PAD 3-D Speech Emotion Recognition Based on Feature Fusion

Xia Pengfei, Zhou Houpan, Zhou Weidong
School of Automation (School of Artificial Intelligence), Hangzhou Dianzi University, Hangzhou Zhejiang 310018, China

Abstract. In order to establish the mutual fusion relationship between the global and timing features of speech and achieve better speech emotion recognition, this paper proposes a PAD 3-D space emotion recognition method based on feature fusion. This paper uses the Opensmile to extract four speech global feature sets: IS09_emotion, IS10_paraling, IS11_speaker_state, IS12_speaker_trait, and establishes a deep learning model combined with CNN, LSTM, and attention mechanism to extract the timing features of each frame simultaneously. Finally, the global features and timing features are fused through the Stacking model. The experimental results show that compared with using global features or timing features alone for speech emotion recognition, the method of using the Stacking model to fuse global features and temporal features can effectively improve the accuracy of speech emotion recognition.

1. Introduction
Speech is a common emotional medium, which contains rich emotional information. With the rapid development of machine learning, deep learning and the continuous emergence of relevant research, speech emotion recognition is more used in education, entertainment, and communications[1-4]. Strengthening the intelligent recognition of speech emotion has become the focus of the next generation of artificial intelligence development. In view of this, the research on speech emotion recognition has important theoretical value and practical significance.

Early researchers mainly modeled speech emotion recognition based on traditional machine learning methods, such as Hidden Markov Model, Decision Tree, and SVM (Support Vector Machine)[5-7]. In recent years, deep learning has shown more powerful feature extraction capabilities under massive data, more and more researchers have begun to use deep learning methods to deal with speech emotion recognition problems. For example, extract spatial local features in the spectrogram based on CNN; and extract timing information based on LSTM and other recurrent neural networks[8-10]. In order to solve the loss of feature information caused by overlong timing information, researchers of speech emotion recognition have introduced an attention mechanism[11 -12].

The PAD 3-D spatial emotion model is a simple and widely used dimensional emotional description model. P (Pleasure-Displeasure) indicates the positive or negative characteristics of an individual's emotional state, A (Arousal-Nonaroused) indicates the degree of individual's neurophysiological activation, D (Dominance-Submissiveness) indicates the individual's subjective control over the environment and others[13-15]. Chen Y L[14] extracted the features of time ignition sequence and ignition position information from the spectrogram to supplement MFCC (Mel Frequency Cepstral Coefficient), and obtained the final predicted result through weighted fusion calculation. The prediction accuracy of this method has been improved to some extent, but the prediction accuracy of its P, A, and D is still low. Sun Y[15] extracted the principal components of the
main features through PCA (Principal Component Analysis), used the principal components as input of SVM. This method improves the prediction accuracy of P, A, and D, but the experiment is only based on three emotions (sadness, anger, and happiness), and it does not consider the difficult emotions such as mildness and disgust, and it cannot be fully reflected problem of emotional dimension.

2. **Algorithm Theory**

2.1. *Introduction to Stacking Model*

As a ensemble model, Stacking is generally divided into two layers. Figure 1 shows the Stacking model. The first layer is composed of multiple base learners, and the second layer is a meta-learner. The meta-learner train all the outputs of the base learners and obtain the final prediction results[16-17]. The training process of the Stacking model is as follows:

1. The original data \((x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\) is divided into training set and test set. K-fold cross-validation is used to divide the training set into K subsets of equal size: Train\(_1\), Train\(_2\) ... Train\(_k\).

2. Use Train\(_2\), Train\(_3\) ... Train\(_k\) as Set\(_1\), Train\(_1\) as the verification set. Based on the Set\(_1\) training base learner, the prediction result Val\(_1\) of the Train\(_1\) is obtained, and the prediction result Test\(_1\) of test set is obtained at the same time.

