Utterance-Based Audio Sentiment Analysis Learned by a Parallel Combination of CNN and LSTM

Ziqian Luo, Hua Xu*, Feiyang Chen

State Key Laboratory of Intelligent Technology and Systems, Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China

Abstract

Audio Sentiment Analysis is a popular research area which extends the conventional text-based sentiment analysis to depend on the effectiveness of acoustic features extracted from speech. However, current progress on audio sentiment analysis mainly focuses on extracting homogeneous acoustic features or doesn't fuse heterogeneous features effectively. In this paper, we propose an utterance-based deep neural network model, which has a parallel combination of Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) based network, to obtain representative features termed Audio Sentiment Vector (ASV), that can maximally reflect sentiment information in an audio. Specifically, our model is trained by utterance-level labels and ASV can be extracted and fused creatively from two branches. In the CNN model branch, spectrum graphs produced by signals are fed as inputs while in the LSTM model branch, inputs include spectral features and cepstrum coefficient extracted from dependent utterances in an audio. Besides, Bidirectional Long Short-Term Memory (BiLSTM) with attention mechanism is used for feature fusion. Extensive experiments have been conducted to show our model can recognize audio sentiment precisely and quickly, and demonstrate our ASV are better than traditional acoustic features or vectors extracted from other deep learning models. Furthermore, experimental results indicate that the proposed model outperforms the state-of-the-art approach by 9.33% on Multimodal Opinion-level Sentiment Intensity dataset (MOSI) dataset.

Keywords: `Audio Sentiment Analysis`, Feature Fusion, Neural Network, Signal Processing
2018 MSC: 00-01, 99-00

1. Introduction

Sentiment Analysis is a well-studied research area in Natural Language Processing (NLP) [1], which is the computational study of peoples' opinions, sentiments, appraisals, and attitudes towards entities such as products, services, organizations and so on [2, 3]. Traditional sentiment analysis methods are mostly based on texts [4, 5], with the rapid development of communication technology, abundance of smartphones and the rapid rise of social media, large amounts of data are uploaded by web users in the form of audios or videos [6, 7], rather than texts [8]. Interestingly, a recent study shows that voice-only as modality seems best for humans empathetic accuracy as compared to video-only or audiovisual communication [9]. In fact, audio sentiment analysis is a difficult task due to the complexity of audio signal. It is generally known that speech is the most convenient and natural medium for human communication, not only carries the implicit semantic information, but also contains rich affective information [10]. Therefore, audio sentiment analysis, which aims to correctly analyze the sentiment of the speaker from speech signals, has drawn a great deal of attention of researchers.

*Hua Xu
Email addresses: luoziqian98@gmail.com (Ziqian Luo), xuhua@tsinghua.edu.cn (Hua Xu), chenfeiyang999@gmail.com (Feiyang Chen)

In recent years, there are three main methods for audio sentiment analysis. Firstly, utilizes Automatic Speech Recognition (ASR) technology to convert speech into texts, following by conventional text-based sentiment analysis systems [11]. Secondly, adopts a generative model operating directly on the raw audio waveform [12]. Thirdly, focuses on extracting signal features from the raw audio files [13], which well captures the tonal content of a music, and has been proved to be more effective than original audio spectrums descriptors such as Mel-Frequency Cepstrum Coefficients(MFCC).

However, for converting speech into texts, by recognizing each word said by the person in an audio, change them into the word embedding and use some techniques in NLP, like Term FrequencyInverse Document Frequency (TF-IDF) and Bag of Words (BOW) model [5]. The result is not always accurate, because sentiment detection accuracy depends on being able to reliably detect a very focused vocabulary in the spoken comments [14]. Furthermore, when the voice is transferred to the text, some sentiment-related signal characteristics are also lost, resulting in a decrease in the accuracy of the sentiment classification. As for extracting from the raw audio files through human works and then being put into the Support Vector Machine(SVM) classifier for classification, those methods require lots of human work and are heavily dependent on language types.

Luckily, along with the success of deep learning in many other application domains, deep learning is also popularly used in audio sentiment analysis in recent years [15, 16, 17]. More recently, [18] directly use the raw audio samples to train a Convolutional Recurrent Neural Network (CRNN) to predict continuous arousal /valence space. [19] study the use of deep learning to automatically discover emotionally relevant features from speech. They propose a novel strategy for feature pooling over time which uses local attention in order to focus on specific regions of a speech signal that are more emotionally salient. [20] use an attentive convolutional neural network with multi-view learning objective function and achieved state-of-the-art results on the improvised speech data of IEMOCAP [21]. [22] propose to use Deep Neural Networks (DNN) to encode each utterance into a fixed-length vector by pooling the activations of the last hidden layer over time. The feature encoding process is designed to be jointly trained with the utterance-level classifier for better classification. [23] propose a 3-D attention-based convolutional recurrent neural networks to learn discriminative features for speech emotion recognition, where the Mel-spectrogram with deltas and delta-deltas are creatively used as input. But most of the previous methods still either considered only one single audio feature [23] or high-dimensional vectors [24, 25] from one homogeneous feature [26], and did not effectively extract and fuse audio features.

We believe the information extracted from a single utterance must have dependency on its context. For example, a flash of loud expression may not indicate a person has a strong emotion since it maybe just caused by a cough while continuous loud one is far more likely to indicate the speaker has a strong emotion.

