Research on Traffic Acoustic Event Detection Algorithm Based on Model Fusion

Xiaodan Zhang, Ming Li, and Chengwei Huang

Abstract—**Road traffic monitoring is important for intelligent transportation, and researchers have begun to focus on the detection of traffic events using acoustic information. In this paper, we apply model fusion to traffic acoustic event classification. First, an improved, two-channel convolutional neural network (CNN) model is proposed as the weak classifier for constructing the fusion model. The mel-cepstral feature and its first-order and second-order difference are selected as the input features. Six different input features are constructed after signal preprocessing and segmentation. Second, after training six different CNN models, the voting method and support vector machine (SVM) stacking method are used to construct the final fusion model. Experimental results demonstrate that the detection rate of traffic acoustic events reaches 95.1%, which is higher than that of traditional traffic detection algorithms.**

Index Terms—**traffic acoustic event detection, acoustic feature, two-channel convolutional neural network, model fusion.**

I. Introduction

THE detection of traffic states based on acoustic information is an important research direction for intelligent transportation. Compared to existing monitoring techniques, acoustic signal processing and classification techniques have the advantage of low cost and are unaffected by lighting conditions. Especially in the case of tunnels, where there is insufficient light for visual monitoring, acoustic signals have better coverage. Therefore, it is an important supplement to existing monitoring methods. However, when compared with the laboratory environment, the real traffic environment is complicated. For example, a tunnel is a special traffic environment and is very different from an open road environment. Effectively processing traffic acoustic data remains a challenge. In 1998, Henryk Maciejewski et al. [1] studied and designed a classification system based on wavelet[2] and neural networks. The specific recognition model, based on the sound signal, was constructed for four different vehicles, and the results indicated that the recognition accuracy was 73.68%. Audi Ovox et al. applied sound recognition technology to the field of intelligent transportation [3] and used voice recognition technology in the car. Xianglong Luo et al. [4] used empirical mode decomposition (EMD) and a support vector machine (SVM)[5] to identify the vehicle state. In recent years, some scholars have tried to apply convolutional neural networks (CNNs)[6] to recognize sound events[7]. Compared with traditional classifiers, convolutional neural networks have greatly improved recognition rates. The ConvNet model[8] has improved the accuracy by nearly 20% on the Esc-50 database. The LSTM+CNN model proposed by Bae et al. achieved an 84.1% accuracy rate in the DCASE2016 competition[9].

Model fusion is an important subject in the machine learning field. It refers to training several different models and then integrating the whole model in a manner conducive to its intended function. The model fusion algorithm is easy to understand, simple to implement, and achieves better performance, so it is widely used in industry[10]. For classification tasks, the general workflow of model fusion involves generating N weak classifiers and then combining them with some strategy to form the final strong classifier and enhance the recognition performance of the model. The higher the accuracy and diversity of the weak classifiers are, the better the final model fusion effect will be. With an increase in the number of weak classifiers, the error rate of strong classifiers decreases exponentially.

In this paper, an algorithm based on model fusion is proposed for classification of traffic events, and the improved two-channel CNN model is proposed as the weak classifier. This model uses the mel-cepstral feature and its first-order and second-order difference features to train six different CNN models and improves the robustness through fusion. To verify the performance of the proposed algorithm, we collected a total of 974 audio samples for the experiment, including the sounds of five types of events: braking, a door closing, crashing, a horn, and an engine running. The experimental results show that the detection rate of traffic acoustic events reaches 95.1%. The proposed audio surveillance system may be used to monitor traffic accidents and save valuable time during rescue missions. In addition, the system can be embedded in the automatic driving system, and improves the ability of self-driving cars to adapt to the state of traffic.

II. Traffic Acoustic Event Detection Algorithm Based on Model Fusion

A. Features

For traffic acoustic event detection, the mel-cepstral feature and its first-order and second-order differences are selected as the input features. Among many acoustic features, the mel-cepstral feature is widely used in acoustic event classification and emotion recognition [11]. The mel-cepstral feature is a spectral feature that is calculated based on the nonlinear relationship between the human ear's auditory characteristics and audio signal frequency. Fig. 1 shows the comparison of the mel-cepstral features from the five traffic

Manuscript received January 08, 2021; revised June 16, 2021. This work was supported by Central Public-interest Scientific Institution Basal Research Fund (# 2020-9061).

