3-D Convolutional Recurrent Neural Networks with Attention Model for Speech Emotion Recognition

Mingyi Chen, Xuanji He, Jing Yang, and Han Zhang

***Abstract*—Speech emotion recognition (SER) is a difficult task due to the complexity of emotion. The SER performances are heavily depend on the effectiveness of emotional features extracted from speech. However, most emotional features are sensitive to emotional irrelevant factors, such as the speaker, speaking styles and environment. In this letter, we assume that calculating the deltas and delta-deltas for personalized features not only preserves effective emotional information but also reduce the influence of emotional irrelevant factors, leading to reduce misclassification. In addition, SER often suffers from the silent frames and emotional irrelevant frames. Meanwhile, attention mechanism has exhibited outstanding performances in learning relevant feature representations for specific tasks. Inspired by this, we propose a 3-D attention-based convolutional recurrent neural networks (ACRNN) to learn discriminative features for SER, where the Mel-spectrogram with deltas and delta-deltas are used as input. Experiments on IEMOCAP and Emo-DB corpus demonstrate the effectiveness of the proposed method and achieve the state-of-the-art performance in terms of unweighted average recall.**

***Index Terms*—speech emotion recognition, convolutional recurrent neural networks, attention mechanism.**

I. Introduction

AS an important part of human intelligence, emotional intelligence is considered to be an essential or even the most critical for a person to be succeed. In order to enable the human-machine interfaces(HMI) more harmonious, it is bound to have emotional intelligence. Speech as the most convenient and natural medium for human communication, not only carries the implicit semantic information, but also contains rich affective information [1]. Therefore, speech emotion recognition (SER), which aims to recognize the correct emotional state of the speaker from speech signals, has drawn a great deal of attention of researchers.

In recent years, deep neural networks (DNN) have exhibited outstanding performances in extracting discriminative features for SER. Compared with hand-crafted features, DNN is capable of extracting hierarchical feature representations for a specific task from a large amount of training samples by supervised learning. Schmidt *et al.* [2] employed a deep belief network (DBN) to extract high-level emotional feature representations from the magnitude spectra and showed better performance compared to the traditional acoustic features. Han *et al.* [3] proposed to use the segments with highest energy to train a DNN model to extract effective emotional information. Mao *et al.* [4] first used convolutional neural networks (CNN) to learn affective-salient features for SER and showed excellent performances on several benchmark datasets. Lee *et al.* [5] applied a recurrent neural network (RNN) to learn long-range temporal relationships for SER. More recently, Trigeorgis *et al.* [6] directly used the raw audio samples to train a convolutional recurrent neural network (CRNN) to predict continuous arousal /valence space.

While DNN has achieved great success in SER, researchers are somehow ignoring the issue that DNN still use personalized features as input, since personalized features can be affected by the various styles of speaking, the content of speech and the environment. In general, the speech emotion features that directly reflect numerical values are called personalized features, which carry a lot of personal emotional information, reflect the characteristics of the speaker and can't contain common emotional information that is invariant to the variations of speakers, contents and environment [4]. At present, most studies of SER are based on personalized emotional features and have achieved good recognition performances, especially for specific speakers. Although the improved SER algorithm can be used to obtain better performances, it hinders the practical application of the SER technology in the real context of the speaker-independent. Therefore, it is significant to reduce the numerical difference of personalized features for different speakers and speaking styles.

Inspired by the positive result of MFCC with deltas and delta-deltas in the field of SER [7], we assume that calculate deltas and delta-deltas for personalized features are capable of reflecting the changing process of emotion and retaining effective emotional information while reducing the influence of emotional irrelevant factors, such as the speakers, contents and environment. In this letter, we calculate deltas and delta-deltas for the log Mel-spectrogram (log-Mels) as the convolutional recurrent networks input. Compared with 2-D convolution, 3-D convolution can better capture more effective information for SER, which helps to reduce misclassification.

In addition, attention mechanism based RNN has achieved great success in learning correlated structures between the input and the output sequences [8]. Attention mechanism was first proposed in [9] for machine translation and later applied for document classification [10] and speech recognition [11].

M. Chen is with the School of Information Science and Engineering, Central South University, Changsha, Hunan Province 410083, China (e-mail: MYCCSU@163.com).

X. He is with the School of Information Science and Engineering, Central South University, Changsha, Hunan Province 410083, China (e-mail: xuanjihe@163.com).