In this paper, based on a large number of experiments, we extract the features of each utterance in an audio through the Librosa toolkit, and obtain four most effective features representing sentiment information, merge them by adopting a BiLSTM with attention mechanism. Moreover, we design a novel model called Audio Feature Fusion-Attention based CNN and RNN (AFF-ACRNN) for audio sentiment analysis. Spectrum graphs and selected traditional acoustic features are fed as input in two separate branches, we can obtain a new fusion of audio feature vector before the softmax layer, which we call the Audio Sentiment Vector (ASV). Finally, the output of the softmax layer is the class of sentiment.

Major contributions of the paper are that:

- We propose an effective AFF-ACRNN model for audio sentiment analysis, through combining multiple traditional acoustic features and spectrum graphs to learn more comprehensive sentiment information in audio.
- Our model is language insensitive and pay more attention to acoustic features of the original audio rather than words recognized from the audio.
- Experimental results indicate that the proposed method outperforms the state-of-the-art methods [26] on Multimodal Corpus of Sentiment Intensity dataset(MOSI) and Multimodal Opinion Utterances Dataset(MOUD).

The rest of the paper is organized as follows. In the following section, we will review related work. In Section 3, we will exhibit more details of our methodology. In Section 4, experiments and results are presented, and conclusion follows in Section 5.

2. Related Work

Current state-of-the-art methods for audio sentiment analysis are mostly based on deep neural network. In this section, we briefly present the advances on audio sentiment analysis task by utilizing deep learning, and then we give a summary on the progress of extracting the audio feature representation.

2.1. Long Short-Term Memory (LSTM)

It has been demonstrated that LSTM [27] are well-suited to make predictions based on time series data, by utilizing a cell to remember values over arbitrary time intervals and the three gates(input gate i, output gate o, forget gate f) to regulate the flow of information into and out of the cell, which can be described as follows:

$$\mathrm{f}_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{1}$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{2}$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{3}$$

where $h_t = o_t * \tanh(C_t)$ is the output of the last cell and x_t is the input of current cell. Besides, the current cell state C_t can be updated by the following formula:

$$\widetilde{C}_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \tag{4}$$

$$C_t = f_t * C_{t-1} + i_t * \widetilde{C}_t \tag{5}$$

where C_{t-1} stands for the previous cell state.

One of the most effective variant of LSTM is the bidirectional LSTM. Each input sequence will be fed into both the forward and backward LSTM layers and thus a hidden layer receives an input by joining forward and backward LSTM layers.

2.2. Convolutional Neural Network (CNN)

CNN [28] are well-known for extracting features from a image by using convolutional kernels and pooling layers to emulates the response of an individual to visual stimuli. Moreover, CNN has been successfully used not only for computer vision, but also for speech [29]. For speech recognition, CNN is proved to be robust against noise compared to other DL models [30].

2.3. Audio Feature Representation and Extraction

Researchers have found pitch and energy related features playing a key role in affect recognition [26]. Other features that have been used by some researchers for feature extraction include formants, MFCC, root-mean-square energy, spectral centroid and tonal centroid features. MFCC is the most recognized feature among the four and the mapping between the real frequency scale (Hz) and the perceived frequency scales (mels) is approximately linear below 1 KHz and logarithmic at higher frequency, and such an approximation is usually adopted in speech recognition. There relationship is modeled as the formula suggested below:

$$F_{mel} = 2595 log_{10}(1 + \frac{F_{Hz}}{700}) \tag{6}$$

During the speech production, there are several utterances and for each utterance, the audio signal can be divided into several segments. Global features are calculated by measuring several statistics, e.g., average, mean, deviation of the local features. Global features are the most commonly used features in the literature. They are fast to compute and, as they are fewer in number compared to local features, the overall speed of computation is enhanced [31]. However, there are some drawbacks of calculating global features, as some of them are only useful to detect affect of high arousal, e.g., anger and disgust. For lower arousal, global features are not effective, e.g. global features are less prominent to distinguish between anger and joy. Global features also lack temporal information and dependence between two segments in an utterance. In a recent study [32], a new acoustic feature representation, denoted as deep spectrum features, derived from feeding spectrum graphs through a very deep image classification CNN and forming a feature vector from the activation of the last fully connected layer. Librosa [33] is an open-source python package for music and audio analysis which is able to extract all the key features as elaborated above.

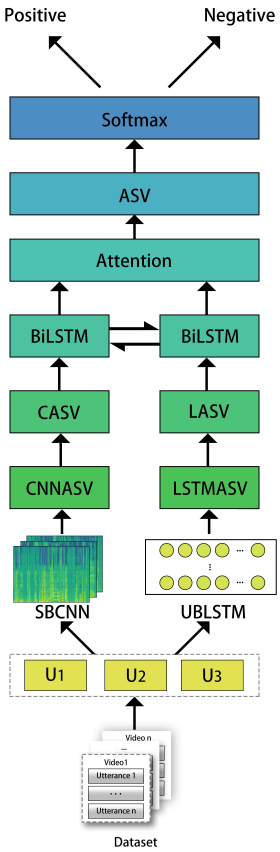

Figure 1: Overview of AFF-ACRNN Model

3. Methodology

In this section, we describe the proposed AFF-ACRNN model for audio sentiment analysis in details. We firstly introduce an overview of the whole neural network architecture. After that, two separate branches of AFF-ACRNN will be explained in details. Finally, we talk about the fusion mechanism used in our model.