Xiaodan Zhang is an associate research fellow of Research Institute of Highway, Ministry of Transport, Beijing 100088, China(phone: (86)15117986667; fax: (010)623-70567; e-mail: zhangdaqing_925@163.com).

Ming Li is a postgraduate student of School of Information Science and Engineering, Southeast University, Nanjing 210096, China (e-mail: 1047583504@qq.com).

Chengwei Huang is a manager of Jiangsu Intever Energy Technology Co. Ltd., Nanjing 210000, China (e-mail: huangcwx@126.com).

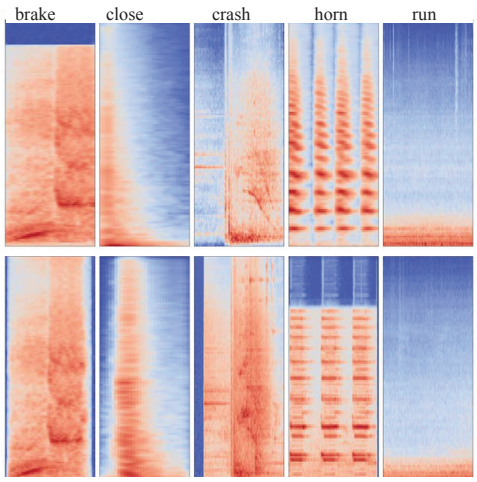

Fig. 1. Mel-cepstral feature of different events

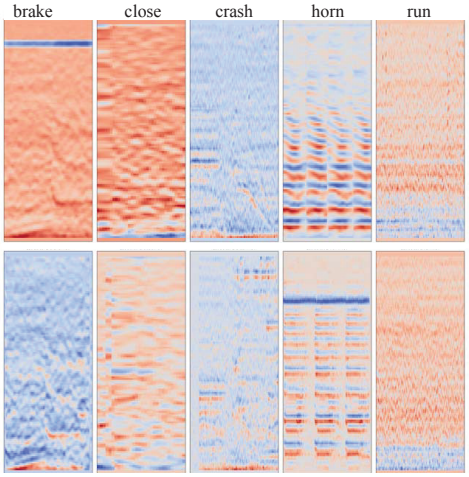

Fig. 2. The first-order differences between mel-cepstral features from different events

acoustic events in the data set. It can be seen from Fig. 1 that the difference between the mel-cepstral features in different categories is obvious, while the difference between the same category is inconspicuous. The standard mel-cepstral feature only reflects the static characteristics of traffic acoustic events. The dynamic characteristics can be described by the difference in mel-cepstral[12], [13]. The difference in mel-cepstral features is further modified to further strip the features and obtain information such as event type changes. Combining the mel-cepstral feature with its differences can improve the recognition performance of the system. Fig. 2 and Fig. 3 show the comparison of the first-order and second-order mel-cepstral differences. For the difference features, the gap between different categories is magnified so the dynamic characteristics are more obvious.

B. Multi-model construction and fusion mode

Model diversity is a key factor that determines the effect of the fusion model, which can be enhanced using the following schemes. First, by increasing the disturbance of the data for a given initial data set, different data can be selected to train different weak classifiers. Second, disturbance of the input features can be added, and different input features can

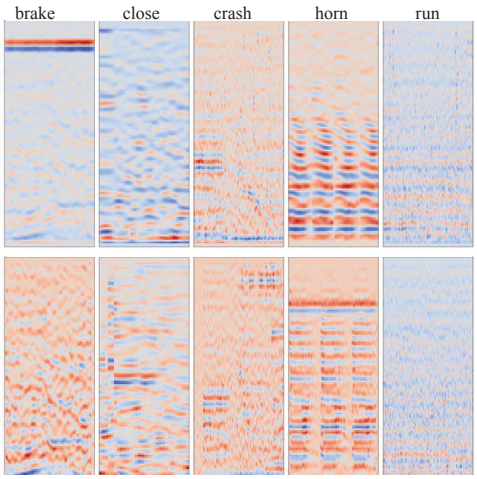

Fig. 3. The second-order differences between mel-cepstral features from different events

be used to train different models. Third, disturbance of the algorithm parameters should be considered: basic algorithms generally have parameters to be set, such as the number of hidden layer neurons and the learning rate of the neural network. By setting different parameters, weak classifiers with large differences can be produced[14].

