial \beta_j}{b_h} f'(\alpha_h - \gamma_h) \\ &= \sum_{j=1}^{\ell} w_{hj} g_j f'(\alpha_h - \gamma_h) \\ &= b_h (1 - b_h) \sum_{j=1}^{\ell} w_{hj} g_j \end{aligned}$$

Update weights

$$\begin{aligned} \Delta w_{hj} &= \eta g_j b_h \\ \Delta v_{ih} &= \eta e_h x_i \end{aligned}$$

Update thresholds

$$\begin{aligned} \Delta \theta_j &= -\eta g_j \\ \Delta \gamma_h &= -\eta e_h \end{aligned}$$

end for

until Termination condition reached

end function

GeForce RTX 2080 Ti graphics processing unit (GPU). For text and speech model training, we fine-tuned the pretrained BERT model and the AlexNet model, respectively, using the AdamW optimizer based on weight attenuation, setting lr to 2e-5 and eps to 1e-8 and using learning rate preheating, which means that the number of training steps was equal to the number of training set samples multiplied by the epoch. We set the number of epochs to 4 for the training of the text model. For speech model training, the number of epochs was set to 100, the dropout rate was set to 0.5, and a cross-entropy loss function was used. For the use of multimodal speech and text information, we loaded the trained models to extract the speech and text features and used a Concat layer for fusion; in the Adam optimizer, the initial learning rate was set to 0.0001, the learning rate reduction strategy was used, step_size was set to 30, gamma was set to 0.1, and the cross-entropy loss function was used.

TABLE I

COMPARISON OF ACCURACY

Method	Input	WA	UA
AlexNet fine-tuning [41]	Spectrogram	67.9%	57.3%
CNN+LSTM [42]	Spectrogram	68.8%	59.4%
FCN+attention [41]	Spectrogram	70.4%	63.9%
Dual RNN [44]	MFCC&Text	71.8%	/
CNN [43]	MFCC	71.6%	59.9%
Our method	Spectrogram&Text	74.0%	65.4%

B. Experimental Results and Analysis

To prove the effectiveness of the proposed method, it was applied to the IEMOCAP data set, which is a multimodal and multiaction database created by the Sail Laboratory at the University of Southern California. During the preparation of this database, participants first performed impromptu actions or scripted scenes, focusing on eliciting emotional expressions, and audio-visual data were recorded, including video, voice, facial motion capture, and text transcription; then, multiple reviewers processed the collected data and assigned category labels, such as anger, happiness, sadness, and neutrality, and dimension labels, such as valence, activation, and dominance. We used the K-fold method to divide the data into a training set and a test set. For a better comparison, we used the WA and UA as indicators, which are defined as follows:

1) WA: the overall accuracy for all samples in the test set.
2) UA: the average accuracy across all categories.

Table I shows the comparison of our results with the results of previous methods in terms of the WA and UA indicators for the SER task. Among the compared methods, the authors of [41] used a fully convolutional network (FCN) + attention structure that was trained to classify the spectrogram features of speech for emotion recognition, achieving a WA of 70.4% and a UA of 63.9%. In a comparative experiment, a fine-tuned AlexNet model was also used to classify and judge the spectrogram characteristics [41], and the results were 67.9% for the WA and 57.3% for the UA. In [42], the authors used a CNN + LSTM model to extract and analyze spectrogram features, and the result was that the WA reached 68.8%, while the UA was 59.4%. In [43], a CNN was used to classify mel-frequency cepstral coefficient (MFCC) features, and the results were that the WA reached 71.6% and the UA reached 59.9%. The above models all perform emotion recognition based on a single modality, that is, speech. In [44], a dual RNN structure was used to conduct multimodal experiments based on MFCC and text features. The effect was generally preferable to that for a single modality, with a final WA of 71.8%. We have comprehensively analyzed the above experimental results and found that when a CNN model is used alone to recognize the speech spectrogram, the accuracy

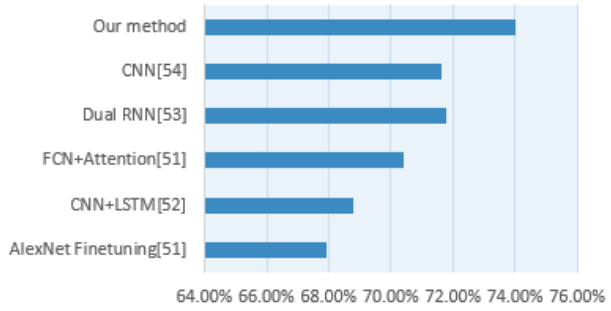


Fig. 10. Comparison of the WA results.