3.1. Model—AFF-ACRNN

We concentrate on a model that has two parallel branches, the Utterance-Based BiLSTM Branch (UB-BiLSTM) and the Spectrum-Based CNN Branch (SBCNN), whose core mechanisms are based on LSTM and CNN. One branch of proposed model uses the BiLSTM to extract temporal information between adjacent utterances, another branch uses the renowned CNN based network to extract features from spectrum graph that sequence model cannot achieve. Furthermore, audio feature vector of each piece of utterance is the input of the proposed neural network that based on Audio Feature Fusion (AFF), we can obtain a new fusion audio feature vector before the softmax layer, which we call the Audio Sentiment Vector (ASV). Finally, the output of the softmax layer produces our final sentiment classification results, as shown in Figure 1.

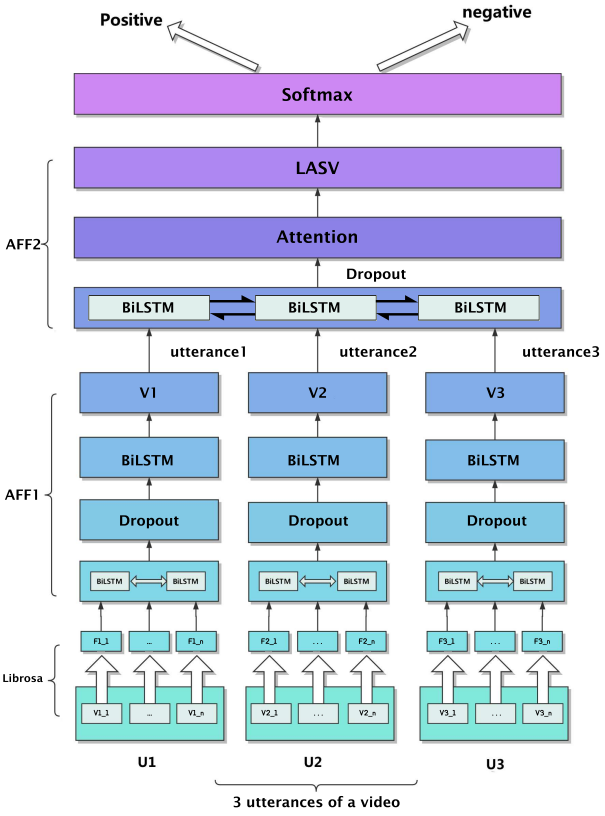

Figure 2: Overview of Our UB-BiLSTM Model

3.2. Audio Sentiment Vector (ASV) from Audio Feature Fusion (AFF)

3.2.1. LSTM Layers

The hidden layers of LSTM have self-recurrent weights. These enable the cell in the memory block to retain previous information [34]. Firstly, we separate the different videos and take three continuous utterances (e.g. u_1, u_2, u_3) in one video at a time. Among them, for each utterance (e.g. u_1), we extract its internal acoustic features through the librosa toolkit, say $f_{1_1}, f_{1_2} ... f_{1_n}$, and then trained by two layers of BiLSTM in AFF1 to obtain the extracted features from the traditional acoustic feature. Therefore, three utterances are corresponding to three more efficient and representative vectors v_1, v_2, v_3, as the inputs to BiLSTM in AFF2. AFF2 effectively combines the contextual information between adjacent utterances, and then subtly acquires the utterance that has the greatest impact on the final sentiment classification through the attention mechanism. Finally, after the dropout layer, a more representative vector, named as LASV is extracted by our LSTM framework before the softmax layer, as shown in Figure 2. The process is described in LSTM branch procedure in Algorithm 1.

Algorithm 1 Related Procedure

procedure LSTM BRANCH
  for i:[0,n] **do**
    $f_i = getAudioFeature(u_i)$
    $ASV_i = getASV(f_i)$
  end for
  for i:[0,M] **do** //M is the number of videos
    $input_i = GetTopUtter(v_i)$
    $u_{f_i} = getUtterFeature(input_i)$
  end for
  $shuffle(v)$
end procedure
procedure CNN BRANCH
  for i:[0,n] **do**
    $x_i \leftarrow$ *get SpectrogramImage*(u_i)
    $c_i \leftarrow$ *CNNModel*(x_i)
    $l_i \leftarrow$ *BiLSTM*(c_i)
  end for
end procedure
procedure FIND CORRESPONDING LABEL
  for i:[0:2199] **do**
    $rename(u_i)$ // for better order in sorting
    $NameAndLabel = createIndex(u_i)$
    // A dictionary [utterance Name: Label]
  end for
  $Label_x = NameAndLabel(u_x)$
end procedure

3.2.2. CNN Layers

Similar to the UB-BiLSTM model proposed above, we extracted the spectrum graph of each utterance through the Librosa toolkit and use it as the input of our CNN branch. After a lot of experiments, we found that the audio feature vector learned by the ResNet152 network structure has the best effect on the final sentiment classification, so we choose the ResNet model in this branch. The convolutional layer performs 2-dimensional convolution between the spectrum graph and the predefined linear filters. To enable the network to extract complementary features and learn the characteristics of input spectrum graph, a number of filters with different functions are used. A more refined audio feature vector is obtained through deep convolutional neural network, and then put into the BiLSTM layer to learn related sentiment information between adjacent utterances. Finally, before the softmax layer, we get another effective vector, named as CASV, extracted by our CNN framework, as shown in Figure 3. The process is described in CNN branch procedure in Algorithm 1.