This paper adopts the second scheme, and different feature extraction methods are adopted to construct a different model. By combining the audio channel separation and audio segmentation methods, different input features are obtained. The audio channel separation method includes non-separation and harmonic-percussive source separation. The audio cutting method includes non-cutting, overlapping cutting and non-overlapping cutting. So, six different features are obtained.

The harmonic component of audio data determines the timbre and can be used to distinguish different sound objects. When an object vibrates and makes noise, it emits many waves of different frequencies, or in other words, harmonics. Different audio data contain different frequencies, loudness and distribution of harmonics, so various acoustic events can be distinguished. Separating the harmonic source from the percussive source and obtaining mel-cepstral features can better judge the tone color of the unique sounding object in traffic acoustic events. By separating the harmonic source and the percussive source, the mel-cepstral feature and its difference can be obtained.

Audio data last for 3 seconds and are divided using three methods: non-cutting, non-overlapping cutting and overlapping cutting. The non-cutting method uses the original audio to extract the mel-cepstral features and their differences; the non-overlapping cutting method divides the audio data into three segments according to 1s, then extracts the features and stacks them to obtain three-channel features for use as input features; the overlapping cutting method divides the speech into two segments according to 1s; and the overlapping cutting method divides data into 1s and then stacks the features to obtain two-channel features for use as input features. As shown in Fig. 4, the above methods are adopted to separate and cut the audio data. The combination of the results can produce six different input features; model fusion is then carried out after training the model separately.

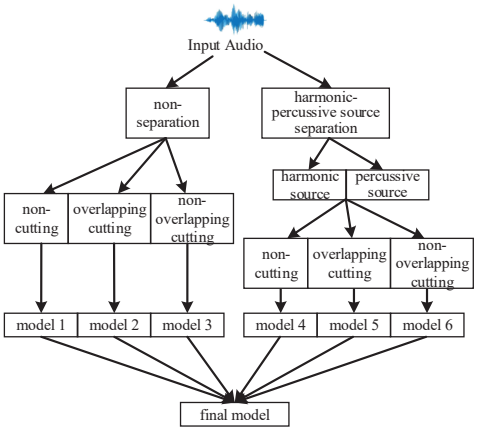

Fig. 4. Model fusion framework

The voting and stacking methods, based on the SVM, are adopted to fuse the output from the six CNN models. The voting method votes on the results from the six models, and the category with the highest number of votes is the final classification result. The stacking method, using an SVM as a strong classifier, uses the output results from the 6 models as the input characteristics for the SVM to train the new model and considers the SVM classification results as the final results.

C. Improved CNN model

Convolutional neural networks (CNNs) are neural networks designed to handle high-dimensional data. Compared with traditional neural networks, CNNs have a greatly improved structure. First, the convolutional layer and the pooling layer are introduced to achieve local and hierarchical feature extraction. Second, weight sharing is used to reduce the difficulty of network optimization. Finally, the ReLU activation function is used to solve the gradient disappearance problem.

