A Broad Learning System to Enhance Performance of Speech Emotion Recognition

Zhuofan Xu\textsuperscript{1,a,†} and Xu Yang\textsuperscript{2,b,†} \\
\textsuperscript{1}School of Computer Science and Technology, Beijing Institute of Technology, Beijing, China. \\
\textsuperscript{2}School of Computer Science and Technology, Beijing Information Science and Technology University, Beijing, China. \\
\textsuperscript{a}120180872@bit.edu.cn, \textsuperscript{b}2017011977@mail.bistu.edu.cn \\
†These authors contributed equally.

Abstract: Speech emotion recognition (SER) is a difficult task because emotions are subjective and recognizing the affective state of the speaker is challenging. To tackle this issue, Broad Learning System is presented to balance the training of networks that are substantially faster than those used previously. Furthermore, we performed experiments on the standard IEMOCAP dataset and achieved the state-of-the-art performance in terms of weighted accuracy and unweighted accuracy. Taken together, the experimental results demonstrated that applying Broad Learning System to SER is reasonable and useful.

1. Introduction

SER has already been widely and effectively applied to electronic equipment, which can help computers in judging human emotions. However, SER is a difficult task because emotions are subjective and recognizing the affective state of the speaker is challenging. Audios is annotated by people with a different perception, and there is no fixed standard for how to classify or measure them.

In the past, different approaches have been explored in SER. Most speech emotion techniques have been utilized to extract discriminative features from speech signals. Features extracted by signal processing techniques (such as cepstral or prosodic contours) have been performed advanced on this task [1, 2]. Moreover, statistical modeling techniques, such as the hidden Markov model (HMM) and the Gaussian mixture model (GMM), have been generally employed for emotion recognition [3]. In Mao et al. and Chen et al.’s papers[4,5], CNN was first utilized to learn affective salient features for SER and achieved superior performances on certain datasets in comparison with other methods. Furthermore, a deep belief network developed by Schmidt et al. [6,7], extracts high-level emotional feature representations from the magnitude spectra, which obtains better performance compared to the traditional acoustic features. Also, Lee et al. used an RNN model to learn long-range temporal relationships in SER [8-10]. Recently, Trigeorgis et al. applied the raw audio samples to train a convolutional RNN (CRNN), which aims to predict continuous arousal/valence space [11, 12].

This paper proposes a novel architecture called Broad Learning System (BLS) for SER. The major contributions of this work are summarized as follows.

1) It is the first time to apply the BLS network in speech emotion recognition.
2) Based on the BLS framework, the model effect of MFCC and Mel-spectrogram as input is explored.

3) The running speed is greatly accelerated, the parameter amount is reduced, and it can run on the CPU, enhancing the real-time performance of network classification.

The rest of this paper is organized as follows. In Section 2, the details of the proposed BLS are given. Section 3 compares the performance of the model using MFCC and Mel-spectrogram as input. Also, experimental results are reported and analyzed. Finally, discussions and conclusions are given in the last section.

2. Model Formulation

2.1. Mel_Spectrogram and Mel Frequency Cepstrum Coefficient

Mel-frequency cepstral coefficients (MFCCs) are coefficients that collectively make up a mel-frequency cepstrum (MFC).

\[ c(n) = \text{IDFT}(\log|E(\omega)| + \log|H(\omega)|) \]  

The mel-frequency cepstral coefficients (MFCCs) of a signal are a small set of features that can be utilized to concisely describe a spectral envelope's overall shape. MFCC alone can be regarded as a feature for speech recognition.

A mel spectrogram is a spectrogram where the frequencies are converted to the mel scale. A piece of audio is firstly Fourier transformed to obtain spectrogram, then triangular filtering is used to obtain mel spectrogram, and finally, cepstrum is used to obtain MFCC. In this process, the dimension of the feature is continuously reduced.

2.2. Broad Learning System

The proposed BLS is constructed based on the characteristics of the flatted functional-link networks. As shown in Fig. 1, Broad Learning Model first maps the inputs to construct a set of mapped features.

\[ \arg\min_{w} \|AW - Y\|_F^2 + \lambda\|W\|_F^2 \]  

with

\[ W = (\lambda I + AA^T)A^T Y \]  

To boost the performance, BLS insert additional enhancement nodes without retraining the whole network from the beginning. The connections of all the mapped features and the enhancement nodes are fed into the output. Ridge regression of the pseudoinverse is designed to find the desired connection weights. After the broad expansion with added mapped features and enhancement nodes via incremental learning, and the structure may have a risk of being redundant due to poor initialization or redundancy in the input data. The simplification can be done using several methods. The overall BLS is given in Fig. 1.