3.2.3. Fusion Layers

Through the LSTM and CNN branches proposed above, we can extract two refined audio sentiment vectors, LASV and CASV for each utterance. We use these two kinds of vectors in parallel as the input of BiLSTM in AFF-ACRNN model. While effectively learning the relevant sentiment information of adjacent utterance, we extract the Audio Sentiment Vector (ASV) that has the greatest influence on the sentiment classification in the three utterances through the action of the attention mechanism. Finally, the final sentiment classification result is obtained by softmax layer. In [35], Long-Term Recurrent Convolution Network (LRCN) model was proposed for visual recognition. LRCN is a consecutive structure of CNN and LSTM. LRCN processes the variable-length input with a CNN, whose outputs are fed into LSTM network,

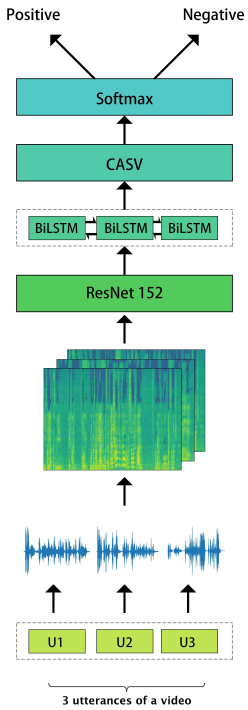

Figure 3: Overview of Our ResNet152 CNN Model

which finally predicts the class of the input. In [36], a cascade structure was used for voice search. Compared to the method mentioned above, the proposed network forms a parallel structure in which LSTM and CNN accept different inputs separately. Therefore, the Audio Sentiment Vector (ASV) can be extracted more comprehensively, and a better classification result can be got.

3.3. Feature Fusion base on Attention Mechanism

Inspired by human visual attention, the attention mechanism is proposed by [37] in machine translation, which is introduced into the Encoder-Decoder framework to select the reference words in source language for words in target language. We use the attention mechanism to preserve the intermediate output of the input sequence by retaining the LSTM encoder, and then a model is trained to selectively learn these inputs and to correlate the output sequences with the model output. Specifically, when we fuse the features, each phoneme of the output sequence is associated with some specific frames in the input speech sequence, so that the feature representation that has the greatest influence on the final sentiment classification can be obtained, and finally obtain a fused Audio Feature Vector. At the same time, attention mechanism behaves like a regulator since it can judge the importance of the contribution by adjacent relevant utterances for classifying the target utterance. Indeed, it is very hard to tell the sentiment of a single utterance if you do not concern its contextual information. However, you will also make a wrong estimation if contextual information is overly concerned. More specifically, in Figure 2, let A_x be the X_{th} attention network for utterance U_x, the corresponding attention weight vector is α_x weighted hidden representation is R_x, we have:

$$P_x = tanh(W_h[x] \cdot H) \tag{7}$$

$$A_x = softmax(w[x]^T \cdot P_x) \tag{8}$$

$$R_x = H \cdot {\alpha_x}^T \tag{9}$$

Final representation for $x_t h$ utterance is:

$${h_x}^* = \tanh(W_m[x] \cdot R_x + W_n[x] \cdot h_x) \tag{10}$$

Where $W_m[x]$ and $W_n[x]$ are weights to be learned while training.

4. Experiments

In this section, we exhibit our experimental results and the analysis of our proposed model. More specifically, our model is trained and evaluated on utterance-level audio from CMU-MOSI dataset [38] and being tested on MOUD [39]. What's more, in order to verify that our proposed model has a strong generalization ability, we also carry out extensive expansion experiments.

4.1. Experiment Setting

Evaluation Metrics We evaluate our performance by weighted accuracy on both 2-class, 5-class and 7-class classification.

$$weighted\ accuracy = \frac{correct\ utterances}{utterances} \tag{11}$$

Additionally, F-Score is used to evaluate 2-class classification.

$$\mathrm{F}_\beta{=}(1+\beta^2) \cdot \frac{precision \cdot recall}{(\beta^2 \cdot precision) + recall} \tag{12}$$

where β represents the weight between precision and recall. During our evaluation process, we set $\beta = 1$ since we regard precision and recall has the same weight thus $F1$-score is adopted.

However, in 5-class and 7-class classification, we use Macro $F1$-Score to evaluate the result.

$$Macro\ F_1{=}\frac{\sum_1^n F_{1n}}{n} \tag{13}$$

where n represents the number of classification and F_{1n} is the $F1$ score on n^{th} category.

4.1.1. Dataset details

CMU-MOSI dataset is rich in sentiment expressions, consisting 2199 opinionated utterances, 93 videos by 89 speakers. The videos address a large array of topics, such as movies, books, and products. Videos were crawled from YouTube and segmented into utterances where each utterance is annotated with scores between -3 (strongly negative) and $+3$ (strongly positive) by five annotators. We took the average of these five annotations as the sentiment polarity and considered three conditions where consists of two classes (positive and negative), five classes (strongly positive, positive, neutral, negative and strongly negative) and seven classes (strongly positive, positive, weakly positive, neutral, strongly negative,negative and weakly negative). Our train/test splits of the dataset are completely disjoint with respect to speakers. In order to better compare with the previous work, similar to [26], we divide the data set by 7:3 approximately, 1616 and 583 utterances are used for training and testing respectively. Furthermore, in order to verify that our model will not be heavily dependent on the language category, we tested it with the Spanish dataset MOUD. MOUD contains product review videos provided by 55 persons. The reviews are in Spanish. The detailed datasets setup is depicted at Table 1.

