Research on Traffic Acoustic Event Detection Algorithm Based on Sparse Autoencoder

Xiaodan Zhang[1,a], and Yongsheng Chen[1] and Guichen Tang[2]

[1]*Research Institute of Highway, Ministry of Transport, Beijing 100088, China*
[2]*School of Communication Engineering, Nanjing Institute of Technology, Jiangsu Nanjing, 211167, China*

Abstract. Road traffic monitoring is very important for intelligent transportation. The detection of traffic state based on acoustic information is a new research direction. A vehicles acoustic event classification algorithm based on sparse autoencoder is proposed to analysis the traffic state. Firstly, the multidimensional Mel-cepstrum features and energy features are extracted to form a feature vector of 125 features; Secondly, based on the computed features, the five-layers autoencoder is trained. Finally, vehicle audio samples are collected and the trained autoencoder is tested. The experimental results show that detection rate of the traffic acoustic event reaches 94.9%, which is 12.3% higher than that of the traditional Convolutional Neural Networks (CNN) algorithm.

1 Introduction

In order to provide auxiliary research for intelligent transportation and traffic safety development, an autoencoder based the acoustic feature for traffic event detection is proposed. In order to integrate the dynamic features of the sound, the algorithm extracts the multidimensional Mel-cepstrum features and energy features, and forms a 110 dimension feature vector. Then the 5-layers autoencoder based on the computed features is trained to improve the robustness. The experimental results show that the detection rate of traffic acoustic events reaches 94.9%, and the recognition rate of collision sounds reaches 97. 9%. The proposed audio surveillance system may be used to monitor traffic accidents and save valuable time in rescue mission. In addition, the system can be embedded in the automatic driving system, which is conducive to the timely response of the self-driving car to the traffic state, greatly improving safety.

The detection of traffic state based on acoustic information has been an important research direction for intelligent transportation. Compared to existing monitoring techniques, acoustic signal processing and classification techniques have the advantage of being low cost and unaffected by lighting conditions. Especially in the case of insufficient light or intermediate obstructions, acoustic signals have higher information coverage. Therefore it is an important supplement to existing monitoring methods. However, compared with the laboratory environment, the real traffic environment is complex. For example, the tunnel is a special traffic environment, that is very different from the open road environment. How to effectively process the traffic acoustic data remains a challenge. In 1998, Henryk Maciejewski et al.[1] studied and designed a classification system based on wavelet and neural network. The specific recognition model based the sound signal was constructed for four different vehicles, and the recognition accuracy was 73.68%. Audi Ovox et al. applied sound recognition technology to the field of intelligent transportation[2] and used voice recognition technology in the car phones. Xianglong Luo et al.[3] used empirical mode decomposition (EMD) and support vector machine (SVM) to identify the vehicle state. In recent years, some scholars have tried to apply convolutional neural networks (CNN) to recognize sound event[4]. Compared with traditional classifiers, convolutional neural networks have greatly improved recognition rate and recognition speed. The ConvNet model[5] has improved the accuracy of nearly 20% on Esc-50 database. The LSTM+CNN model proposed by Bae et al. achieved an 84.1% accuracy rate in the DCASE2016 competition[6]. However, there are still some problems when the traditional CNN model is applied to sound event recognition. The CNN adopts a serial stack structure. The convolutional layer in the network transforms the low-level feature map into a high-level feature map layer by layer. Such a network structure will result in the low-level feature information loss in the final extracted features.

For the above mentioned problems, a sparse autoencoder based on acoustic features is proposed to classify the vehicle state. The original acoustic features are multidimensional Mel-cepstrum features and energy features. The autoencoder generates encoded features to classify the vehicle state from the original acoustic features, which improve the robustness of the algorithm. To verify the performance of the proposed algorithm, we collected a total of 829 samples for experiment, including three types of data: engine running (normal driving), brake and crash. Compared with four algorithms, the recognition rate based on the proposed method can reach 94.9%, which is 11.45% higher than that of the traditional

[a] Corresponding author: zhangdaqing_925@163.com

classification algorithm or 12.3% higher than the traditional CNN algorithm.

2 Acoustic Features

For traffic acoustic event detection, three types of vehicles states are important, which are engine running (normal driving) state, crake state and crash state. The waveform and spectrogram of three kinds of signal are shown in Figure 1. From the graph, three states have some obvious difference. For example, the waveform of the running signal is flat and its frequency range is below 800Hz. In addition, the waveform and the spectrogram of the crash signal have the obvious change. After the analysis, the valid features should be selected to reflect these differences.