Traditional CNNs have only one convolutional channel. For harmonic-percussive source separation, an improved, two-channel CNN model is proposed to process the two-channel input feature by referring to the VGG model[15]. The VGG model uses a continuous 3x3 convolution kernel to replace the larger convolution kernel, which improves the network depth when the kernels have the same perception field. For a given perception field, the small convolution kernel is better than the large convolution kernel because multiple, nonlinear layers can increase the network depth to improve the complexity of the model, and there are fewer parameters. As shown in Fig. 5, separate convolutional layers are used to process the input feature from each channel before the fully connected layers. Each convolutional channel has three convolution blocks, and each convolution block is followed by the max-pooling layer. The convolution block contains two convolution layers followed by the ReLU activation function[16]. A batch normalization layer is added at the end of the convolution block. The number of channels in the convolutional layers is shown in Fig. 5. The output feature map from the convolution layers is spliced into one dimension and input into the fully connected layers. The fully connected layers have two layers, and the number of nodes is [2048, 1024]. The output layer is the softmax layer,

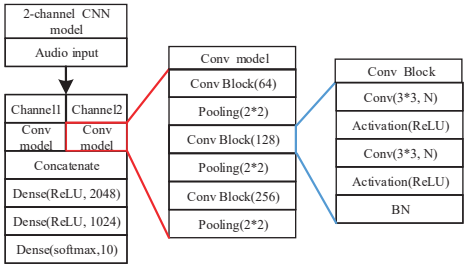

Fig. 5. Two-channel CNN model

which outputs the prediction probability for each category. The models cost function is the cross-entropy loss function. The two-channel CNN model uses different convolutional channels to process the features from different channels to retain each channels characteristics. Therefore, the connection and difference between the two channels can be better explored, which improves the recognition rate of the traffic acoustic event detection system.

For non-separation, the single-channel CNN model is used as a weak classifier. In the single-channel CNN model, the convolutional channel parameters are set in accordance with the parameter settings of one of the two channels, and the parameter settings for the whole connection layer are also consistent with these parameter settings.

III. EXPERIMENT

A. Parameter setting

The data set collected in this experiment contains five common traffic acoustic events: braking, a door closing, a crash, a horn and the sound of an engine running. The data set contains 974 audio data, including 70 for braking, 50 for a door closing, 332 for crashes, 124 for horns, and 398 for engine noise. The effect of the model was evaluated using 5-fold cross-validation.

Audio data last for three seconds, and the sampling frequency is 16 kHz. For harmonic-percussive source separation, the mel-cepstral feature and its difference for each channel are extracted. For non-separation, the input feature is extracted from the audio data directly. First, the audio data are divided into frames, with a frame length of 50 ms and a frame shift of 20 ms. Second, an FFT is calculated for each data frame, and the number of FFT points is 1024. Third, the log mel-cepstral feature is obtained using mel-filter banks with 80 subband filters. Fourth, the first-order and second-order mel-cepstral feature differences are calculated and the multi-channel input feature is obtained. The features obtained from the three different cutting methods have different sizes. For non-cutting, the final input feature size is (149, 80, 3). For non-overlapping cutting, the size of the mel-cepstral feature for each 1s is (49, 80, 3), and the final input feature size obtained by stacking the feature of each 1s is (49, 80, 9). For overlapping cutting, the size of the mel-cepstral feature for each 2s is (99, 80, 3), and the final input feature size obtained by stacking the feature of each 2s is (99, 80, 6).

The experimental equipment includes an Nvidia Tesla K20C graphics card (5 GB memory) and an Intel Xeon CPU e5-2675 v3 processor. The Adam optimizer is used; the exponential decay rate of the first-order moment estimation

TABLE I
THE AVERAGE ACCURACY OF THE SIX WEAK CLASSIFIERS

Scene label	M1/%	M2/%	M3/%	M4/%	M5/%	M6/%
brake	53.3	64.0	68.4	92.8	72.2	72.7
close	57.1	70.6	43.6	85.7	45.0	60.0
crash	87.5	95.8	87.7	88.0	81.6	90.2
horn	84.6	78.8	83.3	80.4	73.8	80.6
run	98.8	99.2	96.7	97.5	88.8	99.1
average	88.2	91.1	87.4	90.8	87.1	89.8

TABLE II
THE EXPERIMENTAL RESULTS OF MODEL FUSION

Model	Accuracy/%
M1	88.2
M2	91.1
M3	87.4
M4	90.8
M5	89.1
M6	89.8
voting fusion method	93.4
stacking fusion method with SVM	95.1

is 0.9 and that of the second-order moment estimation is 0.999. The learning rate is 1e-3. The batch size is 64 and the epoch is 100.