Datasets	Train		Test	
	utterance	video	utterance	video
MOSI	1616	74	583	19
MOSI→MOUD	2199	93	437	79

Table 1: Datasets Setting. The right arrow means the model is trained and validated on the MOSI and tested on the MOUD

4.1.2. Network structure parameter

Our proposed architecture is implemented based on the open-source deep learning framework Keras. More specifically, for proposed UB-BiLSTM framework, after a lot of experiments, we extracted the most four representative audio features of each utterance in a video through Librosa toolkit, which are MFCC, spectral_centroid, chroma_stft and spectral_contrast respectively. In data processing, we make each utterance one-to-one correspondence with the label and rename the utterance. Accordingly, we extend each utterance to a feature matrix of $256 * 33$ dimensions. The output dimension of the first layer of BiLSTM is 128, and the second layer is 32. The output dimension of the first layer of Dense is 200, and the second is 2.

For the proposed CNN framework, the input images are warped into a fixed size of $512 * 512$. If the bounding box of the training samples provided, we firstly crop the images and then warp them to the fixed size. To train the feature encoder, we follow the fine-tuning training strategy.

In all experiments, our networks are trained by Adam or SGD optimizer. In the LSTM branch, we initiate the learning rate to be 0.0001, and there are 200 epochs in the training part with batch size equals to 30 in each epoch. In the CNN branch, we initiate the learning rate to be 0.001, and there are 200 epochs in training Resnet-152 with batch size equals to 20 in each epoch.

Best Feature Combination	Model	Accuracy(%)		
		2-class	5-class	7-class
Single Type	LSTM	55.12	23.64	16.99
	BiLSTM	55.98	23.75	17.24
Two Types	LSTM	62.26	28.23	21.54
	BiLSTM	63.76	29.77	22.92
Thress Types	LSTM	66.36	32.98	24.66
	BiLSTM	67.02	33.75	25.80
Four Types	**LSTM**	**68.23**	**33.15**	**26.27**
	BiLSTM	**68.72**	**34.27**	**26.82**
Five Types	LSTM	67.86	31.29	25.79
	BiLSTM	67.97	32.66	26.01
Six Types	LSTM	67.88	32.23	26.07
	BiLSTM	68.61	33.97	26.78
Seven Types	LSTM	68.01	33.06	25.99
	BiLSTM	68.67	34.18	26.12

Table 2: Comparison of different feature combinations

4.2. Performance Comparison

4.2.1. Comparison of different feature combinations.

Firstly, we have considered seven types of acoustic features that can best represent an audio,which mainly includes MFCC, root-mean-square energy, spectral and tonal features. A lot of experiments have been done in order to get the best feature combinations with different models on three types of classification. In the binary classification, for example, in order to find which are the best four features among the seven, we have carried out C_7^4 sets of experiments and only the best result of any 4-combination is recorded in bold in Table

Methods	2-class		5-class		7-class	
	Acc(%)	F1	Acc(%)	Macro F1	Acc(%)	Macro F1
LeNet	56.75	55.62	23.67	21.87	15.63	15.12
AlexNet	58.71	57.88	26.43	23.19	19.21	18.79
VGG16	57.88	55.97	27.37	25.78	17.34	16.25
ZFNet	55.37	53.12	21.90	21.38	12.82	11.80
ResNet18	58.94	56.79	25.26	24.63	18.35	17.89
ResNet50	62.52	61.21	28.13	27.04	20.21	20.01
ResNet152	**65.42**	**64.86**	**28.78**	**28.08**	**21.56**	**20.57**

Table 3: Comparison of SBCNN with different structure

Methods	2-class		5-class		7-class	
	Acc(%)	F1	Acc(%)	Micro F1	Acc(%)	Micro F1
UB-Res18	67.19	66.37	33.83	31.97	26.78	25.83
UB-Res50	67.83	66.69	34.21	33.78	27.75	26.41
UB-Res152	68.64	67.94	35.87	34.11	28.15	27.03
UBBi-Res18	68.26	66.25	35.43	33.52	27.63	26.09
UBBi-Res50	69.18	68.22	36.93	34.67	28.11	27.54
UBBi-Res152	**69.64**	**68.51**	**37.71**	**35.12**	**29.26**	**28.45**

Table 4: Comparison of different combinations between SBCNN and UB-BiLSTM

2. What's more, we have also compared the different performance of our LASV extracted from LSTM-based and BiLSTM-based fusion model. As the Table 2 shows, the performance of LASV that extracted from BiLSTM-based model behaves better, since the acoustic information behind may also have impact on the acoustic information previous. It can be seen that the best number of feature combination is four and those four features are MFCC, spectral_centroid, spectral_contrast and chroma_stft. That means the other three features, which are root-mean-square energy, spectral_contrast and tonal centroid may introduce some noises or misleading in our sentiment analysis since all seven types of features do not have the best result.

