2022 21st International Symposium on Distributed Computing and Applications for Business Engineering and Science (DCABES)

Road Rage Recognition System Based on Speech Features

Yang Li[1],Wenjing Wang[2],Xinmin Xu[1]
[1]College of Information Science & Electrical Engineering, Zhejiang University, Hangzhou, China 310027
[2]Jinhua Institute of Zhejiang University,Jinhua, China 321032
Email: xuxm@zju.edu.cn

***Abstract*—A major cause of traffic accidents is road rage. How to identify road rage is an important problem that needs to be solved urgently. Road rage recognition is different from traditional emotion recognition. The sound signal to be recognized contains complex traffic environment noise, and the recognition target is a single anger emotion. This paper extracts high robustness, high generalization, and anger features from speech signals. A convolutional neural network (CNN) and multi-headed self-attention criterion bi-directional long-short-term memory network (Multi-headed Self-Attention Bi-LSTM) fusion decision model is proposed to realize anger emotion recognition.**

Keywords-Anger emotion recognition, road rage, speech signal processing, deep learning, feature fusion

I. Introduction

Road rage is defined as the driver's anger while driving the vehicle. The psychological community defines road rage according to the specific physiological information of the driver, which is expressed as anger and negative emotions related to rapid heartbeat, shortness of breath, and capillary spout, which form aggressive psychological motivation, and finally cause excessive driving to occur.

The main research contents of speech-based anger emotion recognition are divided into the following three categories, dataset establishment, feature extraction, and emotion recognition algorithm. Different from traditional emotion recognition, which requires the recognition of multiple emotion classifications, the research object of this paper is a single anger emotion, so it is necessary to solve the problem of how to effectively extract the speech features representing anger emotion. At the same time, the datasets on which existing emotion recognition is based often do not contain noise, which is different from the actual emotion recognition environment. System noise reduction is often improved in three directions, including extracting robust features, building speech enhancement algorithms, and building acoustic adaptive algorithms to accomplish various tasks. Therefore, this paper improves the feature and algorithm model to improve the robustness of the recognition system. Finally, a complete anger emotion recognition system for speech signal acquisition, preprocessing, noise reduction, and emotion recognition in actual scenes are constructed.

The purpose of this paper is to study the detection of anger in the noise environment based on speech features. At the same time, the anger detection algorithm proposed in this paper is applied to the road rage emotion diagnosis system, to use the road rage information to assist driving behavior.

II. Basics of Speech Emotion Recognition

A. Emotional Model

Emotions are any relatively short-lived conscious experience characterized by intense mental activity, divided into model emotional models and dimensional emotional models. The vertical and horizontal axes of the arousal-valence dimension space theory are the arousal dimension and the valence dimension, respectively. The arousal dimension, that is, the vertical axis, expresses the intensity of emotions, and the valence dimension, that is, the horizontal axis, expresses the degree of positive and negative emotions. All emotions are the same distance from the origin.

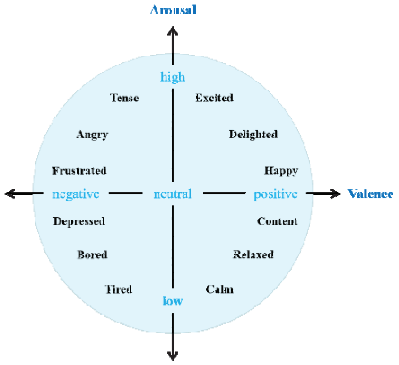

Figure 1. arousal-valence dimension space theory.

B. Speech signal preprocessing

In order to realize anger emotion recognition, the input speech signal needs to be preprocessed first. The preprocessing can reduce the negative effects caused by the voice signal itself, the aliasing phenomenon, and the loss of high-order harmonics formed by the voice signal itself, the device for receiving the voice signal, and make it uniform and smooth.

C. Feature extraction of speech signal

The three elements of sound are: loudness, pitch, and timbre, and the acoustic features are extracted according to the three elements. The features extracted from speech signals can generally be divided into the following four categories, prosody, timbre, spectral correlation, and deep learning features. During the whole process of emotion recognition, the acoustic features can be calculated for each frame using statistical functions. The statistical functions used include mean, variance, etc.