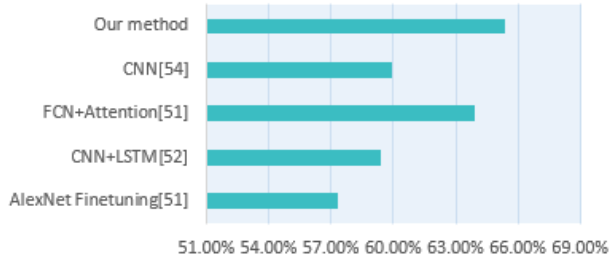


Fig. 11. Comparison of the UA results.

is not high. However, by fine-tuning the BERT preprocessing model, extracting text features, and fusing them with voice spectrogram features, the accuracy of emotion recognition can be significantly improved, with a final WA of 74.0% and a final UA of 65.4%. Visual comparisons of the performance of the different methods are shown in Fig. 10 and Fig. 11. Compared with the above methods, our method achieves the best performance.

As shown in Fig. 12, we obtained the confusion matrix diagram of our AlexNet + BERT multimodal emotion recognition method for the specific classification of 4 emotions: angry, happy, sad and neutral. From this figure, we can observe that the recognition rates for the neutral and happy emotions are relatively high. On the given test set, the correct rates for these emotions can reach approximately 78.6% and 75.4%, respectively. In contrast, the classification accuracies for anger and sadness are lower, reaching only approximately 72.7% and 67.5%, respectively. Further investigation of the data set and our model revealed the following possible reasons for these results. On the one hand, the features extracted by our model for sadness and anger are not sufficient, making the model prone to misclassify these emotions. On the other hand, it may be that the language experts who participated in creating the data set were more accurate in identifying neutral and happy emotions, while the identification of other emotions was more subjective, leading to deviations in their annotation.

In addition to the above performance indicators, receiver operating characteristic (ROC) curves are introduced for the visualization of multiclass performance. A ROC chart depicts the relative trade-off between the true positive rate (TPR) and the false positive rate (FPR). As shown in Fig. 13, the area under the ROC curve (AUC) for each emotion is greater than

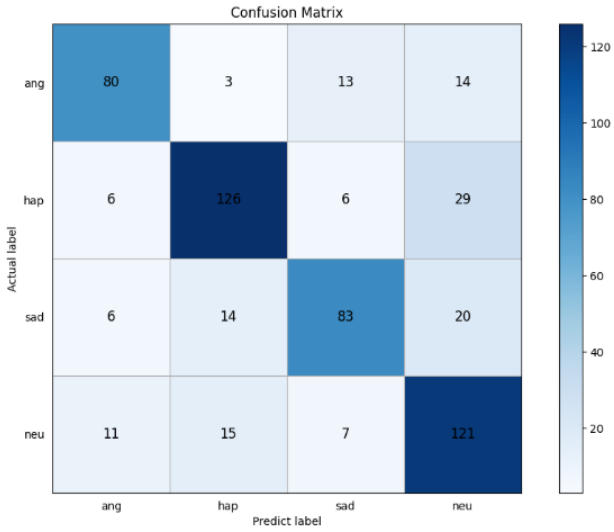


Fig. 12. Confusion matrix of our method.