4.2.2. Comparison of several renowned CNN-based model.

We have compared our CASV performance extracted from the spectral map with several genres of popular models of CNN and its variants: LeNet [40], AlexNet [41], VGG16 [42], ZFNet [43], ResNet [44]. The results are listed in Table 3. As the neural network goes deeper, more representative features can be got from the spectrum graph and that is why ResNet152 has the best performance. It is benefited from the residual unit which will guarantee the network will not degrade when the network goes deeper.

4.2.3. Comparison of different combinations between SBCNN and UB-Bilstm

At last, we have performed fusion experiments between several best SBCNN and UB-BiLSTM and UB-LSTM. More accurately, we choose the three best SBCNN, which are ReSNet18, ResNet50 and ResNet152 to combine with the two kinds of utterance dependent LSTM. The best combination is UB-BiLSTM with Res152. The final result is shown in Table 4.

4.2.4. Comparison with traditional method.

Apart from training deep neural network, a bunch of traditional binary classifiers has been used for sentiment analysis. In order to demonstrate the effectiveness of our model, we firstly compare our model with those traditional methods.

[45] introduced a text-based SVM and Naive Bayes model for binary sentiment classification, thus we test their model on MOUD, rather than MOSI, to make comparison with our model because MOUD has only two sentiment level and each utterance has text record in the dataset.

[46] In this paper, except for SVM, the feature vectors like Mel Frequency Discrete Wavelet Coefficients (MFDWC), MFCC and Linear Predictive Cepstral Coefficients (LPCC) extracted from original record signal are trained with classification algorithm such as Dynamic Time Warping (DTW), Hidden Markov Model (HMM) and Gauss Mixture Model (GMM).

As shown in Table 5, we use weighted accuracy (ACC) and F1-Score to evaluate our results. Especially, for the ACC on MOUD, our proposed model outperforms the best model, SVM classifier, by 11.51%.

Model	MOUD	
	ACC(%)	F1
SVM	57.23	54.83
Naive Bayes	55.72	52.14
GMM	54.66	52.89
HMM	56.63	55.84
DTW	53.92	53.06
AFF-ACRNN	**68.74**	**66.37**

Table 5: Comparison with traditional methods on MOUD

4.2.5. Comparison with the state-of-art.

[26] has introduced a LSTM-based model to utilize the contextual information extracted form each utterance in an video. However, the input of the neural network model only has one type of feature, which is MFCC. This means all the utterance information is merely represented by one single feature. The acoustic information contained by the feature is somewhat duplicated and is bound to omit much sentiment information that might be hidden in many other useful features. What's worse, one type of feature means the input vector should be large enough to make sure that it carries enough information before it is fed into the neural network. This will undoubtedly increase the parameters to be trained in the network and meanwhile, it is time consuming and computation costly.

Our proposed model not only extracts the feature or sentiment vector from four types of traditional recognized acoustic features, have considered utterance dependency, but also extracts the feature from the spectrum graph, which may reveal some sentiment information that acoustic features cannot reflect. The final AFF-ACRNN consists of the best combination of SBCNN and UB-BiLSTM and outperforms the state-of-the-art approach by 9.33% in binary classification on MOSI dataset and by 10.54% on MOUD. The results are shown in Table 6.

Model	ACC(%)		
	MOSI	$\rightarrow$	MOUD
State-of-the-art	60.31		47.20
AFF-ACRNN	**69.64**		**59.74**

Table 6: Comparison with state-of-art result (Poria et al.2017) . The right arrow means the model is trained and validated on the MOSI and tested on the MOUD

We have also run our model on one audio whose length is 10s for 1000 times and the average time to get the sentiment calssification result from input is only 655.94ms which thanks to our concentrated ASV extracted from AFF-ACRNN.

Datasets	Train		Test	
	utterance	video	utterance	video
Sichuan	1734	11	434	3
Cantonese	4123	36	1031	9
Mandarin	19764	75	4941	19

Table 7: Expansion Experiment Datasets Setting

4.3. Experiment Expansion

Furthermore, considering the impact that different languages may have on the generalization capabilities of the proposed model, we experimented with the three largest language-related datasets in the world, which are Chinese, English, and Spanish. As mentioned above, CMU-MOSI is a English dataset, which is rigorously annotated with labels for subjectivity, sentiment intensity, per-frame and per-opinion annotated visual features, and per-milliseconds annotated audio features, but in this paper we only discuss annotated audio features. MOUD is a Spanish dataset, which collects a set of videos from the social media web site YouTube, using several keywords likely to lead to a product review or recommendation. It is worth mentioning that, although Chinese is the most used language in the world, there is currently no public dataset available for experimentation. In this paper, our another contribution is to collate three different Chinese datasets. As is known to us, Chinese is wide-ranging and profound. Therefore, we compiled the three most representative Chinese datasets, which are Mandarin, Cantonese and Sichuan respectively. The detailed datasets setup is depicted at Table 7.

Our Chinese datasets come from major online social media platforms or live video sites, including Weibo, bilibili, Tik tok and so on. The content covers product reviews, movie reviews, shopping feedback and many other aspects of daily life. At the same time, taking into account the complexity of human emotions in reality and the differences in individual emotions, we draw on the annotation method of the CMU-MOSI dataset to find five people with psychology-related professional backgrounds to independently mark the scores, and finally 5 annotations. The scores are averaged to give the final emotional score. Finally, we divide the Chinese dataset into three categories, which are positive, neutral, and negative respectively. The detailed experiment results are shown in Table 8.