2473-3636/22/$31.00 ©2022 IEEE
DOI 10.1109/DCABES57229.2022.00070

III. ANGER EMOTION RECOGNITION BASED ON FUSION FEATURES OF F-MGCC

This chapter mainly improves and integrates MFCC and GFCC. Since there are more Mel filters in the low-frequency region and fewer in the high-frequency region, it can better characterize the low-frequency speech signal. However, anger signals are often high-frequency signals, so a Mel filter bank needs to be designed to better characterize high-frequency signals. At the same time, only using splicing and fusion for MFCC and GFCC will lead to a lot of redundancy. Therefore, it is necessary to remove redundant features through algorithms to improve subsequent calculation efficiency and accuracy.

A. Fisher's ratio criterion

Fisher's ratio criterion borrows the thinking of equal variance analysis to project the data in the spatial dimension. After projection, it is hoped that the projection center point of each class of data is as close as possible, and the distance between the class centers of each class of data is Also as big as possible.

In anger emotion recognition, the features of the speech signal are projected, so that the feature points of both anger and non-anger emotions are more concentrated in their respective projection areas in the multi-dimensional space, and the feature points of anger and non-anger are in the multi-dimensional space. The distance between the projection areas is scattered, as in

$$J(k)=\frac{S_B^{(k)}}{S_W^{(k)}}=\frac{\sum_{c=1}^{C}(m_c^{(k)}-m^k)^2}{\sum_{c=1}^{C}[\frac{1}{n}\sum_{a^{(k)}\in\omega_c}(q^{(k)}-m^{(k)})^2]},1\leq c\leq C \quad .(1)$$

$S_B^{(k)}$ is the between-class variance of the kth feature; $S_W^{(k)}$ is the intra-class variance of the k th feature. The sample set of each dataset is ω_c ; $m^{(k)}$ represents the average value of the kth feature of a sample $q^{(k)}$ in the angry emotion speech dataset; $m_c^{(k)}$ represents the average value of the kth feature of all samples in the cth class of the anger emotion dataset; n represents how many samples the dataset contains strip samples.

B. Splicing MFCC

MFCC and IMFCC perform signal characterization at low frequency and high frequency, respectively. The 1-6 order coefficients of the MFCC and the 7-12 order coefficients of the IMFCC can be spliced to obtain a 12-order spliced MFCC.

C. Hybrid MFCC

In the low (1-4000Hz) and high frequency bands (4000-8000Hz), the 12th-order Mel filter bank and the 12th-order inverse Mel filter bank are respectively used to generate mixed MFCC features, which can be compared in both high and low frequency bands. A good representation of the characteristics, results in a 12-order hybrid MFCC.

D. F-MFCC

The Fisher ratios of all cepstral coefficients of each order are sorted from high to low, and the highest ranked 12-dimensional feature is selected, and the 12-dimensional features are spliced and combined into F-MFCC features.

E. F-MGCC

For the F-MGCC extraction process, first extract the MFCC, IMFCC, and GFCC characteristic parameters for each speech sample, then obtain the Fisher ratio of the three cepstral coefficients and carry out the Fisher ratios of all cepstral coefficients from high to low. This paper sorts and selects the 18-dimensional feature with the highest ranking, then splices the 18-dimensional feature to form the F-MGCC feature.

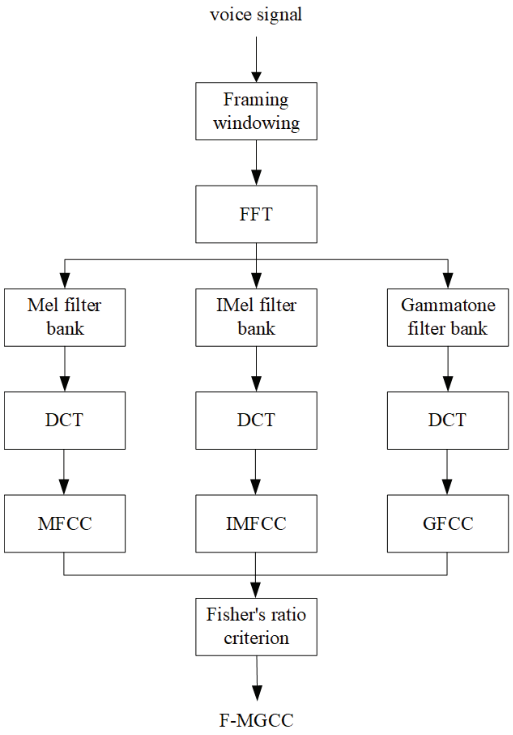

Figure 2. F-MGCC extraction process.