3. Use Train\(_1\), Train\(_3\) ... Train\(_k\) as Set\(_2\), Train\(_2\) as the verification set. Based on the Set\(_2\) training base learner, the prediction result Val\(_2\) of the Train\(_2\) is obtained, and the prediction result Test\(_2\) of test set is obtained at the same time.

4. Train\(_3\), Train\(_4\)...Train\(_k\) are used as the verification set respectively, and the prediction results are Val\(_3\), Val\(_4\)... Val\(_k\), and the prediction results of the test set are Test\(_3\), Test\(_4\)... Test\(_k\).

5. Matrix stitching of Val\(_1\), Val\(_2\)... Val\(_k\) to get 1 column of Val\(_{all}\), and use this data as the training data of the meta-learner. At the same time, the average value as Test\(_{avg}\) of Test\(_1\), Test\(_2\)... Test\(_k\) is obtained, and Test\(_{avg}\) is used as the test data on the meta-learner.

![Fig. 1 Schematic diagram of Stacking model](image)

2.2. *Introduction to LightGBM*

Light Gradient Boosting Machine (LightGBM) is a type of Gradient Boosting Decision Tree (GBDT), which was proposed by Microsoft in 2015[18]. In the GBDT, the most time-consuming step is to use the Pre-Sorted method to enumerate all possible feature points on the ordered feature values, and then find the optimal partition point. The LightGBM uses the histogram algorithm to replace the traditional Pre-Sorted to reduce memory consumption.

LightGBM abandoned the Level-wise method and adopted the Leaf-wise algorithm. Leaf-wise finds the maximum gain split method from the existing leaves in each iteration, and loops until the
given maximum depth is reached. This method effectively avoids unnecessary overhead and improves the calculation rate[19]. As shown in the figure below is the Leaf-wise algorithm.

![Fig.2 Schematic diagram of Leaf-wise](image)

### 2.3. Introduction to Adaboost Regression

Adaboost is an iterative algorithm whose idea is to train different classifiers against the same training set, and then combine these weak classifiers to form a stronger classifier. When the Adaboost handles the regression problem, the focus is on how to update the sample weights and classifier weights. The following is the AdaBoost R2 regression algorithm[20]:

1. For the Mth weak learner, calculate its maximum error on the training set:
   \[
   E_m = \max |y_i - G_m(x_i)|, \quad i = 1, 2, ... k
   \]  
   (1)

2. Calculate the relative error of each sample:
   \[
   e_{mi} = \frac{|y_i - G_m(x_i)|}{E_m}
   \]  
   (2)

3. Get the error rate of the Mth weak classifier:
   \[
   e_m = \sum_{i=1}^{k} w_{mi} e_{mi}
   \]  
   (3)

4. Get the weak learner weight coefficient  \( \alpha_m \):
   \[
   \alpha_m = \frac{e_m}{1 - e_m}
   \]  
   (4)

5. For the updated sample weight \( D \), the \( m+1 \)th weak learner sample weight set coefficients are:
   \[
   w_{m+1}, i = \frac{w_{mi}}{Z_m} \alpha_m^{1 - e_{mi}}
   \]  
   (5)

   \( Z_m \) is the normalization factor:
   \[
   Z_m = \sum_{i=1}^{k} w_{mi} \alpha_m^{1 - e_{mi}}
   \]  
   (6)

6. Using the combination strategy of weighted average method, the final strong regressor is:
   \[
   f(x) = \sum_{i=1}^{m} \left( \ln \frac{1}{\alpha_m} G_m(x_i) \right)
   \]  
   (7)