Datasets	Model	Accuracy(%)		
		Train	Dev	Test
Sichuan	AFF-ACRNN	60.41	57.11	56.25
Cantonese	AFF-ACRNN	66.29	61.72	60.31
Mandarin	AFF-ACRNN	68.84	64.10	64.08

Table 8: Expansion Experiment on Sichuan, MOSI and Mandarin Datasets

4.4. Discussion

The above experimental results have already shown us that the proposed method has a great improvement in the performance of audio sentiment analysis. In order to get the best structure of our AFF-ACRNN model, we have tested two separate branches respectively, and compare the final AFF-ACRNN with traditional or state-of-art method. Weighted accuracy and F1-Score, Macro F1-Score are used as metrics to evaluate the model's performance. In the UB-Bilstm branch, a lot of experiments have shown that four types of heterogeneous traditional features trained by BiLSTM will have the best result, whose weighted accuracy is 68.72% on MOSI. In the SBCNN branch, we have carried out seven experiments to prove the ResNet152 used in SBCNN will have the best result, for instance, with the weighted accuracy of 65.42% on MOSI, due to its extreme depth and the helpful residual units used to prevent degradation. We selected six best

combinations of SBCNN and UB-BiLSTM and find that the best is ResNet152 used in SBCNN with UB-Bilstm, whose weighted accuracy is 69.42% on MOSI and outperforms not only the traditional classifier like SVM, but also the state-of-the-art approach by 9.33% on MOSI dataset. Attention mechanism is used in both branch to subtly combine the heterogeneous acoustic features and choose the feature vectors that have the greatest impact on the sentiment classification. Furthermore, in the experiment of using MOSI as training set and verification set and MOUD as test set, it also shows that our proposed model has strong generalization ability.

5. Conclusion and Future Work

In this paper, we propose a novel utterance-based deep neural network model termed AFF-ACRNN, which has a parallel combination of CNN and LSTM based network, to obtain representative features termed ASV, that can maximally reflect sentiment information in an utterance from an audio. We extract several traditional heterogeneous acoustic features by Librosa toolkit and choose the four most representative features through a large number of experiments, and regard them as the input of the neural network. We can get CASV and LASV from the CNN branch and the LSTM branch respectively, and finally merge the two branches to obtain the final ASV for sentiment classification of each utterance. Besides, BiLSTM with attention mechanism is used for feature fusion. The experiment results show our model can recognize audio sentiment precisely and quickly, and demonstrate our heterogeneous ASV are better than traditional acoustic features or vectors extracted from other deep learning models. Furthermore, experiment results indicate that the proposed model outperforms the state-of-the-art approach by 9.33% on MOSI dataset. We have also tested our model on MOUD to prove the model won't heavily depend on language types. In the future, we will combine the feature engineering technologies to further discuss the fusion dimension of audio features and consider the fusion of different dimensions of different categories of features, and even apply them to multimodal sentiment analysis.

6. Acknowledgement

This paper has been funded by the project (Grant No: 61673235) supported by National Natural Science Foundation of China.

References

References

[1] B. Pang, L. Lee, et al., Opinion mining and sentiment analysis, Foundations and Trends® in Information Retrieval 2 (1–2) (2008) 1–135.
[2] B. Liu, Sentiment analysis: Mining opinions, sentiments, and emotions, Cambridge University Press, 2015.
[3] J. Tao, F.-L. Chung, S. Wang, On minimum distribution discrepancy support vector machine for domain adaptation, Pattern Recognition 45 (11) (2012) 3962–3984.
[4] P. de Souza, Texture recognition via autoregression, Pattern Recognition 15 (6) (1982) 471–475.
[5] N. Passalis, A. Tefas, Neural bag-of-features learning, Pattern Recognition 64 (2017) 277–294.
[6] A. Reihanian, M.-R. Feizi-Derakhshi, H. S. Aghdasi, Overlapping community detection in rating-based social networks through analyzing topics, ratings and links, Pattern Recognition 81 (2018) 370–387.
[7] G. Lin, K. Liao, B. Sun, Y. Chen, F. Zhao, Dynamic graph fusion label propagation for semi-supervised multi-modality classification, Pattern Recognition 68 (2017) 14–23.
[8] S. Poria, E. Cambria, R. Bajpai, A. Hussain, A review of affective computing: From unimodal analysis to multimodal fusion, Information Fusion 37 (2017) 98–125.
[9] M. W. Kraus, Voice-only communication enhances empathic accuracy., American Psychologist 72 (7) (2017) 644.
[10] S. Zhang, S. Zhang, T. Huang, W. Gao, Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching, IEEE Transactions on Multimedia 20 (6) (2018) 1576–1590.
[11] S. Ezzat, N. El Gayar, M. M. Ghanem, Sentiment analysis of call centre audio conversations using text classification, International Journal of Computer Information Systems and Industrial Management Applications 4 (1) (2012) 619–627.
[12] A. Van Den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. W. Senior, K. Kavukcuoglu, Wavenet: A generative model for raw audio., in: SSW, 2016, p. 125.