J. Yang and H. Zhang are with the School of Information Science and Engineering, Central South University, Changsha, Hunan Province 410083, China (e-mail: mllxy54@163.com; zhanghan131@gmail.com).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/LSP.2018.2860246, IEEE Signal Processing Letters

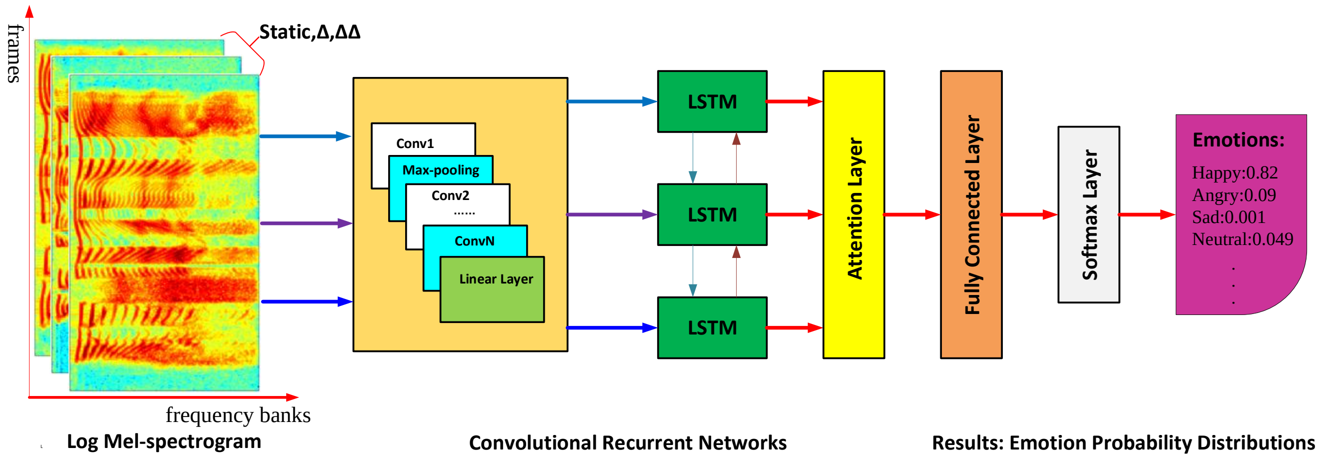


Fig. 1. Illustration of the proposed 3-D attention-based convolutional recurrent network (ACRNN) architecture to generate discriminative features for speech emotion recognition. This architecture consists of six phases: (1)Log-Mels (static, deltas and delta-deltas) are extracted from the speech signal as the ACRNN input. (2) 3-D CNN is employed for local invariant features extraction with log-Mels. (3) Bidirectional recurrent neural networks (BLSTM) are used to learn temporal dependencies between different time-step local invariant features. (4) An attention layer is employed to generate utterance-level features by discovering emotional relevant parts of the CRNN features. (5) Utterance-level features are fed into a fully connected layer to obtain higher-level feature representations for better classification. (6) The higher-level features are feed into a softmax layer for the final classification.

Attention mechanism based RNN fits the SER tasks well. First, speech is essentially the sequential data with a varied length. Second, in most speech emotion corpus, the emotion labels are annotated at the utterance-level, while the utterance often contains many silence periods and in many cases emotions are associated with only a few words. Thus, it is important to select emotion relevant frames for SER. In the field of SER, several works [12]-[15] have studied the effectiveness of attention mechanism in discovering emotion relevant regions for utterance-level emotion prediction and obtained promising performances in several datasets. In this letter, we extend our model with an attention mechanism based CRNN to produce affective-salient features for the final emotion classification.

In this letter, we propose a novel architecture, called 3-D attention-based convolutional recurrent neural networks (ACRNN) for SER by combining CRNN with an attention model. The major contributions of this letter are summarized as:

1) By calculating deltas and delta-deltas for the log-Mels, the numerical differences for emotion irrelevant factors, such as the style of speaking, speech contents and the environment, can be effectively reduced. We propose a novel 3-D CRNN for SER, which helps to better capture the time-frequency relationship of the log-Mels and leads to a stable SER performances.
2) To deal with silent frames and emotion irrelevant frames, we employ an additional attention model to automatically focus on the emotion relevant frames and produce discriminative utterance-level features for SER.
3) Experimental results indicate that the proposed method outperforms the baseline methods by 13.5% and 11.26% for IEMOCAP and Emo-DB, respectively.