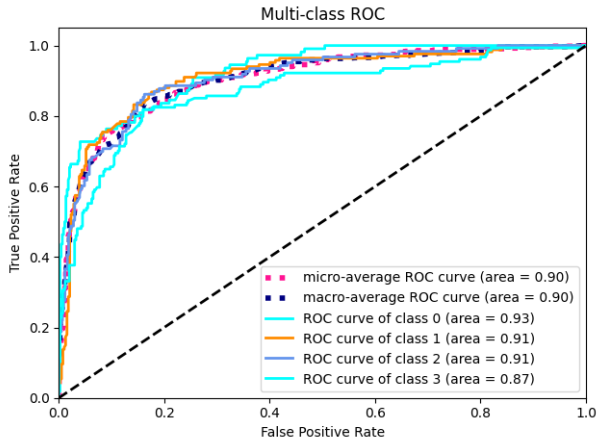


Fig. 13. ROC curve of our method.

0.85. In particular, the AUC for class 0 (meaning anger) reaches 0.93, while that for class 3 (meaning neutrality) is relatively low, reaching only 0.87; nevertheless, the areas under the microaveraged ROC curve and the macroaveraged ROC curve are both equal to 0.9. It can be concluded that the results of the proposed model are stable and can be used for SER.

VI. Conclusion

Next-generation ITSs represent the effective comprehensive application of advanced science and technology (information technology, computer technology, data communication technology, sensor technology, electronic control technology, automatic control theory, operations research, AI, etc.) for transportation and service control. Taking suitable measures during vehicle manufacturing can strengthen the connections between vehicles, roads, and users to support the formation of a comprehensive transportation system that can guarantee safety, improve efficiency, reduce environmental impacts, and save energy. As the foundation for an ITS, the in-vehicle communication network makes use of advanced wireless

communication technology to realize vehicle-to-vehicle and vehicle-to-road communication and the organic integration of traffic participants, vehicles and their environment to improve the safety and efficiency of the transportation system. However, due to deployment, coverage, and capacity issues, ground-based networks alone cannot serve in-vehicle applications well under diverse conditions. At the same time, autonomous vehicles are facing a fundamental shift in human–computer interaction logic. Today's users are eager to interact with their devices in a closer and more conversational manner. However, the accuracy rate of current in-vehicle voice emotion recognition technology is insufficient to meet the needs of a more conversational and personalized interactive experience between users and cars or to enhance safe driving and user immersion. Hence, this paper proposes a multimodal emotion recognition model based on a 5G-enabled SAGIN that combines voice and text to solve these two problems.

Our future work will promote innovation, use 5G technology to improve the service capabilities of SAGINs, and further improve the accuracy of vehicle voice emotion recognition.

REFERENCES

[1] N. Chen, M. Wang, N. Zhang, and X. Shen, "Energy and information management of electric vehicular network: A survey," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 2, pp. 967–997, 2nd Quart., 2020, doi: 10.1109/COMST.2020.2982118.

[2] C. Feng, K. Yu, M. Aloqaily, M. Alazab, Z. Lv, and S. Mumtaz, "Attribute-based encryption with parallel outsourced decryption for edge intelligent IoV," *IEEE Trans. Veh. Technol.*, vol. 69, no. 11, pp. 13784–13795, Nov. 2020, doi: 10.1109/TVT.2020.3027568.

[3] Y. Zhang, R. Wang, M. S. Hossain, M. F. Alhamid, and M. Guizani, "Heterogeneous information network-based content caching in the Internet of Vehicles," *IEEE Trans. Veh. Technol.*, vol. 68, no. 10, pp. 10216–10226, Oct. 2019, doi: 10.1109/COMST.2020.2982118.

[4] Y. Zhang, Y. Li, R. Wang, M. S. Hossain, and H. Lu, "Multi-aspect aware session-based recommendation for intelligent transportation services," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 7, pp. 4696–4705, Jul. 2021, doi: 10.1109/TITS.2020.2990214.

[5] K. Yu, L. Lin, M. Alazab, L. Tan, and B. Gu, "Deep learning-based traffic safety solution for a mixture of autonomous and manual vehicles in a 5G-enabled intelligent transportation system," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 7, pp. 4337–4347, Jul. 2021, doi: 10.1109/TITS.2020.3042504.