[13] T. Bertin-Mahieux, D. P. Ellis, Large-scale cover song recognition using hashed chroma landmarks, in: Applications of Signal Processing to Audio and Acoustics (WASPAA), 2011 IEEE Workshop on, IEEE, 2011, pp. 117–120.
[14] L. Kaushik, A. Sangwan, J. H. Hansen, Automatic audio sentiment extraction using keyword spotting, in: Sixteenth Annual Conference of the International Speech Communication Association, 2015.
[15] K. Hong, G. Liu, W. Chen, S. Hong, Classification of the emotional stress and physical stress using signal magnification and canonical correlation analysis, Pattern Recognition 77 (2018) 140–149.
[16] F. Mandanas, C. Kotropoulos, M-estimators for robust multidimensional scaling employing 2, 1 norm regularization, Pattern Recognition 73 (2018) 235–246.
[17] W. C. F. Mariel, S. Mariyah, S. Pramana, Sentiment analysis: a comparison of deep learning neural network algorithm with svm and naïve bayes for indonesian text, in: Journal of Physics: Conference Series, Vol. 971, IOP Publishing, 2018, p. 012049.
[18] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, S. Zafeiriou, Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network, in: Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on, IEEE, 2016, pp. 5200–5204.
[19] S. Mirsamadi, E. Barsoum, C. Zhang, Automatic speech emotion recognition using recurrent neural networks with local attention, in: Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on, IEEE, 2017, pp. 2227–2231.
[20] M. Neumann, N. T. Vu, Attentive convolutional neural network based speech emotion recognition: A study on the impact of input features, signal length, and acted speech, arXiv preprint arXiv:1706.00612.
[21] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, S. S. Narayanan, Iemocap: Interactive emotional dyadic motion capture database, Language resources and evaluation 42 (4) (2008) 335.
[22] Z.-Q. Wang, I. Tashev, Learning utterance-level representations for speech emotion and age/gender recognition using deep neural networks, in: Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on, IEEE, 2017, pp. 5150–5154.
[23] M. Chen, X. He, J. Yang, H. Zhang, 3-d convolutional recurrent neural networks with attention model for speech emotion recognition, IEEE Signal Processing Letters 25 (10) (2018) 1440–1444.
[24] J. Lee, D.-W. Kim, Scls: Multi-label feature selection based on scalable criterion for large label set, Pattern Recognition 66 (2017) 342–352.
[25] K. Kim, J. Lee, Sentiment visualization and classification via semi-supervised nonlinear dimensionality reduction, Pattern Recognition 47 (2) (2014) 758–768.
[26] S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh, L.-P. Morency, Context-dependent sentiment analysis in user-generated videos, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Vol. 1, 2017, pp. 873–883.
[27] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural computation 9 (8) (1997) 1735–1780.
[28] Y. LeCun, B. E. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. E. Hubbard, L. D. Jackel, Handwritten digit recognition with a back-propagation network, in: Advances in neural information processing systems, 1990, pp. 396–404.
[29] T. N. Sainath, B. Kingsbury, G. Saon, H. Soltau, A.-r. Mohamed, G. Dahl, B. Ramabhadran, Deep convolutional neural networks for large-scale speech tasks, Neural Networks 64 (2015) 39–48.
[30] D. Palaz, R. Collobert, et al., Analysis of cnn-based speech recognition system using raw speech as input, in: Proceedings of INTERSPEECH, no. EPFL-CONF-210029, 2015.
[31] M. El Ayadi, M. S. Kamel, F. Karray, Survey on speech emotion recognition: Features, classification schemes, and databases, Pattern Recognition 44 (3) (2011) 572–587.
[32] N. Cummins, S. Amiriparian, G. Hagerer, A. Batliner, S. Steidl, B. W. Schuller, An image-based deep spectrum feature representation for the recognition of emotional speech, in: Proceedings of the 2017 ACM on Multimedia Conference, ACM, 2017, pp. 478–484.
[33] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, O. Nieto, librosa: Audio and music signal analysis in python, in: Proceedings of the 14th python in science conference, 2015, pp. 18–25.
[34] S. H. Bae, I. Choi, N. S. Kim, Acoustic scene classification using parallel combination of lstm and cnn, in: Proceedings of the Detection and Classification of Acoustic Scenes and Events 2016 Workshop (DCASE2016), 2016, pp. 11–15.
[35] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, T. Darrell, Long-term recurrent convolutional networks for visual recognition and description, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 2625–2634.
[36] T. N. Sainath, O. Vinyals, A. Senior, H. Sak, Convolutional, long short-term memory, fully connected deep neural networks, in: Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on, IEEE, 2015, pp. 4580–4584.
[37] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, arXiv preprint arXiv:1409.0473.
[38] A. Zadeh, R. Zellers, E. Pincus, L.-P. Morency, Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos, arXiv preprint arXiv:1606.06259.
[39] V. Pérez-Rosas, R. Mihalcea, L.-P. Morency, Utterance-level multimodal sentiment analysis, in: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Vol. 1, 2013, pp. 973–982.
[40] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proceedings of the IEEE 86 (11) (1998) 2278–2324.
[41] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in neural information processing systems, 2012, pp. 1097–1105.

[42] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556.
[43] M. D. Zeiler, R. Fergus, Visualizing and understanding convolutional networks, in: European conference on computer vision, Springer, 2014, pp. 818–833.
[44] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.
[45] S. Maghilnan, M. R. Kumar, Sentiment analysis on speaker specific speech data, in: Intelligent Computing and Control (I2C2), 2017 International Conference on, IEEE, 2017, pp. 1–5.
[46] C. Bakir, D. S. Jarvis, Institutional entrepreneurship and policy change, Policy and Society.