II. Proposed Methodology

In this section, we introduce our ACRNN method for SER. First, we generate the log-Mels (static, deltas and delta-deltas) from speech signals as ACRNN input. Then, we introduce the architecture of ACRNN, which integrates CRNN with an attention model, followed by a fully connected layer and a softmax layer for SER, as shown in Fig.1.

A. 3-D Log-Mels generation

In recent years, CNN has exhibited excellent performances in the field of SER [1], [4], [6]. Chan *et al.* [16] found that 2-dimensional convolution outperforms 1-dimensional convolution with limited data, and the time domain convolution is as important as the frequency domain convolution. However, the above CNN models use personalized features as input, and the SER performances vary greatly among different speakers and various speaking styles. To address this issue, in this letter, we use the log-Mels with deltas and delta-deltas as the ACRNN input, where the deltas and delta-deltas reflect the process of emotional change.

Given a speech signal, zero mean and unit variance are done for reducing the variations between different speakers, and split the signal into short frames with Hamming windows of 25 *ms* and a 10 *ms* shift. Then, we calculate the power spectrum for each frame by using Discrete Fourier Transform (DFT). By passing the power spectrum through the Mel-filter bank i to produce output p_i. Then, as shown by (1), the log-Mels m_i is produced by taking the logarithm of p_i. To calculate the deltas features m_i^d of the log-Mels, we use the formula is given by (2). A popular choice for N is 2. Similarly, the delta-deltas features m_i^{dd} are calculated by taking the time derivative of the deltas, as shown in (3). After computing the log-Mels with deltas and delta-deltas, we can obtain a 3-D feature representation $X \in R^{t \times f \times c}$ as the CNN input, where t denotes the time (frame) length and f denotes the number of Mel-filter bank, and c denotes the number of feature channels. In this task, we set f as 40, the same as speech recognition [17], and set the value of c as *3*, representing the static, deltas, and delta-deltas respectively.

$$m_i = \log(p_i) \tag{1}$$

$$m_i^d = \frac{\sum_{n=1}^{N} n(m_{i+n} - m_{i-n})}{2\sum_{n=1}^{N} n^2} \tag{2}$$

$$m_i^{dd} = \frac{\sum_{n=1}^{N} n(m_{i+n}^d - m_{i-n}^d)}{2\sum_{n=1}^{N} n^2} \tag{3}$$

B. Architecture of ACRNN

In this section, we combine CRNN with an attention model to analyze 3-D log-Mels for SER. First, we perform 3-D CNN on the whole log-Mels, different from [12], [17], which convolution in a patch that only contains several frames. Next, 3-D CNN sequential features are fed into LSTM for temporal summarization. Then, the attention layer takes a sequence of high-level features as input to produce utterance-level features. Finally, utterance-level features are used as the fully connected layer input to obtain higher-level features for SER.

1) CRNN Model: Given a 3-D log-Mels, CRNN is used to extract high-level features for SER. In this letter, CRNN consists of several 3-D convolution layers, one 3-D max-pooling layers, one linear layer and one LSTM layer. Specifically, the first convolutional layer has 128 feature maps, while the remaining convolutional layers have 256 feature maps, and the filter size of each convolutional layer is 5×3, where 5 corresponds to the time axis, and 3 corresponds to the frequency axis. We only perform max-pooling after the first convolutional layer and the pooling size is 2×2. According to [18], the model parameters can be reduced effectively with no loss in accuracy by adding a linear layer before passing 3-D CNN features into the LSTM layer. Thus we add a linear layer after 3-D CNN for dimension reduction and we find that the linear layer with 768 output units is appropriate. After 3-D CNN is performed, we feed the 3-D CNN sequence features into a bi-directional recurrent neural network with long short-term memory cells (BLSTM) [19] for temporal summarization, and each direction contains 128 cells, then we can obtain a sequence of 256-dimensional high-level feature representations.

2) Attention Layer : With a sequence of high-level representations, an attention layer is employed to focus on emotion relevant parts and produce discriminative utterance-level representations for SER, since not all frame-level CRNN features contribute equally to the representation of the speech emotion. In this letter, we use an attention model to score the importance of a sequence of high-level representations to the final utterance-level emotion representations, instead of simply performing a mean/max pooling over time.