### 2.4. Introduction to CNN-LSTM-Attention

The deep learning neural network can be mapped to arbitrary functions theoretically and can solve complex problems. At present, deep learning algorithms such as CNN, LSTM and Attention are widely used in the field of speech emotion recognition [8-12]. \( [X_1, X_2, X_3 ... X_T] \) can be obtained by extracting the timing features for each audio data, where T represents the time step. The following is based on the CNN-LSTM-Attention deep learning algorithm formula definition:

\[
S_i = CNN(X_i), \quad i \in \{1,2,3 ... T\}
\]  
(8)

\[
H_i = LSTM(S), \quad i \in \{1,2,3 ... T\}
\]  
(9)

\[
f(H_i) = \text{tanh}(w^T H_i + b), \quad i \in \{1,2,3 ... T\}
\]  
(10)

\[
\nu_i = \frac{e^{f(H_i)}}{\sum_{t=1}^{T} e^{f(H_t)}}
\]  
(11)

\[
e_i = \nu_i H_i
\]  
(12)

Where \( S_i \) is the result of convolutional pooling of the timing features by the convolutional neural network, \( H_i \) represents the state of the hidden layer of LSTM, \( f(H_i) \) is the correlation function, \( w \) and \( b \) are the training parameters of the model, \( \nu_i \) represents the attention weight parameter calculated for the input vector \( H_i \), and \( e_i \) is the output obtained after the attention layer is weighted.
2.5. Introduction to Ridge Regression
Ridge regression is essentially an improved Least Squares Method, which is a biased estimation regression method dedicated to collinear data analysis. Ridge regression adds L2 regularization in the loss function, and the final loss function is as follows:

\[
J = \frac{1}{n} \sum_{i=1}^{n} (f(x_i) - y_i)^2 + \lambda ||w||_2^2
\]

(13)

3. Test Results and Discussions

3.1. Description of database
The speech databases selected in this experiment are EMODB and RAVDESS, which are widely used in speech emotion recognition[21-22]. EMODB database is obtained by 10 actors, voice emotions include neutral, anger, fear, joy, sadness, disgust, a total of 535 sentences. RAVDESS database is described by 24 actors in a neutral North American accent. Voice emotions include neutral, clam, happy, sad, angry, fear and disgust and surprised, a total of 1440 sentences.

The PAD values corresponding to the basic emotion types are shown in Table 1. According to the original PAD table developed by Mehrabian and the correspondence between the simplified Chinese PAD value revised by the Chinese Academy of Sciences and the basic emotion type, the PAD value of each emotion type in the database can be obtained[13-15].

| Emotion Type   | P    | A    | D    |
|----------------|------|------|------|
| Neutral        | 0.27 | -0.07| -0.14|
| Joy            | 0.66 | 0.74 | 0.32 |
| Unhappy        | -0.28| -0.36| -0.78|
| Anger          | -0.86| 0.66 | 0.91 |
| Startle        | -0.33| 0.65 | -0.72|
| Sadness        | -0.46| 0.58 | -0.13|
| Disgust        | -0.44| 0.22 | 0.36 |
| Fear           | -0.33| 0.65 | -0.72|

The EMODB and RAVDESS data are divided into training data and test data respectively, and the division ratio of each emotion type is 4:1. As shown in Table 2 below.

| Database   | Total data | Training data | Test data |
|------------|------------|---------------|-----------|
| EMODB      | 535        | 428           | 107       |
| RAVDESS    | 1440       | 1152          | 288       |

3.2. Description of Feature Extraction
In this paper, IS09_emotion, IS10_paraling, IS11_speaker_state, and IS12_speaker_trait are extracted through the Opensmile toolkit. These four speech feature sets are derived from the Challenges held by InterSpeech Journal from 2009 to 2012[23-27]. At the same time, this paper uses a 300ms window and a 150ms step to extract speech timing features. Since the speech fragments are not the same length, the number of timing features of each sentence is fixed to 401 during the experiment using padding. Extracting the timing features of each frame includes MFCC, Chroma_cens, Zero_crossing_rate, and melspectrogram. A total of 150-dimensional timing features are extracted, in which the sampling frequency is 16khz. Table 3 below shows the feature introduction.

| Global features | Number of features | Timing features | Number of features |
|-----------------|--------------------|----------------|--------------------|


### 3.3. Description of Experiment Procedure

The schematic diagram of the experimental process is shown in Figure 3. After extracting four global features and timing features from the original data, the feature set is divided into training data and test data respectively, and perform 4-fold cross-validation on the training data.