![Illustration of overall Broad Learning System](image-url)
The functional-link network model is illustrated in Fig. 1, denoted by the matrix $X | \xi(W_h + \beta_h)$, where $X$ is the expanded input matrix consisting of all input vectors combined with enhancement components. Compared with the classic model, this the advantage of our model is simple, fast, and easy to update.

2.3. Summary and Optimization
In the overall modeling process, we first obtain info in sound fragments and joint them with corresponding labels extracted from text files. Subsequently, we read the wave data and extract detailed features with Mel_Spectrogram and Mel Frequency Cepstrum Coefficient. Hence, we can obtain two different models, with relevant data stored in mat form files, which can offer abundant data for contrastive analysis between them. In this process, we adjust seven groups of frame numbers operated after Fast Fourier Transform.

According to the statistics and analysis of our datasets, and considering our analyzing efficiency, we abort adding feature enhance nodes in our Broad Learning System frame. To fit our broad learning system, we reshape our originated data and adjust the corresponding regularization coefficient to get a better training effect. To monitor the prediction effect faced with sound fragments containing different emotions, we use the same dataset as the training data and set fragments with different emotion labels as the test data. In addition, we introduce specific evaluation metrics to overview details of each prediction by our model. We also use weighed accuracy and unweighted accuracy to judge the performance of the two models. The best performance of the proposed model is achieved as the following conditions: including Mel Frequency Cepstrum Coefficient, frame number operated after Fast Fourier Transform at 512 and without adding feature enhance node in our Broad Learning System frame, which has a giant progress compared with traditional model with Mel_Spectrogram and Deep Learning System model.

3. Experiments

3.1. Datasets
In this work, we use the Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset, which contains abundant multimodal emotion descriptions of natural utterances. The corpus includes five sessions of utterances between two speakers (one male and one female). The emotional category for each utterance was annotated by three people. First, we extract all data that were labeled as the specified emotions. Then, we do one-hot coding on the four emotion labels. The final dataset contains 4,441 utterances in total (1,059 angry, 595 happy, 1,084 sad, and 1,703 natural), and the detailed results are visually displayed in Fig. 2.

![Figure.2 Ratios of sample number](image)
3.2. Evaluation Metrics

To illustrate the possibility for further extension, we introduce a confusion matrix to reveal the detail of the training result. The ratio between the number of the predicted label speech set on the corresponding actual label speech set and the number of the corresponding actual label speech set is marked on each cube and measured in shades of green. As shown in Fig. 3, the number on the left oblique line reveals the ratio of precision.

\[
\text{Precision} = \frac{TP}{(TP + FP_1 + FP_2 + FP_3)}
\]  

(4)

And it is apparent that the precision of prediction on “Angry”, “Natural” and “Sad” is pretty high, which means/indicates the model has obtained a promising fitting effect. Furthermore, all results are achieved with each frame taking Fast Fourier Transform of set at 512.

![Confusion matrix](image)

Figure.3 Confusion matrix of model performance, where each row presents the truth emotion.

In contrast, Mel_Spectrogram and Mel Frequency Cepstrum Coefficient were introduced to develop two independent models and combined them with the modified Broad Learning System. As a consequence, the model applied with Mel Frequency Cepstrum Coefficient obtain better performance in predicting the sound fragments with “Angry”, “Natural” and “Happy”. Interestingly, the model introducing Mel_Spectrogram has a slight advantage in discerning sad sound fragments.

Sightseeing refocusing on a single model with MFCC, we can apparently find happy sound fragments are much imperceptible compared with sound fragments with other labels. The reason of the model being inaccurate in the happy class can be mainly summarized as two aspects: Firstly, we have fewer samples with “Happy” label in Interactive Emotional Dyadic Motion Capture datasets, which increases the difficulty of fragment discrimination. Additionally, the characteristic with happy fragments is short, and they are similar to the characteristic in natural sound fragments. Thus, these fragments are easily recognized to be neutral and smooth by our model.