Specifically, as shown in Fig.2, with the LSTM output $h_t = [\overrightarrow{h_t}, \overleftarrow{h_t}]$ at time step *t*, we first compute the normalized importance weight α_t by a softmax function as (4). Then we calculate the utterance-level representations *c* by performing a weighted sum on h_t according to the weights, as shown in (5).

Finally, we pass the utterance-level representations into a fully connected layer with 64 output units to obtain higher-level representations that help the softmax classifier to better map the utterance representations into *N* different spaces, where *N* denotes the number of emotion classes. Batch normalization (BN) [20] is applied to the fully connected layer to accelerate training and improve the generalization performance.

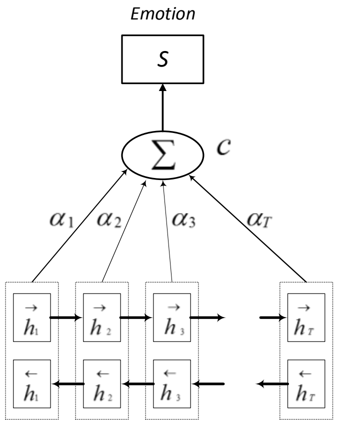


Fig. 2. The working process of the attention layer.

$$\alpha_t = \frac{\exp(W \cdot h_t)}{\sum_{\tau=1}^{T} \exp(W \cdot h_t)} \tag{4}$$

$$c = \sum_{t=1}^{T} \alpha_t h_t \tag{5}$$

III. Experiments

A. Experiment Setup

To evaluate the performance of our proposed model, we perform speaker-independent SER experiments on the Interactive Emotional Dyadic Motion Capture database (IEMOCAP) [21] and the Berlin Emotional Database (Emo-DB) [22]. IEMOCAP consists of 5 sessions and each session is displayed by a pair of speakers (female and male) in scripted and improvised scenarios, contains a total of 10039 utterances, with an average duration of 4.5 seconds, and the sample rate is 16 kHz. Emo-DB consists of 535 utterances that displayed by ten professional actors, covering seven emotions (*neutral, fear, joy, angry, sadness, disgust and boredom*), and this database sampled at 44.1 kHz, and later downsample to 16 kHz.

In this letter, for IEMOCAP, we only consider the improvised data with four emotional categories: *happy, angry, sad and neutral*, and use all seven emotions for Emo-DB. Since both databases contain 10 speakers, we employ a 10-fold cross-validation technique in our evaluations. Specifically, for each evaluation, 8 speakers are selected as the training data and 1 speaker is select as the validation data, while the remaining 1 speaker is used as the test data. With different parameter initializations, we can obtain wide range of results. Thus, we repeat each evaluation for five times with different random seeds and report the average and standard deviation to get more reliable results. Since the test class distributions are imbalanced, we report unweighted average recall (UAR) on the test set. Note that, all model architectures, including the number of epochs are selected by maximizing the UAR on the validation set.

For better parallel acceleration, we split the speech signal into equal-length segments of 3 seconds, and zero-padding is applied for the utterances whose duration less than 3 seconds. In the training phase, each sub-segment is used to predict one emotion, while in the test phase, we evaluate the whole sentences prediction by applying max pooling on the posterior probability of each sub-sentence. The log-Mels are extracted by the *openEAR* toolkit [23] with the window size of 25 *ms* and a 10 *ms* shift, and both the training and test log-Mels are

TABLE I
SER Average and Standard Deviation of Different ACRNN Architectures on IEMOCAP and Emo-DB in Terms of UAR

Network	IEMOCAP	Emo-DB
1 convolutional layer and LSTM	64.20±5.73	81.53±6.96
2 convolutional layers and LSTM	64.18±5.57	81.72±6.32
3 convolutional layers and LSTM	64.11±5.23	82.74±5.03
4 convolutional layers and LSTM	64.62±5.27	81.75±6.22
5 convolutional layers and LSTM	63.80±5.45	**82.82±4.99**
6 convolutional layers and LSTM	**64.74±5.44**	82.69±5.53
7 convolutional layers and LSTM	63.95±5.12	80.84±5.09

normalized by the global mean and the standard deviation of the training set.