Represent the contribution of each order feature in different MFCCs, IMFCCs, and MGCCs, select the 1st, 2nd, 3rd, 4th, 5th, and 8th orders in MFCC, the 1st, 2nd, 6th, and 7th orders in IMFCC, and the first, 2, 6, and 7 orders in GFCC. The 1st, 2nd, 3rd, 6th, 7th, 8th, 10th, and 12th orders constitute a new 18th-order F-MGCC feature.

IV. ANGER EMOTION RECOGNITION BASED ON DEEP LEARNING

The CNN network will be used to extract the spatial association of low-level features. The attention mechanism and Bi-LSTM network will be used to capture the time complexity in the feature sequence, and Softmax will be used to classify, and finally realize the recognition of anger.

A. Network Architecture Design

The model in this paper mainly uses CNN to obtain high-level feature vectors of spatial dimensions in speech feature parameters. The Bi-LSTM is used to extract the high-level feature vector of the time dimension in the speech feature parameters combined with the multi-head self-attention mechanism. The output of Multi-headed Self-Attention Bi-LSTM is spliced with the output of CNN, then the fully connected layer is used, and finally the Softmax function is called to complete the anger emotion recognition result.

In order to enhance the relevance of the speech context, it is necessary to prevent the complementation of the speech signal in the back of the time series to the speech signal in the front, such as: when an angry emotion is generated, the subsequent emotion is enhanced. The LSTM is trained from the time series from front to back and from back to front, respectively. At t time , the state values A_t and $A_t^{'}$ output on both sides are spliced.

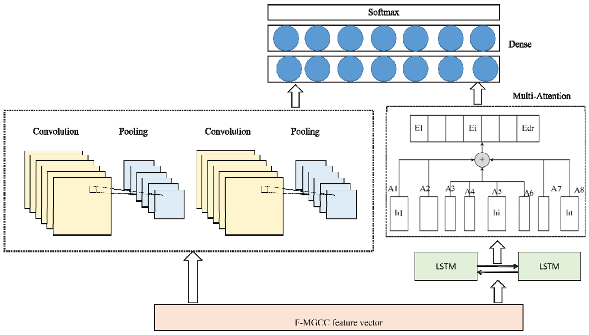

Figure 3. CNN+Multi-Attention Bi-LSTM fusion decision model structure diagram.

B. Network Model Optimization Strategy

1) Loss function selection

If the root mean square error is used, it is often found when the model is just trained that when the output probability of the model is close to 0 or 1, the partial derivative value will become very small, resulting in slow parameter update speed and model is unable to converge. So, this paper chooses cross entropy as the loss function.

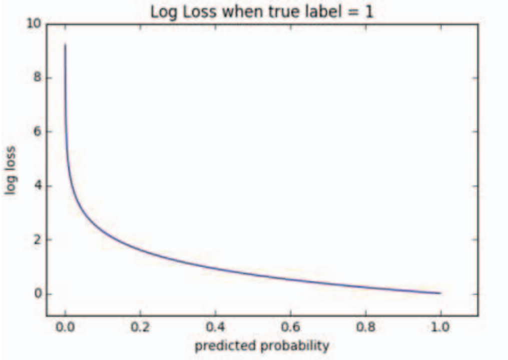

Figure 4. Cross-entropy loss function and prediction probability relationship

Cross-descendence loss function:

$$L = \frac{1}{N}\sum_i -[y_i\, log(p_i) + (1\text{-}y_i)\log(1\text{-}p_i)] \quad . \qquad (2)$$

y_i indicates the label of the sample i , the positive class is 1, and the negative class is 0; p_i represents the probability that the sample i is predicted to be a positive class.