B. Validation of the two-channel CNN model

Table I shows the experimental results of the six weak classifiers, and the settings of models 1-6(M1-M6) are shown in Fig. 4. The results verify that the two-channel CNN model has a positive effect on improving accuracy. For non-cutting, the average accuracy of the one-channel CNN model is 88.2%, and that of the two-channel CNN model is 91.1%, which is 2.9% higher. For non-overlapping cutting, the average accuracy of the one-channel CNN model is 87.4%, and that of the two-channel CNN model is 90.8%, which is 3.4% higher. For overlapping cutting, the average accuracy of the one-channel CNN model is 87.1%, and that of the two-channel CNN model is 89.8%, which is 2.7% higher. The two-channel CNN model significantly improves the accuracy of the three different cutting methods. Overall, the accuracy of the two-channel CNN model is 2%-4% higher than that of the one-channel CNN model. Harmonic-percussive source separation can better identify the timbre of sound objects and is conducive to mining information.

C. Validation of model fusion

Table II integrates the above experimental results and presents the experimental results of model fusion. As shown in the table, the positive effect on accuracy is obvious. The accuracy of the vote method for model fusion is 93.4%, and the accuracy of the stacking method for model fusion, based on SVMs, is 95.1%. These accuracies are 2.3% and 4.0% higher than the maximum accuracy of the single model, respectively. Fig. 6 and Fig. 7 show the confusion matrixes of the voting method for model fusion and the stacking method for model fusion with SVMs; these figures also comprehensively show the classification for each category. It can be seen from the figure that in some categories, the

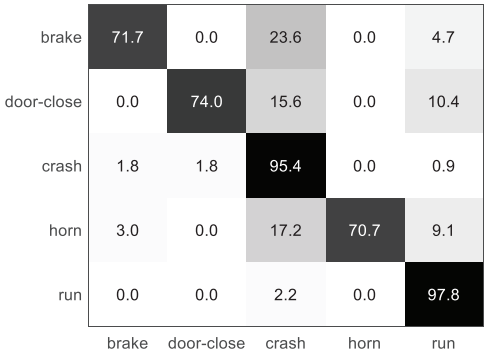

Fig. 6. The confusion matrix of voting method model fusion

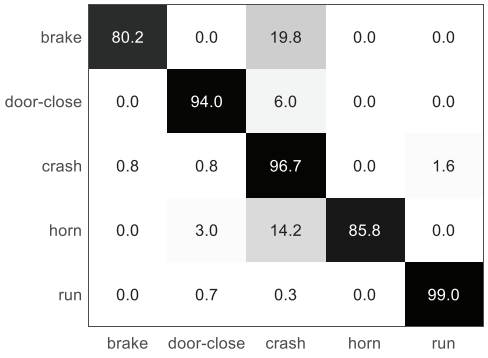

Fig. 7. The confusion matrix of stacking method model fusion with SVM

model fusion algorithm achieves quite a high accuracy. For example, in the traffic acoustic event of engine noise, the two fusion methods have the highest recognition rate, with the voting method for model fusion reaching an accuracy of 97.8% and the stacking method for model fusion with SVMs reaching an accuracy of 99.0%. In addition, in the traffic acoustic event of a crash, the accuracy for both methods is over 85%. These results indicate that the mel-cepstral feature and its difference can accurately reflect the acoustic characteristics of the engine noise and crash events, and the CNN model can effectively excavate the characteristics. At the same time, fewer data and uneven data distributions affect the experimental results. The data size of the crash crashes and engine noise events is larger than that for other events, so the accuracies are higher.

D. Comparative experiment results

To better analyze the performance of the model, four different classifiers were introduced and compared: the Gaussian mixture model (GMM)[17], support vector machine model (SVM)[18], k-nearest neighbor model (KNN)[19], and deep neural network model (DNN)[20]. The Gaussian kernel with a penalty coefficient of 1.5 was adopted as the SVM model kernel function and the OVR mode was adopted for classification, the number of neighbor for the KNN model was k=15, and the DNN model used four fully connected layers where the number of neurons in each layer was [988, 1024, 512, 6].