![Fig.3 Schematic diagram of the experimental process](image)

Adaboost and LightGBM are used as the basic learners of Stacking respectively, for training global features and obtain prediction results of global features. In the experiment, the Adaboost model selects AdaBoostRegressor, in which the weak learner is a CART decision tree, and the parameters of the CART decision tree are selected by default, and the learning rate of Adaboost is 0.05. At the same time, the LightGBM model selects LGBMRegressor, where the promotion type is GDBT, the number of leaf node is 31, the learning rate is 0.05, the feature selection ratio of tree-building is 0.9, and the sample sampling ratio of tree-building is 0.8.

CNN-LSTM-Attention is used as the base learner to train the timing features and obtain the prediction results of the timing features. The CNN used in the experiment is a one-dimensional convolution. The first convolutional layer has 64 filters, 20 kernels, and relu activation function, the pooling layer is max-pooling. And the second layer has 32 filters, and the others are the same. The number of units of the hidden layer of LSTM is 512, the dropout of LSTM is 0.2, and the hidden state is weighted average through attention, and finally the result is input to the fully connected layer. Adam is used to optimize the model during the training process, and the learning rate is 0.00001.

Ridge regression as a meta-learner of Stacking to fuse global features and time-series features, and get the final prediction results. Finally, MSE (Mean Squared Error) and Pea (Pearson Score) are used to evaluate the performance of the Stacking model.

\[
\text{MSE} = \frac{1}{m} \sum_{i=1}^{m} (y_i - \hat{y}_i)^2 \tag{14}
\]

This indicator calculates the average of the sum of squared errors of the predicted data and the original data. The smaller of the value, the better the prediction effect.
This indicator calculates the correlation between the predicted data and the original data. The larger of the value, the more significant the correlation coefficient. The correlation coefficient is between \([-1, 1]\).

### 3.4. Description of Experimental Results

Table 4 and Table 5 are the prediction results of EMODB and RAVDESS databases respectively. Including the results obtained by training four global feature sets based on LightGBM, the results obtained by training four global feature sets based on Adaboost, the results obtained by training timing features based on CNN-LSTM-Attention deep learning model, and the results obtained by training four global feature sets and timing feature based on Stacking model. Experiments based on two speech databases show that the prediction effect after fusing global features and time series features based on the Stacking model is significantly improved. At the same time, the experiments based on the two databases show that the prediction effect based on the fusion of global features and time series features based on the Stacking model is significantly improved. For example, the Pearson correlation coefficients of P, A, and D in the EMODB database are increased by 10%, 4%, and 6%, and the Pearson correlation coefficients of P, A, and D in the RAVDESS database were increased by 10%, 6%, and 11%, respectively. (Note: The experimental method LGBM-09 in the table refers to the result obtained by using LightGBM training based on the IS09_emotion global feature set, and the other is the same).

| Evaluation | MSE   | Pearson |
|------------|-------|----------|
|            | P     | A        | D     |
| **EMODB**  |       |          |       |
| LGBM-09    | 0.156 | 0.036    | 0.113 |
| LGBM-10    | 0.153 | 0.029    | 0.105 |
| LGBM-11    | 0.135 | 0.030    | 0.122 |
| LGBM-12    | 0.170 | 0.030    | 0.082 |
| AD-09      | 0.163 | 0.034    | 0.108 |
| AD-10      | 0.167 | 0.028    | 0.099 |
| AD-11      | 0.143 | 0.029    | 0.087 |
| AD-12      | 0.139 | 0.027    | 0.086 |
| C-L-Att    | 0.136 | 0.030    | 0.096 |
| Stacking   | 0.102 | 0.021    | 0.071 |