[6] Q. Ali, N. Ahmad, A. Malik, G. Ali, and W. Rehman, "Issues, challenges, and research opportunities in intelligent transport system for security and privacy," *Appl. Sci.*, vol. 8, no. 10, p. 1964, Oct. 2018, doi: 10.3390/app8101964.

[7] M. Y. Vadwala and A. Y. Vadwala. (2017). *The User has Requested Enhancement of the Downloaded File*. Accessed: Dec. 27, 2020. [Online]. Available: https://www.researchgate.net/publication/320547133

[8] C. Y. Loh, K. L. Boey, and K. S. Hong, "Speech recognition interactive system for vehicle," in *Proc. IEEE 13th Int. Colloq. Signal Process. Appl.*, Oct. 2017, pp. 85–88, doi: 10.1109/CSPA.2017.8064929.

[9] F. Ding, G. Zhu, M. Alazab, X. Li, and K. Yu, "Deep-learning-empowered digital forensics for edge consumer electronics in 5G HetNets," *IEEE Consum. Electron. Mag.*, early access, Dec. 28, 2020, doi: 10.1109/MCE.2020.3047606.

[10] Z. Guo, K. Yu, Y. Li, G. Srivastava, and J. C.-W. Lin, "Deep learning-embedded social Internet of Things for ambiguity-aware social recommendations," *IEEE Trans. Netw. Sci. Eng.*, early access, Jan. 5, 2021, doi: 10.1109/TNSE.2021.3049262.

[11] G. Lugano, "Virtual assistants and self-driving vehicles," in *Proc. Int. Conf. Telecommun.* Jul. 2017, pp. 1–5, doi: 10.1109/ITST.2017.7972192.

[12] A. McKenna, L. Daniel, and T. Thomas, "The future of autonomous vehicles: Risk with privacy and tracking," Envista Forensics, Deerfield, IL, USA, Tech. Rep., 2018.

[13] L. Zhen, Y. Zhang, K. Yu, N. Kumar, A. Barnawi, and Y. Xie, "Early collision detection for massive random access in satellite-based Internet of Things," *IEEE Trans. Veh. Technol.*, vol. 70, no. 5, pp. 5184–5189, May 2021, doi: 10.1109/TVT.2021.3076015.

[14] Y. Gong, L. Zhang, R. Liu, K. Yu, and G. Srivastava, "Nonlinear MIMO for industrial Internet of Things in cyber–physical systems," *IEEE Trans. Ind. Informat.*, vol. 17, no. 8, pp. 5533–5541, Aug. 2021, doi: 10.1109/TII.2020.3024631.

[15] J. Liu, Y. Shi, Z. M. Fadlullah, and N. Kato, "Space-air-ground integrated network: A survey," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 4, pp. 2714–2741, 4th Quart., 2018.

[16] P. Harar, R. Burget, and M. K. Dutta, "Speech emotion recognition with deep learning," in *Proc. 4th Int. Conf. Signal Process. Integr. Netw. (SPIN)*, Feb. 2017, pp. 137–140, doi: 10.1109/SPIN.2017.8049931.

[17] K. Yu, L. Tan, L. Lin, X. Cheng, Z. Yi, and T. Sato, "Deep-learning-empowered breast cancer auxiliary diagnosis for 5GB remote E-health," *IEEE Wireless Commun.*, vol. 28, no. 3, pp. 54–61, Jun. 2021, doi: 10.1109/MWC.001.2000374.

[18] Z. Guo, K. Yu, A. Jolfaei, A. K. Bashir, A. O. Almagrabi, and N. Kumar, "A fuzzy detection system for rumors through explainable adaptive learning," *IEEE Trans. Fuzzy Syst.*, early access, Jan. 15, 2021, doi: 10.1109/TFUZZ.2021.3052109.

[19] K. Yu, Z. Guo, Y. Shen, W. Wang, J. C.-W. Lin, and T. Sato, "Secure artificial intelligence of things for implicit group recommendations," *IEEE Internet Things J.*, early access, May 12, 2021, doi: 10.1109/JIOT.2021.3079574.