The ACRNN architecture is implemented with TensorFlow toolkit [24], and the parameters of the model were optimized by minimizing the cross-entropy objective function, with a mini-batch of 40 samples, using the Adam optimizer [25] with Nestorov momentum. The initial learning rate is set to 10^{-4} and the momentum is set to *0.9*.

B. Baselines

We compare our approach with several baselines:

1） DNN-ELM [3]. According to [3], DNN-ELM consists of three hidden layers with 256 hidden units. Utterance-level features are obtained by applying statistical functions on the segment-level probabilities and further feed into an Extreme Learning Machine (ELM) for the final decision.

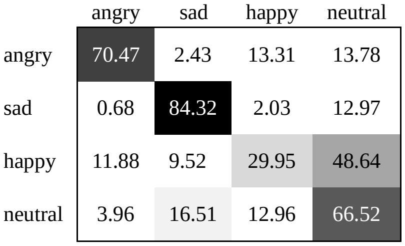
2） 2-D ACRNN. Different from our proposed 3-D ACRNN, we train the ACRNN on the log-Mels with the same setting and call the “2-D ACRNN”.

C. Comparison of Network Architectures

In this section, we use ARCNN to execute SER on the Log-Mels with increasing convolutional layer. As shown in Table I, we observe that ACRNN achieves the best performance on IEMOCAP when it contains six convolutional layers and achieves the best performance when it contains five convolutional layers on Emo-DB. The result reveal that the best SER architecture heavily depends on the type and size of the training data, this finding is of great significance for the development of SER systems on new datasets. Besides, we have evaluated ACRNN with more LSTM layers, but unfortunately, we don’t get any improvement in UAR.

D. Experiment Results

Table II shows the comparison of our proposed method with baselines in terms of UAR. First, we compare our method with the state-of-the-art DNN-ELM method described in [3]. The result shows that ACRNN outperforms DNN-ELM and obtains an absolute improvement of 13.5% and 11.26% for IEMOCAP and Emo-DB, respectively. Next, we investigate the effectiveness of 3-D CNN. Compared with 2-D ACRNN, 3-D ACRNN obtains an absolute improvement of 2.34% and 3.44% for IEMOCAP and Emo-DB, respectively. In addition, the standard deviation of the 3-D ACRNN is smaller than 2-D ACRNN. This indicates that calculating deltas and delta-deltas for personalized features (e.g., the Log-Mels, MFCC) can retain effective emotional information while reducing the influence of the speakers, speaking styles and other emotional irrelevant factors.



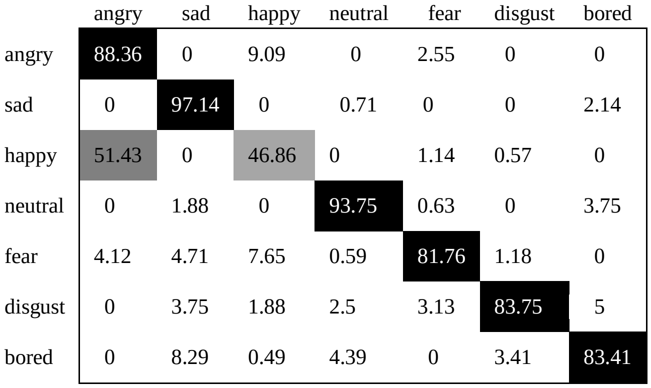
	angry	sad	happy	neutral
angry	70.47	2.43	13.31	13.78
sad	0.68	84.32	2.03	12.97
happy	11.88	9.52	29.95	48.64
neutral	3.96	16.51	12.96	66.52

Fig. 3. Confusion matrix of 3-D ACRNN with an average recall of 64.74% on the IEMOCAP dataset, where each row presents the confusion of the ground truth emotion during prediction.