The 988-dimensional input feature was extracted with the OpenSmile tool[21]. The features are primarily composed of 26 low-level descriptors (LLDs) and their corresponding differences, including intensity, loudness, MFCC 1-12, LSP 0-7, the zero-crossing rate, voiced probability, fundamental

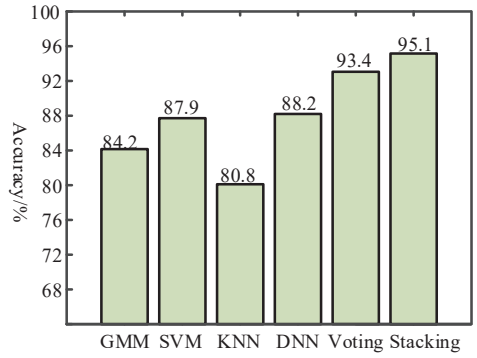

Fig. 8. Comparisons of experimental results

frequency, and fundamental frequency envelope. Then,19 statistical functions are calculated for 26 LLDS, so features extracted from each data have a total of 988 dimensions[22]. To display the results of the comparative experiment more intuitively, a histogram of the experimental results is drawn as follows. As shown in Fig. 8, it is obvious that the fusion algorithm comprising the two models proposed in this paper demonstrates a great improvement over other algorithms, and the recognition rate is 5.2% and 6.9% higher than that of the top-performing DNN.

IV. CONCLUSION

This paper proposes a model fusion algorithm for traffic acoustic event detection. Combining audio channel separation and audio segmentation methods, six different input features are extracted, and six different models are trained. The voting method and stacking method based on SVMs are adopted to fuse the output results from the six CNN models. To better extract the timbre features of the audio data, the harmonic source and percussive source are separated, and an improved two-channel CNN model is proposed to process the two-channel input feature. Before the fully connected layers, the two-channel CNN model uses separate convolutional layers to process different channels input features. Experimental results show that the two-channel CNN model plays a key role in improving detection accuracy. Moreover, the model fusion algorithm with a two-channel convolutional neural network as the weak classifier has good performance in acoustic event detection and will have good application prospects in road traffic monitoring.

REFERENCES

[1] Maciejewski Henryk, Mazurkiewicz Jacek, Skowron Krzysztof, Walkowiak Tomasz, "Neural Networks for Vehicle Recognition," in*Proc. of the 6th International Conference on Microelectronics for Neural Networks, Evolutionary and Fuzzy Systems*, New York, 1998, pp.292C296.

[2] C.X. Shi, Z.J. Zhou, H.M. Zhao, and M.R. Zhou, "Envelope extraction of underwater acoustic echo using wavelet and hilbert transforms," *IAENG International Journal of Computer Science,* vol. 47, no. 2, pp. 207-213, 2020.

[3] D.Y. Zhang, J. Jin, Zh.Zh. Guo, "Exploration into Road Traffic Accident Prevention Research System,"*China Safety Science Journal,* Vol.17, No.7, pp.132-138, 2007.

[4] X.L. Luo, G.H. Niu, "Vehicle recognition by acoustic signals based on EMD and SVM,"*Applied Acoustics,* Vol.29, No.3, pp.178-183, 2010.

[5] Q. Zheng, X. Tian, M. Yang, and H. Su, "The email author identification system based on Support Vector Machine (SVM) and Analytic Hierarchy Process (AHP)," *IAENG International Journal of Computer Science,* vol. 46, no. 2, pp. 178-191, 2019.

[6] R. F. Rachmadi, I. K. Eddy Purnama, M. H. Purnomo, and M. Hariadi, "A systematic evaluation of shallow convolutional neural network on CIFAR dataset," " *IAENG International Journal of Computer Science,* vol. 46, no. 2, pp. 365-376, 2019.