[20] H. Qu, X. Xu, J. Zhao, and P. Yue, "An SDN-based space-air-ground integrated network architecture and controller deployment strategy," in *Proc. IEEE 3rd Int. Conf. Comput. Commun. Eng. Technol. (CCET)*, Aug. 2020, pp. 138–142, doi: 10.1109/CCET50901.2020.9213109.

[21] G. Wang, S. Zhou, S. Zhang, Z. Niu, and X. Shen, "SFC-based service provisioning for reconfigurable space-air-ground integrated networks," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 7, pp. 1478–1489, Oct. 2020, doi: 10.1109/JSAC.2020.2986851.

[22] S. Zhou, G. Wang, S. Zhang, Z. Niu, and X. S. Shen, "Bidirectional mission offloading for agile space-air-ground integrated networks," *IEEE Wireless Commun.*, vol. 26, no. 2, p. 3845, 2019, doi: 10.1109/MWC.2019.1800290.

[23] N. Kato, "Optimizing space-air-ground integrated networks by artificial intelligence," *IEEE Wireless Commun.*, vol. 26, no. 4, pp. 140–147, Aug. 2019, doi: 10.1109/MWC.2018.1800365.

[24] C.-Q. Dai, X. Li, and Q. Chen, "Intelligent coordinated task scheduling in space-air-ground integrated network," in *Proc. 11th Int. Conf. Wireless Commun. Signal Process. (WCSP)*, Oct. 2019, p. 16, doi: 10.1109/WCSP.2019.8928112.

[25] P. Z. Li, "Flow control and scheduling algorithm of air-space-ground integrated networks," *J. Phys. Conf. Ser.*, vol. 1087, no. 2, 2018, Art. no. 022012, doi: 10.1088/1742-6596/1087/2/022012.

[26] M. Veres and M. Moussa, "Deep learning for intelligent transportation systems: A survey of emerging trends," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 8, pp. 3152–3168, Aug. 2020, doi: 10.1109/TITS.2019.2929020.

[27] F. Zhu, Y. Lv, Y. Chen, X. Wang, G. Xiong, and F.-Y. Wang, "Parallel transportation systems: Toward IoT-enabled smart urban traffic control and management," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 10, pp. 4063–4071, Oct. 2020, doi: 10.1109/TITS.2019.2934991.

[28] A. Ferdowsi, U. Challita, and W. Saad, "Deep learning for reliable mobile edge analytics in intelligent transportation systems: An overview," *IEEE Veh. Technol. Mag.*, vol. 14, no. 1, pp. 62–70, Mar. 2019, doi: 10.1109/MVT.2018.2883777.

[29] N. Zhang, S. Zhang, P. Yang, O. Alhussein, W. Zhuang, and X. S. Shen, "Software defined space-air-ground integrated vehicular networks: Challenges and solutions," *IEEE Commun. Mag.*, vol. 55, no. 7, pp. 101–109, Jul. 2017, doi: 10.1109/MCOM.2017.1601156.

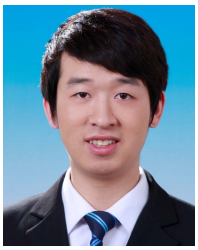
[30] H. Wu *et al.*, "Resource management in space-air-ground integrated vehicular networks: SDN control and AI algorithm design," *IEEE Wireless Commun.*, vol. 27, no. 6, pp. 52–60, Dec. 2020.

[31] C. Zhao, M. Shi, M. Huang, and X. Du, "Authentication scheme based on hashchain for space-air-ground integrated network," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2019, p. 16, doi: 10.1109/ICC.2019.8761821.

[32] C. Feng, B. Liu, Z. Guo, K. Yu, Z. Qin, and K.-K.-R. Choo, "Blockchain-based cross-domain authentication for intelligent 5G-enabled Internet of Drones," *IEEE Internet Things J.*, early access, Sep. 17, 2021, doi: 10.1109/JIOT.2021.3113321.