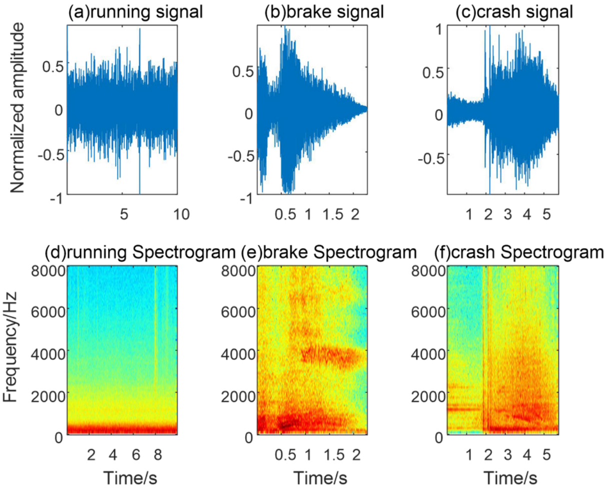


Figure 1. The representation of three types of signals

Among many acoustic features, the Mel-Cepstral feature is widely used in sound event classification and speaker recognition. The Mel-Cepstral is a spectral feature which is calculated base on the non-linear relationship between the human ear's auditory characteristics and signal frequency. However, the standard Mel-Cepstral only reflects the static characteristics of speech parameters. The dynamic characteristics of speech can be described by the differential spectrum of these features. The extracted features are differentiated to further strip the features, and obtain the information such as type and speed changes of the acoustic event. Combining the Mel-Cepstral feature with its difference can improve the recognition performance of the system.

In addition, different types of sounds have different energy values and change trends, so short-time energy and their statistical features are also computed. Table 1 describes the adopted 110 features.

Table 1. List of Acoustic Features.

No.	features
1-95	Mel-Frequency Cepstral Coefficients (MFCC-0 to MFCC-12), averages of their 1st and 2nd order differences, their maximums, minimums, ranges, and standard deviations.
96-110	Short-term energy, averages of its 1st and 2nd order difference, its maximum, minimum, range, and standard deviation.

3 Autoencoder for traffic acoustic event detection

The autoencoder is the neural network composed of several hidden layers and is an unsupervised learning algorithm. The general data representation from the input will be obtained by setting the target value as the input[7]. As shown in Fig.1, the autoencoder includes two parts: encoder and decoder. A number of hidden layers on the left of the graph form an encoder. The right hidden layers form a decoder, which has the same output as the encoder. In the middle of Figure 2, the output of the middle layer is the coded feature. In the process of the input reconfiguration, the data distribution is learned by the encoding and decoding process. At last, the data is compressed and coded, so the more compact features[8] are obtained.

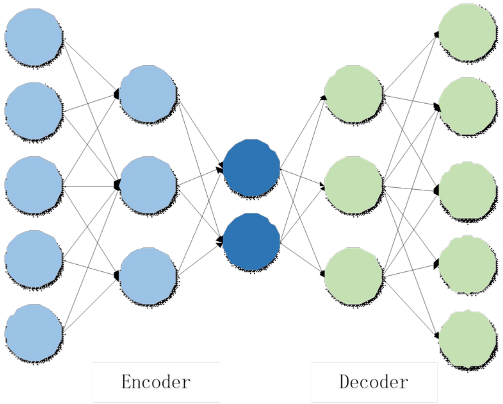


Figure 2. The architecture of autoencoder

For an autoencoder, x_i is the input, y_i is the output and $\boldsymbol{\theta}$ represents the parameters. The input x_i represents the 110-dimensional audio feature we extracted, and the output y_i is the optimized feature after the sparse autoencoder, which is also 110-dimensional. During the training, the goal of parameter optimization is:

$$\min_{\boldsymbol{\theta}} \sum_{i=1}^{N} \left\| x_i - y_i \right\|^2 \qquad (1)$$

By limiting the expected activation of the hidden unit to sparsity, a regularization term[9] is added to punish the deviations between the expected activation degree of the hidden unit and the target sparsity ρ. So equation (1) is:

$$\min_{\boldsymbol{\theta}} \sum_{i=1}^{N} \left\| x_i - y_i \right\|^2 + \beta \sum_{j=1}^{m} sp(\rho \,\|\, \hat{\rho}_j) \qquad (2)$$

where $sp(\rho\|\hat{\rho}_j)$ is the sparsity penalty term, which is computed as:

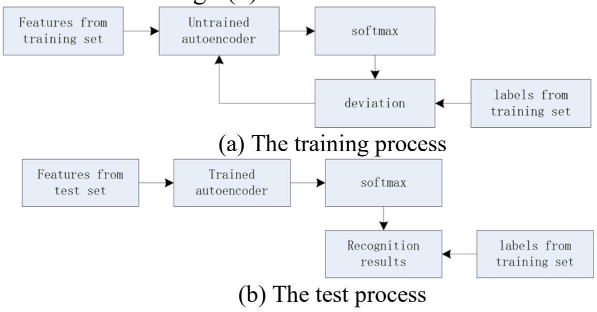
$$sp(\rho\|\hat{\rho}_j)=\rho\log\frac{\rho}{\hat{\rho}_j}+(1-\rho)\log\frac{1-\rho}{1-\hat{\rho}_j} \qquad (3)$$

Here, $\hat{\rho}_j$ is the average activation degree for all neurons in the hidden layer, ρ is sparsity, β is the penalty coefficient, and m is the number of all neurons. By adding the sparsity limitation[10], most of the neurons in the autoencoder are suppressed, only a small number of neurons are active, which reduces the redundancy[11] of the network and increases the robustness of the model.

Table2. Parameters Setting.

Component	No. Layer	#Neurons	Activation
Encoder	1 Fully Connected Layer	110	ReLU
	2 Fully Connected Layer	96	ReLU
	3 Fully Connected Layer	64	ReLU
Decoder	4 Fully Connected Layer	96	ReLU
	5 Fully Connected Layer	110	-

The 5-layers autoencoder is constructed for learning the latent feature representation of traffic acoustic event, and its parameter settings are shown in Table 2. The workflow of the proposed autoencoder is shown in Fig.3. The training process is shown in Fig.3(a). At first, the autoencoder is trained to study the latent representation of the original acoustic features from the training set. The learning strategy is to reconstruct input signals to obtain the latent representation, and use gradient descent method to minimize the error between the reconstructed signal and the input signal. When the deviation reaches the set requirements, a trained autoencoder is constructed. Then, a softmax classifier is added to compute the deviation between its output and the real labels. The gradient of the deviation is calculated and the back propagation algorithm is used to tune the parameters of each layer. After the training is finished, its performance of trained autoencoder can be estimated by the test data. The test flow is shown in Fig.3(b).



(a) The training process

(b) The test process

Figure 3. The workflow of the proposed model

4 Experimental analysis

4.1. Experimental data and parameter setting

In this experiment, a total of 829 samples were collected for the three types of sound, such as engine running(normal driving), braking and crashing. Among them, there are 442 engine running (normal driving) sounds, 176 brake sounds and 211 crash sounds.

In order to improve the effectiveness of the algorithm, the voice activity detection is firstly done on the sample to remove the silent segment. The sample is then resampled at a frequency of 16000 Hz. Next, the sample is framed and the FFT is calculated, the number of FFT points is 512, and the frame overlap rate is 50%.

The current general deep learning framework Caffe is used to build and train the network. Because the selection of hyperparameters in neural networks has a great impact on the training and the convergence state of the network, the final network hyperparameters are determined through multiple experiments and comparisons.

The comparison classifier and its parameters are: 1) Random forest[12], the maximum depth is 6, and the number of base estimators is 100; 2) CovNet[13], a two-layer convolutional layer with convolution kernels of (110, 6) and (1, 3), with a full layer of 1,000 neurons; 3) k-nearest neighbors (KNN); 4) the support vector machine.

4.2 Comparison of state-of-the-art recognition algorithms

The experimental evaluate the algorithm performance in a five-fold cross-validation manner. For a more valid assessment, Random Forest[14], KNN[15], CNN[16] and ConvNet performed the same experiments and comparisons on the same dataset.

Table 3 shows the experimental results for the five classifiers on the data set. Compared with random forest and KNN, the autoencoder improved the accuracy by 12.2% and 17.4%, respectively. It can be seen that the performance of autoencoder is better than the traditional classifier. The reason is that the data samples are very comprehensive, and there are many recorded data under various environments, while the generalization ability of the traditional classifier is bad. In addition, through data analysis, the traditional classifier has the lowest recognition rate for the brake event category, the highest recognition rate for the driving event category, followed by the collision event category. So, it may also be caused by the fact that the brake data is too small, and the data needs to be supplemented later for further verification.

Compared with traditional CNN and ConvNet, the recognition accuracy is improved by 12.3% and 3.9%, respectively. The possible reason is that the autoencoder uses encoded features to classify the state. These encoded features are more robust than the original features.