| Evaluation | MSE   | Pearson |
|------------|-------|----------|
|            | P     | A        | D     |
| **RAVDESS**|       |          |       |
| LGBM-09    | 0.119 | 0.076    | 0.157 |
| LGBM-10    | 0.113 | 0.071    | 0.148 |
| LGBM-11    | 0.118 | 0.075    | 0.156 |
| LGBM-12    | 0.125 | 0.072    | 0.153 |
| AD-09      | 0.163 | 0.086    | 0.223 |
| AD-10      | 0.139 | 0.082    | 0.207 |
| AD-11      | 0.151 | 0.082    | 0.217 |

\[
Pea = \frac{\text{cov}(XY)}{\delta x \delta y} = \frac{\mathbb{E}[(X-\mu_x)(Y-\mu_y)]}{\delta x \delta y}
\]
4. Conclusion

Based on the results and discussions presented above, the conclusions are obtained as below:

(1) Based on the EMODB database show that Adaboost as the base learner trains the four global feature sets and obtains the best prediction results, and CNN-LSTM-Attention as the base learner trains timing features and obtains better prediction results.

(2) Based on the RAVDESS database show that LightGBM as the base learner trains the four global features and obtains the best prediction results, and CNN-LSTM-Attention as the base learner trains timing features and obtains better prediction results.

(3) At the same time, experiments based on two speech databases show that the prediction effect after fusing global features and time series features based on the Stacking model is significantly improved. At the same time, the experiments based on the two databases show that the prediction effect based on the fusion of global features and time series features based on the Stacking model is significantly improved. For example, the Pearson correlation coefficients of P, A, and D in the EMODB database are increased by 10%, 4%, and 6%, and the Pearson correlation coefficients of P, A, and D in the RAVDESS database were increased by 10%, 6%, and 11%, respectively.

Acknowledgments

This work was supported by teachers from the School of Automation (School of Artificial Intelligence) and the Smart City Center of Hangzhou Dianzi University.

References

[1] Kerkeni L, Serrestou Y, Mbarki M, et al. A review on speech emotion recognition: Case of pedagogical interaction in classroom[C]// International Conference on Advanced Technologies for Signal & Image Processing. IEEE, 2017.

[2] Boril H, Sadjadi SO, Kleinschmidt T, Hansen JHL. Analysis and detection of cognitive load and frustration in drivers’ speech. In: Kobayashi T, Hirose K, Nakamura S, eds. Proc. of the 2010 INTERSPEECH. Chiba: ISCA, 2010. 502-505.

[3] Wang WS, Wu JB. Emotion recognition based on CSO&SVM in e-learning. In: Proc. of the 2011 7th Int'l Conf. on Natural Computation (ICNC). Beijing: IEEE Computer Society, 2011. 566-570. [doi: 10.1109/ICNC.2011.6022071]

[4] Marchi E, Schuller B, Battiner A, Fridenzon S, Tal S, Golan O. Emotion in the speech of children with autism spectrum conditions: Prosody and everything else. In: Proc. of the 2012 Workshop on Child, Computer and Interaction (WOCCI 2012), Satellite Event of INTERSPEECH. Portland: ISCA, 2012

[5] Nwe T L, Foo S W, De Silva L C. Speech emotion recognition using hidden Markov models[J]. Speech Communication, 2003, 41(4): 603-623.

[6] Mao Q, Wang X, Zhan Y. Speech emotion recognition method based on improved decision tree and layered feature selection[J]. International Journal of Humanoid Robotics, 2010, 7(02): 245-261.

[7] Jain M, Narayan S, Balaji P, et al. Speech emotion recognition using support vector machine[J]. arXiv preprint arXiv:2002.07590, 2020.

[8] Kim J, Saurous R A. Emotion Recognition from Human Speech Using Temporal Information and Deep Learning[C]// Interspeech 2018. 2018.

[9] Zhao J, Mao X, Chen L. Speech emotion recognition using deep 1D & 2D CNN LSTM networks[J]. Biomedical Signal Processing and Control, 2019, 47: 312-323.