[33] L. Tan, K. Yu, N. Shi, C. Yang, W. Wei, and H. Lu, "Towards secure and privacy-preserving data sharing for COVID-19 medical records: A blockchain-empowered approach," *IEEE Trans. Netw. Sci. Eng.*, early access, Aug. 4, 2021, doi: 10.1109/TNSE.2021.3101842.
[34] Y. Su, "Research on vehicle control system based on speech recognition technology," in *Proc. 2nd Int. Conf. Robot., Control Autom. Eng.*, Nov. 2019, p. 8892, doi: 10.1145/3372047.3372105.
[35] M. F. M. Idros, A. H. A. Razak, S. Al Junid, A. K. Halim, and N. Khairudin, "Capability of voice recognition system for automatic signal in autonomous vehicle (AV) application," in *Proc. IEEE 5th Int. Conf. Smart Instrum., Meas. Appl. (ICSIMA)*, Nov. 2018, p. 16, doi: 10.1109/ICSIMA.2018.8688755.
[36] A. Mahtab, M. Singh, V. K. Sharma, and A. Kumar, "Anti-collision vehicle with voice recognition," in *Proc. Int. Conf. Electron., Commun. Aerosp. Technol. (ICECA)*, Apr. 2017, pp. 126–129, doi: 10.1109/ICECA.2017.8212778.
[37] S. Wang, J. Cao, K. Sun, and Q. Li, "SIEVE: Secure in-vehicle automatic speech recognition systems," in *Proc. 23rd Int. Symp. Res. Attacks, Intrusions Defenses (RAID)*, 2020, pp. 365–379.
[38] K. Zhang, L. Chen, Y. An, and P. Cui, "A QoE test system for vehicular voice cloud service," *Mobile Netw. Appl.*, vol. 26, pp. 700–715, Dec. 2019, doi: 10.1007/s11036-019-01415-3.
[39] L. Yu, K. Zhang, J. Man, H. Yu, Y. Yao, and L. Chen, "A test system for vehicular speech cloud service," in *Social-Informatics and Telecommunications Engineering* (Lecture Notes of the Institute for Computer Sciences), vol. 295. New York, NY, USA: Springer, Jul. 2019, pp. 346–352, doi: 10.1007/978-3-030-32216-8_33.
[40] S. Amman, J. Huber, F. Charette, B. Richardson, and J. Wheeler, "The impact of microphone location and beamforming on in-vehicle speech recognition," *SAE Int. J. Passeng. Veh.-Electron. Electr. Syst.*, vol. 10, no. 2, pp. 430–434, Mar. 2017, doi: 10.4271/2017-01-1692.
[41] Y. Zhang, J. Du, Z. Wang, J. Zhang, and Y. Tu, "Attention based fully convolutional network for speech emotion recognition," in *Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, Mar. 2019, pp. 1771–1775, doi: 10.23919/APSIPA.2018.8659587.
[42] A. Satt, S. Rozenberg, and R. Hoory, "Efficient emotion recognition from speech using deep learning on spectrograms," in *Proc. Conf. Int. Speech Commun. Assoc.*, 2017, pp. 1089–1093.
[43] S. Tripathi, A. Kumar, A. Ramesh, C. Singh, and P. Yenigalla, "Deep learning based emotion recognition system using speech features and transcriptions," 2019, *arXiv:1906.05681*. [Online]. Available: http://arxiv.org/abs/1906.05681
[44] S. Yoon, S. Byun, and K. Jung, "Multimodal speech emotion recognition using audio and text," in *Proc. IEEE Spoken Lang. Technol. Workshop (SLT)*, Dec. 2018, pp. 112–118, doi: 10.1109/SLT.2018.8639583.

Liang Tan received the B.A. and Ph.D. degrees in computer science from the University of Electronic Science and Technology of China in 2002 and 2007, respectively. He is currently a Professor with the College of Computer Science, Sichuan Normal University, and a Post-Doctoral Research Associate with the Institute of Computing Technology, Chinese Academy of Sciences. His research interests include cloud computing, big data, trusted computing, and network security. For the TPM2.0 Key Migration Protocol, he proposed a migration protocol based on duplication authority. It uses the duplication authority (DA) as an authentication and control center to divide the key migration process into an initialization phase, an authentication and attribute acquisition phase, and a control and execution phase. The DA determines the migration process by the duplicate attributes and types of the migration key and by the handle type of the new parent key. He considered a variety of reasonable combinations of attributes and designed 12 different migration processes before settling on the protocol analyzed and simulated. The results show that the protocol is not only fully compliant with the "TPM-Rev-2.0-Part-1-Architecture-01.38," but also has integrity, confidentiality, and supports authentication.