Table 3. Recognition rate of five models on the data set.

Classifier	Overall recognition rate	Collision recognition rate
Random Forests	82. 7%	84. 8%
DNN	77. 5%	85. 6%
CNN	82. 6%	90. 1%
ConvNet	91. 0%	92. 3%
Proposed method	94. 9%	97. 9%

In addition, Table 3 additionally compares the recognition efficiency of various algorithms for collision sounds. The experimental results show that the proposed algorithm's recognition rate of collision sound is 3% more than the overall recognition rate, reaching 97. 9%. This shows that the algorithm can be effectively applied to the traffic incident warning system to realize the timely alarm function of traffic accidents based on traffic state detection.

5 Conclusions

In order to provide auxiliary research for intelligent transportation and traffic safety development, an autoencoder based the acoustic feature for traffic event detection is proposed. In order to integrate the dynamic features of the sound, the algorithm extracts the multidimensional Mel-cepstrum features and energy features, and forms a 110 dimension feature vector. Then the 5-layers autoencoder based on the computed features is trained to improve the robustness. The experimental results show that the detection rate of traffic acoustic events reaches 94.9%, and the recognition rate of collision sounds reaches 97. 9%. The proposed audio surveillance system may be used to monitor traffic accidents and save valuable time in rescue mission. In addition, the system can be embedded in the automatic driving system, which is conducive to the timely response of the self-driving car to the traffic state, greatly improving safety.

Acknowledgments

The work was supported by Science and Technology Innovation Project of Research Institute of Highway, Ministry of Transport (2018-E0021).

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, 1998, pp.292–296.
2. Zhang Dian Ye, Jian Prof Jin, Zhi-Zheng Assoc Prof Guo. Exploration into Road Traffic Accident Prevention Research System. China Safety Science Journal, Vol.17, No.7, 2007, pp.132-138.
3. Luo Xiang Long, Niu. Vehicle recognition by acoustic signals based on EMD and SVM. Applied Acoustics, Vol.29, No.3, 2010, pp.178-183.
4. Salamon Justin, Bello Juan. Deep Convolutional Neural Networks and Data Augmentation for Environmental Sound Classification. IEEE Signal Processing Letters, Vol. 99, 2016, pp.1-4.
5. Piczak Karol J. Environmental sound classification with convolutional neural networks. IEEE International Workshop on Machine Learning for Signal Processing, 2015, pp.1-4.
6. 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, 2016, pp.11-15.
7. Xu J, Xiang L, Liu Q, et al. Stacked Sparse Autoencoder (SSAE) for Nuclei Detection on Breast Cancer Histopathology Images, IEEE Transactions on Medical Imaging, Vol.35, No.1, 2016, pp.119-130.
8. Chandar A P S, Lauly S, Larochelle H, et al. An autoencoder approach to learning bilingual word representations, International Conference on Neural Information Processing Systems. MIT Press, 2014, pp.1853-1861.
9. Goodfellow I J, Le Q V, Saxe A M, et al. Measuring invariances in deep networks, International Conference on Neural Information Processing Systems, 2009, pp.646-654.
10. Mairal J, Bach F, Ponce J, et al. Online Learning for Matrix Factorization and Sparse Coding, Journal of Machine Learning Research, Vol.11, No.1, 2009, pp.19-60.
11. Hinton G E, Salakhutdinov R R. Reducing the Dimensionality of Data with Neural Networks, Science, 313(5786), 2006, pp.504-507.
12. Phan Huy, Maaß Marco, Mazur Radoslaw, Mertins Alfred. Random Regression Forests for Acoustic Event Detection and Classification. IEEE/ACM Transactions on acoustic Speech & Language Processing, Vol.23, No.1, 2015, pp.20-31.
13. Garcia-Pedrajas N, Hervas-Martinez C, Munoz-Perez J. COVNET: a cooperative coevolutionary model for evolving artificial neural networks. IEEE Transactions on Neural Networks, Vol.14, No.3, 2003, pp.575-596.
14. Pal, M. Random forest classifier for remote sensing classification. International Journal of Remote Sensing, 2005, 26(1):217-222.
15. Zhang M L , Zhou Z H . ML-KNN: A lazy learning approach to multi-label learning. Pattern Recognition, 2007, 40(7):2038-2048.
16. Wei Y , Zhao Y , Lu C , et al. Cross-Modal Retrieval With CNN Visual Features: A New Baseline. IEEE Transactions on Cybernetics, 2017, 47(2):449-460.