3.3. Experimental Result Analysis

After exploring the detailed prediction result with the assistance of evaluation metrics, we introduce more evaluation indicators to give a panoramic display of our results. Overall, we put our statistics into a simple table below. For either the weighed accuracy in prediction combined with all sound fragments or the unweighted accuracy in prediction sound fragments with different labels, the model with Mel Frequency Cepstrum Coefficient is more advanced than the model Mel_Spectrogram. In Table 1, the weighted accuracy is higher than the unweighted accuracy, which indicates we still have a possible promotion in further model modification to balance the stability when we extrapolate different emotions from different sound fragments.
Table.1 The WA and UA of mel-spec and MFCC

| Type     | WA    | UA    |
|----------|-------|-------|
| mel_spec | 0.65030 | 0.58230 |
| MFCC     | 0.69106 | 0.62257 |

Besides, another important consideration to modify our model is the frame number operated after Fast Fourier Transform. It is difficult to explain how to choose an appropriate parameter, and we introduce a bar chart to make a thorough inquiry. As illustrated in Fig. 4, the interval between the binary logarithm of our investigation value is one, and we find for both two models, the best-fit value is 512.

![Figure.4 Accuracy results with MFCC and mel_spec as the feature input](image)

4. Conclusion
This paper creates an innovative combination of Mel Frequency Cepstrum Coefficient and Broad Learning System to characterize different emotions hidden behind tremendous numbers of sound fragments. The experimental results demonstrate that our accuracy can be improved up to 69.11% with each frame taking Fast Fourier Transform of N equals 512. In the future, different goals related to Broad Learning System can be achieved based on the current work. For example, in the investigation and research on emotion analysis, Broad Learning System can distinguish various emotions from these four basic emotions. Furthermore, better use of the Broad Learning System can assist us in traditional machine vision areas to access progress.

References
[1] Weihui Dai, Dongmei Han, Yonghui Dai, and Dongrong Xu, “Emotion recognition and affective computing on vocal social media,” *Information & Management*, vol. 52, no. 7, pp. 777–788, 2015.
[2] C. Jan et al., “Attention-based models for speech recognition,” *Comput. Sci.*, vol. 10, no. 4, pp. 429–439, 2015.
[3] Chi-Chun Lee, Emily Mower, Carlos Busso, Sungbok Lee, and Shrikanth Narayanan, “Emotion recognition using a hierarchical binary decision tree approach,” *Speech Communication*, vol. 53, no. 9-10, pp. 1162–1171, 2011.
[4] Q. Mao, M. Dong, Z. Huang, and Y. Zhan, “Learning salient features for speech emotion recognition using convolutional neural networks,” *IEEE Trans. Multimedia*, vol. 16, no. 8, pp. 2203–2213, 2014.
[5] C. L. P. Chen, C.-Y. Zhang, L. Chen, and M. Gan, “Fuzzy restricted Boltzmann machine for the enhancement of deep learning,” *IEEE Trans. Fuzzy Syst.*, vol. 23, no. 6, pp. 2163–2173, 2015.
[6] Z. Yu et al., “Incremental semi-supervised clustering ensemble for high dimensional data clustering,” *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 3, pp. 701–714, 2016.
[7] Weihui Dai, Dongmei Han, Yonghui Dai, and Dongrong Xu, “Emotion recognition and affective computing on vocal social media,” *Information & Management*, vol. 52, no. 7, pp. 777–788, 2015.

[8] W. Yoon and K. Park, “Building robust emotion recognition system on heterogeneous speech databases,” *IEEE Transactions on Consumer Electronics*, vol. 57, no. 2, pp. 747-750, 2011.

[9] S. Ramamohan and S. Dandapat, “Sinusoidal modal-based analysis and classification of stressed speech,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 3, pp. 737-746, 2006.

[10] C. Wu and W. Liang, “Emotion recognition of affective speech based on multiple classifiers using acoustic-prosodic information and semantic labels,” *IEEE Transactions on Affective Computing*, vol. 2, no. 1, pp. 10-21, 2006.

[11] Y. Xie *et al.*, “Speech emotion classification using attention-based LSTM,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 11, pp. 1675-1685, 2019.

[12] Carlos Busso *et al.*, “Iemocap: Interactive emotional dyadic motion capture database,” *Language resources and evaluation*, vol. 42, no. 4, pp. 335, 2008.