Keping Yu (Member, IEEE) received the M.E. and Ph.D. degrees from the Graduate School of Global Information and Telecommunication Studies, Waseda University, Tokyo, Japan, in 2012 and 2016, respectively.

He was a Research Associate and a Junior Researcher with the Global Information and Telecommunication Institute, Waseda University, from 2015 to 2019 and from 2019 to 2020, respectively, where he is currently a Researcher. He has hosted and participated in more than ten projects, is involved in many standardization activities organized by ITU-T and ICNRG of IRTF, and has contributed to ITU-T Standards Y.3071 and Supplement 35. He has authored more than 100 publications including papers in prestigious journals/conferences, such as the IEEE WIRELESS COMMUNICATIONS, *IEEE Communications Magazine* (ComMag), *Net-Mag Magazine* (NetMag), IEEE INTERNET OF THINGS JOURNAL (IoTJ), IEEE TRANSACTIONS ON FUZZY SYSTEMS (TFS), IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS (TII), IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS (TITS), IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY (TVT), IEEE TRANSACTIONS ON NETWORK SCIENCE AND ENGINEERING (TNSE), IEEE TRANSACTIONS ON GREEN COMMUNICATIONS AND NETWORKING (TGCN), *IEEE Consumer Electronics Magazine* (CEMag), *IoT Magazine* (IoTMag), ICC, and GLOBECOM. His research interests include smart grids, information-centric networking, the Internet of Things, artificial intelligence, blockchain, and information security. He received the Best Paper Award from ITU Kaleidoscope 2020 and the Student Presentation Award from JSST 2014. He has served as the General Co-Chair and the Publicity Co-Chair for the IEEE VTC2020-Spring 1st EBTSRA Workshop, the General Co-Chair for IEEE ICCC2020 2nd EBTSRA Workshop, the General Co-Chair for IEEE TrustCom2021 3nd EBTSRA Workshop, the Session Chair for IEEE ICCC2020, the TPC Co-Chair for SCML2020, the Local Chair for MONAMI 2020, the Session Co-Chair for CcS2020, and the Session Chair for ITU Kaleidoscope 2016. He has been the Lead Guest Editor of *Sensors*, *Peer-to-Peer Networking and Applications*, *Energies*, *Journal of Internet Technology*, *Journal of Database Management*, *Cluster Computing*, *Journal of Electronic Imaging*, *Control Engineering Practice*, and *Sustainable Energy Technologies and Assessments*; and the Guest Editor of IEEE TRANSACTIONS ON COMPUTATIONAL SOCIAL SYSTEMS, *IEICE Transactions on Information and Systems, Computer Communications*, *IET Intelligent Transport Systems*, *Wireless Communications and Mobile Computing*, *Soft Computing*, and *IET Systems Biology*. He is also an Associate Editor of IEEE OPEN JOURNAL OF VEHICULAR TECHNOLOGY, *Journal of Intelligent Manufacturing*, and *Journal of Circuits, Systems and Computers*.

Long Lin received the bachelor's degree in engineering from Sichuan Normal University, Chengdu, Sichuan, China, in 2018, where he is currently pursuing the master's degree in software engineering. He worked on a digital teaching patent. His research interest includes emotion recognition.

Xiaofan Cheng was born in Neijiang, Sichuan, China, in 1997. He received the bachelor's degree in computer science and technology from Chengdu University in 2019. He is currently pursuing the master's degree in software engineering with Sichuan Normal University. His research interest includes medical image analysis.