[7] Salamon Justin, Bello Juan,"Deep Convolutional Neural Networks and Data Augmentation for Environmental Sound Classification,"*IEEE Signal Processing Letters*, Vol. 99, pp.1-4, 2016.

[8] Piczak Karol J, "Environmental sound classification with convolutional neural networks,"*IEEE International Workshop on Machine Learning for Signal Processing,* pp.1-4,2015.

[9] H Bae S, I Choi, S Kim N, "Acoustic scene classification using parallel combination of LSTM and CNN,"*Proceedings of the Detection and Classification of Acoustic Scenes and Events,* pp.11-15,2016.

[10] Wang F, Cachecho P, Zhang W, et al. "Bayesian Model Fusion: Large-Scale Performance Modeling of Analog and Mixed-Signal Circuits by Reusing Early-Stage Data,"*Design Automation Conference,* 2013.

[11] C. Huang, B. Song, L. Zhao. "Emotional speech feature normalization and recognition based on speaker-sensitive feature clustering,"*International Journal of Speech Technology,* Vol.19, No.4, pp. 805-816, 2016.

[12] Eronen A J, Peltonen V T, Tuomi J T, et al. "Audio-based context recognition," *IEEE Transactions on Audio, Speech, and Language Processing,* Vol.14, No.1, pp.321-329, 2006.

[13] Eghbal-Zadeh H, Lehner B, Dorfer M, et al. "CP-JKU submissions for DCASE-2016: a hybrid approach using binaural i-vectors and deep convolutional neural networks,"*IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE),* 2016.

[14] Sakashita Y, Aono M. "Acoustic scene classification by ensemble of spectrograms based on adaptive temporal divisions,"*IEEE AASP Challenge on DCASE 2018 technical reports,*2018.

[15] Simonyan K, Zisserman A. "Very Deep Convolutional Networks for Large-Scale Image Recognition,"*Computer Science,* 2014.

[16] Li Y, Yuan Y. "Convergence Analysis of Two-layer Neural Networks with ReLU Activation,"*Proceedings of the 31st International Conference on Neural Information Processing Systems,* pp.597-607, 2017

[17] Hansen L P. "Large Sample Properties of Generalized Method of Moments Estimators,"*Econometrica,* Vol.50, No.4, pp.1029-1054, 1982.

[18] Cortes C, Vapnik V. "Support-vector networks,"*Machine learning,* Vol.20, No.3,pp.273-297, 1995.

[19] Hastie T, Tibshirani R . "Discriminant adaptive nearest neighbor classification.,"*IEEE Transactions on Pattern Analysis and Machine Intelligence,* Vol.18, No.6, pp.607-616,1996.

[20] Hinton G E, Osindero S, Teh Y W. "A Fast Learning Algorithm for Deep Belief Nets,"*Neural Computation,*Vol.18, No.7,pp.1527-1554, 2014.

[21] Zeng Z, Pantic M, Roisman G I, et al. "A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions,"*IEEE Trans Pattern Anal Mach Intell,* Vol.31, No.1,pp.39-58,2009.

[22] Schuller B, Steidl S, Batliner A, et al. "The INTERSPEECH 2010 paralinguistic challenge,"*Eleventh Annual Conference of the International Speech Communication Association,* 2010.

Xiaodan Zhang received Ph.D. degree from Southeast University, China, in 2013. Her major research areas are traffic safety and smart freeway, and she is specialized on modern intelligent data analytics based on Big Data. She works as the associate research fellow with Ministry of Transport Research Institute of Highway (RIOH) since August 2013.

Ming Li was born in Xuzhou, Jiangsu Province on August 28, 1994. She holds a master's degree in Southeast University. The main research direction is speech signal processing.

Chengwei Huang received his undergraduate degree in 2006, and Ph.D. for speech emotion recognition from Southeast University (China) in 2013. He conducted research on big data technologies as CTO of Sugon (Nanjing) Institute of Chinese Academy of Sciences. He is currently directing the A.I. research department in Intever Energy Technology Co. Ltd. (http://www.intelever.com/) His research interests include affective computing, signal processing and data mining.