2) Optimizer selection

The Adam algorithm has both the advantages of the Momentum and RMSprop algorithms, and improves the gradient calculation method and learning rate, which is equivalent to RMSprop with a momentum term, so this paper uses the Adam optimizer.

C. Experimental results

1) Feature and Model Generalization Ability Analysis

Through the calculation of the noise in the car and the voice power of the human voice, the experiment is carried out on the composition of F-MGCC in the pure Clean speech, 0dB, -10dB and -20dB signal-to-noise ratio speech in this chapter. At the same time, in order to take into account the high and low signal-to-noise ratios, we explored the accuracy of speech recognition of anger under other signal-to-noise ratio conditions when F-MGCC was used as the input for network training with a single signal-to-noise ratio.

TABLE 1. GENERALIZED F-MGCC ANGER EMOTION RECOGNITION ACCURACY

Correct rate (%)	The correct rate of anger recognition results(%)				
	Clean	0dB	-10dB	-20dB	Ave
Clean	**92.71**	87.15	86.03	64.24	82.53
0dB	86.11	**92.01**	90.97	87.5	89.15
-10dB	83.3	89.93	**92.71**	88.85	88.70
-20dB	86.11	87.15	88.89	**92.36**	88.63

2) Comparison Experiment of Anger Recognition Models

On the two datasets, the accuracy of anger recognition is 2.47%, 2.26%, 2.51%, and 2.40% higher than the prediction accuracy of the traditional single LSTM and CNN models, respectively. Compared with the serial models of CNN and LSTM, the accuracy of the model in this paper is improved by 1.98%, 2.50%, 1.41% and 2.34% respectively.

TABLE 2. MODEL TRAINING PARAMETERS

learning rate	total epoch	batch size
0.001	200	128

TABLE 3. THE CORRECT RATE OF CNN+MULTI-ATTENTION BI-LSTM IN IEMOCA AND CASIA ANGER RECOGNITION

	RAVDESS	CASIA
LSTM	91.36	91.86
CNN	91.32	91.72
LSTM+CNN	92.71	92.71
CNN+LSTM	88.54	91.78
CNN+Multi-Attention Bi-LSTM	**93.83**	**94.12**

V. CONCLUSION AND DISCUSSION

This paper mainly studies anger emotion recognition based on speech features, and applies the anger emotion recognition model to the "road rage" emotion diagnosis system to achieve statistical analysis of the frequency and degree of "road rage" during driving. assisted driving behavior. The main contents of this paper are as follows:

- The reading materials have an understanding of the research background and recognition methods of the "road rage" problem, and comparatively analyze the related technologies of speech emotion recognition at home and abroad. The second chapter introduces the emotion model, and speech signal related technology, including preprocessing and feature classification and comparison, to prepare for the subsequent feature extraction.
- Because MFCC cannot characterize high-frequency signals well, IMFCC is introduced due to the large frequency domain of anger. At the same time, in order to solve the anger recognition in a noisy environment, a robust GFCC is introduced. The characteristics, extraction process, and methods of the above three features are introduced in detail.
- For the fusion feature construction of MFCC and IMFCC, the splicing method, filter combination method, and Fisher ratio criterion method are adopted to obtain 12-dimensional splicing MFCC, hybrid MFCC, and F-MFCC respectively. For MFCC, IMFCC, and GFCC, Fisher's ratio criterion method is used to obtain the 18-dimensional features with the highest contribution to speech-based anger recognition, which constitutes an 18-dimensional F-MGCC. And in the speech signal experiments with different signal-to-noise ratios, the distribution of MFCC, IMFCC and GFCC is analyzed and counted, and a 20-dimensional generalized F-MGCC is obtained. Experiments show that the F-MGCC constructed by Fisher improves the accuracy of anger emotion recognition by 7.53% compared with splicing MFCC, hybrid MFCC, GFCC, MFCC, and IMFCC. The generalized F-MGCC training model for single-template SNR speech can obtain an average accuracy of 87.25% on the SNR speech signals of Clean, 0dB, -10dB, and -20dB.
- Construct CNN+Multi-headed Self-Attention Bi-LSTM fusion decision model to complete the recognition and classification task of anger. CNN extracts spatial domain depth feature vectors, and LSTM extracts time domain depth feature vectors. The model is improved by using Bi-LSTM and multi-head attention mechanism respectively. Mixing different signal-to-noise ratio speeches in equal proportions, extracting the improved generalized F-MGCC fusion feature parameters for model training, the accuracy of anger recognition rate in a noisy environment is 93.83 on RAVDESS and CASIA datasets respectively % and 94.72%.