Gautam Srivastava (Senior Member, IEEE) received the B.Sc. degree from Briar Cliff University, Sioux City, IA, USA, in 2004, and the M.Sc. and Ph.D. degrees from the University of Victoria, Victoria, BC, Canada, in 2006 and 2011, respectively. He then taught for three years at the Department of Computer Science, University of Victoria, where he was regarded as one of the top undergraduate professors in the Computer Science Course Instruction at the University. In 2014, he joined a Tenure-Track position at Brandon University, Brandon, MB, Canada, where he currently is active in various professional and scholarly activities. He was promoted to the rank as an Associate Professor in January 2018. He is popularly known as an active in research in the field of data mining and big data. In his eight-year academic career, he has published a total of 60 papers in high-impact conferences in many countries and in high-status journals (SCI and SCIE) and has also delivered invited guest lectures on big data, cloud computing, the Internet of Things, and cryptography at many Taiwanese and Czech universities. He received the Best Oral Presenter Award in FSDM 2017 which was held at the National Dong Hwa University (NDHU), Shoufeng, Taiwan, in November 2017. He also has active research projects with other academics in Taiwan, Singapore, Canada, Czech Republic, Poland, and USA. He is constantly looking for collaboration opportunities with foreign professors and students. He is also an editor of several international scientific research journals.

Jerry Chun-Wei Lin (Senior Member, IEEE) is currently a Full Professor with the Department of Computer Science, Electrical Engineering and Mathematical Sciences, Western Norway University of Applied Sciences, Bergen, Norway. He has published more than 400 research articles in refereed journals, such as [IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING (TKDE), IEEE TRANSACTIONS ON CYBERNETICS (TCYB), IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS (TII), IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS (TITS), IEEE TRANSACTIONS ON NETWORK SCIENCE AND ENGINEERING (TNSE), IEEE TRANSACTIONS ON EMERGING TOPICS IN COMPUTATIONAL INTELLIGENCE (TETCI), IEEE SYSTEMS JOURNAL (SysJ), IEEE SENSORS JOURNAL (SensJ), IEEE INTERNET OF THINGS JOURNAL (IoTJ), *ACM Transactions on Knowledge Discovery from Data* (TKDD), *ACM Transactions on Data Science* (TDS), *ACM Transactions on Management Information Systems* (TMIS), *ACM Transactions on Internet Technology* (TOIT), and *ACM Transactions on Intelligent Systems and Technology* (TIST) and international conferences, such as IEEE ICDM**,** IEEE ICDE, PAKDD, and KDD. He has filed and held 33 patents, including three U.S. patents. He is also the Co-Leader of the well-known SPMF project, and also the Founder and the Leader of PPSF Project. He is also a Senior Member of ACM and a fellow of IET. Moreover, he has been awarded as the Most Cited Chinese Researcher respectively in 2018, 2019, and 2020 by Elsevier/Scopus. He is also the Editor-in-Chief of *Data Science and Pattern Recognition* (DSPR) journal, and an Editor/Associate Editor/Guest Editor of many journals, such as IEEE TRANSACTIONS ON FUZZY SYSTEMS (TFS), IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS (TII), *ACM Transactions on Management Information Systems* (TMIS), and *ACM Transactions on Internet Technology* (TOIT).

Wei Wei (Senior Member, IEEE) received the M.S. and Ph.D. degrees from Xian Jiaotong University in 2005 and 2011, respectively. He is currently an Associate Professor with the School of Computer Science and Engineering, Xi'an University of Technology, Xi'an, China. He ran many funded research projects as a principal investigator and a technical member. He has published around 100 research papers in international conferences and journals. His research interests include wireless networks, wireless sensor networks application, image processing, mobile computing, distributed computing, pervasive computing, the Internet of Things, and sensor data clouds. He is also a TPC Member of many conferences and a Regular Reviewer of the IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS (TPDS), the IEEE TRANSACTIONS ON IMAGE PROCESSING (TIP), the IEEE TRANSACTIONS ON MOBILE COMPUTING (TMC), the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS (TWC), and many other Elsevier journals. He is also an Editorial Board Member of *Future Generation Computer Systems* (FGCS), *Ad Hoc and Sensor Wireless Networks: Architectures, Algorithms and Protocols* (AHSWN), *IEICE Transactions on Information and Systems*, and *KSII Transactions on Internet and Information Systems*.