[10] Yoon S, Byun S, Jung K. Multimodal speech emotion recognition using audio and text[C]// 2018
IEEE Spoken Language Technology Workshop (SLT). IEEE, 2018: 112-118. 

[11] Chen M, He X, Yang J, et al. 3-D Convolutional Recurrent Neural Networks With Attention Model for Speech Emotion Recognition[J]. IEEE Signal Processing Letters, 2018, 25(10):1440-1444. 

[12] Xu H, Zhang H, Han K, et al. Learning alignment for multimodal emotion recognition from speech[J]. arXiv preprint arXiv:1909.05645, 2019. 

[13] Wen-Jing H, Hai-Feng L I, Hua-Bin R, et al. Review on Speech Emotion Recognition[J]. Journal of Software, 2014. 

[14] Chen Y L, Cheng Y F, Chen X Q et al. Speech emotionestimation in PAD 3D emotion space [J]. Journal of Harbin Institute of Technology, 2018, 05(011):160-166. (in chinese) 

[15] Sun Y, Hun Y X, Zhang X Y et al. Emotion dimension PAD prediction for emotion speech recognition [J]. Journal of Zhejiang University: Engineering Edition, 2019, 53(10). (in chinese) 

[16] Wu D P, Zhang Z L, Cao T T. Research on Stability Classifier Combination Algorithm Based on Stacking Strategy[J]. Small Microcomputer System, 2019, 040(005):1045-1049.(in chinese) 

[17] Wolpert, David H. "Stacked generalization." Neural networks 5.2 (1992): 241-259. 

[18] Ke G L, Meng Q, Finley T, et al. LightGBM: A Highly Efficient Gradient Boosting Decision Tree [C]//Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4 — 9 December 2017, Long Beach, CA, USA 

[19] Andres B. Gorka U, Perez J M, et al. Smart optimization of a friction-drilling process based on boosting ensembles[J]. Journal of Manufacturing Systems, 2018, 48: 108 — 121 

[20] Feng-Qi S. New Rock Burst Prediction Modeling Based on Ensemble Neural Network[J]. journal of jilin university(information science edition), 2009. 

[21] Burkhardt F, Paeschke A, Rolfes M, et al. A database of German emotional speech [C]//Proceedings of Interspeech 2005. Lisbon: ISCA, 2005: 1 517-1 520. 

[22] Livingstone S R, Peck K, Russo F A. Ravdess: the ryerson audio-visual database of emotional speech and song [C]//Proceedings of the 22nd Annual Meeting of the Canadian Society for Brain, Behaviour and Cognitive Science (CSBBCS). Kingston:CSBBCS, 2012:71-72. 

[23] Eyben F, Wollmer M, Schuller B. openSMILE—The Munich versatile and fast open-source audio feature extractor. In: Proc. of the 2010 ACM Multimedia. Firenze, 2010. 1459-1462. [doi: 10.1145/1873951.1874246] 2017: 2227-2231. 

[24] Björn Schuller, Stefan Steidl, Anton Batliner. The INTERSPEECH 2009 Emotion Challenge[J]. interspeech, 2009:312--315. 

[25] Björn Schuller, Stefan Steidl, Anton Batliner, et al. The INTERSPEECH 2010 paralinguistic challenge[C]//Proceeding of the 11th Annual Conference of the International Speech Communication Association. Makuhari, Japan:ISCA, 2010:2795-2798. 

[26] Björn Schuller, Stefan Steidl, Anton Batliner, et al. INTERSPEECH 2011 The INTERSPEECH 2011 Speaker State Challenge *[C]// Interspeech, Conference of the International Speech Communication Association. DBLP, 2013. 

[27] Björn Schuller, Stefan Steidl, Anton Batliner, et al. The INTERSPEECH 2012 speaker trait challenge. 13th Annual Conference of the International Speech Communication Association 2012, INTERSPEECH 2012. 1.