The shortcomings of this experiment are: Firstly, the real "Road Rage" speech dataset is established. The dataset used in this paper is generated based on the superposition of noise and public datasets according to a certain signal-to-noise ratio, which is different from the actual "Road Rage" voice dataset. According to the noise type, the speech signal-to-noise ratio is classified and constructed. Secondly, Supplement the features of the speech signal spectrum. This paper mainly focuses on the fusion and optimization of the relevant features of the speech spectrum. Therefore, in the follow-up research, when constructing features, prosody can be integrated. academic and sound quality characteristics.

References

[1] Stephens A N and Groeger J A . “Anger-congruent behaviour transfers across driving situations.” Cognition and Emotion, 2011, vol. 25, pp.1423-1438.

[2] Vydana H K , Kumar P P , Krishna K , et al. “Improved emotion recognition using GMM-UBMs.” International Conference on Signal Processing & Communication Engineering Systems. IEEE, 2015.

[3] Yang S , Yang G . “Emotion Recognition of EMG Based on Improved L-M BP Neural Network and SVM.” Journal of Software, 2011, vol. 6, pp.1529-1536.

[4] Fahad M S , Deepak A , Pradhan G , et al. “DNN-HMM-Based Speaker-Adaptive Emotion Recognition Using MFCC and Epoch-Based Features.” Circuits, Systems, and Signal Processing, 2021, vol. 40, pp.466-489.

[5] Pitaloka D A , Wulandari A , Basaruddin T , et al. “Enhancing CNN with Preprocessing Stage in Automatic Emotion Recognition.” Procedia Computer Science, 2017.

[6] Zhang T , Wu J . “Speech emotion recognition with i-vector feature and RNN model.” 2015 IEEE China Summit and International Conference on Signal and Information Processing (ChinaSIP). IEEE, 2015, pp.524-528.

[7] Lee J , Tashev I . “High-level Feature Representation using Recurrent Neural Network for Speech Emotion Recognition.” Interspeech. 2015.

[8] Badshah A M , Ahmad J , Rahim N , et al. “Speech Emotion Recognition from Spectrograms with Deep Convolutional Neural Network.” 2017 International Conference on Platform Technology and Service (PlatCon). IEEE, 2017.

[9] Zhao Z , Bao Z , Zhao Y , et al. “Exploring Deep Spectrum Representations via Attention-Based Recurrent and Convolutional Neural Networks for Speech Emotion Recognition.” IEEE Access, 2019.

[10] Geravanchizadeh M , Akhtari-Khosroshahi S , Zakeri S . “Audio Scene Classification Based on New Hybrid Feature and Bidirectional Long Short-Term Memory.” 11th International Conference on Information Technology, Computer & Telecommunication. 2021.

[11] Papa panagiotou, Vasileios, Diou C , Delopoulos A . “Self-Supervised Feature Learning of 1D Convolutional Neural Networks with Contrastive Loss Using In-Ear Microphone Audio for Eating Detection.” 2021.

[12] Durk T , Doty T J , Woldorff M G . “Selective attention and audiovisual integration: is attending to both modalities a prerequisite for early integration.” Cerebral Cortex, 2007, vol.17, pp.679-690.

[13] Gemmeke J F , Ellis D , Freedman D , et al. “Audio Set: An ontology and human-labeled dataset for audio events.” IEEE International Conference on Acoustics. IEEE, 2017.

[14] Li P , Yan S , Mcloughlin I , et al. “An Attention Pooling Based Representation Learning Method for Speech Emotion Recognition.” Interspeech 2018. 2018.

[15] Chen M , X He, Jing Y , et al. “3-D Convolutional Recurrent Neural Networks With Attention Model for Speech Emotion Recognition.” IEEE Signal Processing Letters, 2018, vol.25, pp.1440-1444.