TABLE II
SER Average and Standard Deviation for Different Methods on IEMOCAP and Emo-DB in Terms of UAR

Method	IEMOCAP	Emo-DB
DNN-ELM[3]	51.24±7.24	71.56±8.43
2-D ACRNN	62.40±6.70	79.38±7.78
3-D ACRNN	**64.74±5.44**	**82.82±4.99**

Finally, we present the confusion matrix to further analyze the SER performances of 3-D ACRNN. According to Fig. 3 and 4, we observe that on both IEMOCAP and Emo-DB datasets, *sad* obtains the highest recognition rate, and *happy* obtains the lowest recognition rate. In addition, we find that on the IEMOCAP, 48.64% *happy* samples are misclassified as *neutral*, while on the Emo-DB, 51.43% *happy* samples are misclassified as *angry*. We attribute these mistakes to the similar activation level of *angry* and *happy*, while *neutral* is at the center of the activation-valence space.



	angry	sad	happy	neutral	fear	disgust	bored
angry	88.36	0	9.09	0	2.55	0	0
sad	0	97.14	0	0.71	0	0	2.14
happy	51.43	0	46.86	0	1.14	0.57	0
neutral	0	1.88	0	93.75	0.63	0	3.75
fear	4.12	4.71	7.65	0.59	81.76	1.18	0
disgust	0	3.75	1.88	2.5	3.13	83.75	5
bored	0	8.29	0.49	4.39	0	3.41	83.41

Fig. 4. Confusion matrix of 3-D ACRNN with an average recall of 82.82% on the Emo-DB dataset, where each row presents the confusion of the ground truth emotion during prediction.

IV. Conclusion

In this letter, we proposed a 3-D attention-based convolutional recurrent neural networks (ACRNN) for SER. We first extract log-Mels (static, deltas and delta-deltas) from speech signals as the 3-D CNN input. Next, we combine 3-D CNN with LSTM for high-level features extraction. Finally, an attention layer is used to focus on the emotional relevant parts and produce utterance-level affective-salient features for SER. Experiments on the IEMOCAP and Emo-DB databases show the superiority of our proposed approach compared with the baselines in terms of UAR.

REFERENCES

[1] S. Zhang, S. Zhang, T. Huang and W. Gao, "Speech Emotion Recognition Using Deep Convolutional Neural Network and Discriminant Temporal Pyramid Matching," in *IEEE Transactions on Multimedia*, vol. PP. 99 (2017):1-1.

[2] E. M. Schmidt and Y. E. Kim, “Learning emotion-based acoustic features with deep belief networks,” in *Proc. IEEEWorkshop Appl. Signal Process. Audio Acoust.*, New Paltz, NY, USA, 2011, pp. 65–68.

[3] K. Han, D. Yu, and I. Tashev, “Speech Emotion Recognition using Deep Neural Network and Extreme Learning Machine,” in *Proceedings of Interspeech*, 2014.

[4] Q. Mao, M. Dong, Z. Huang, and Y. Zhan, “Learning salient features for speech emotion recognition using convolutional neural networks,” *IEEE Transactions on Multimedia*, vol. 16, no. 8, pp. 2203–2213, Dec 2014.

[5] Jinkyu Lee and Ivan Tashev, “High-level feature representation using recurrent neural network for speech emotion recognition,” in *Interspeech*, 2015.

[6] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, “ADIEU Features? End-to-end Speech Emotion Recognition using A Deep Convolutional Recurrent Network,” in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2016, pp. 5200–5204.

[7] Schuller, Björn, et al. "Acoustic emotion recognition: A benchmark comparison of performances." *Automatic Speech Recognition & Understanding, 2009. ASRU 2009. IEEE Workshop on* IEEE, 2010: 552-557.

[8] Huang, Che Wei, and S. S. Narayanan. "Attention Assisted Discovery of Sub-Utterance Structure in Speech Emotion Recognition." *INTERSPEECH* 2016:1387-1391.

[9] Bahdanau, Dzmitry, K. Cho, and Y. Bengio. "Neural Machine Translation by Jointly Learning to Align and Translate." *Computer Science* (2014).

[10] Yang, Zichao, et al. "Hierarchical Attention Networks for Document Classification." *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* 2017:1480-1489.

[11] Chorowski, Jan, et al. "Attention-Based Models for Speech Recognition." *Computer Science* 10.4(2015):429-439.

[12] C. W. Huang and S. S. Narayanan, "Deep convolutional recurrent neural network with attention mechanism for robust speech emotion recognition," *2017 IEEE International Conference on Multimedia and Expo* (*ICME*), Hong Kong, 2017, pp. 583-588.

[13] Huang, Che Wei, and S. S. Narayanan. "Attention Assisted Discovery of Sub-Utterance Structure in Speech Emotion Recognition." *INTER-SPEECH* 2016:1387-1391.

[14] S. Mirsamadi, E. Barsoum, and C. Zhang, “Automatic speech emotion recognition using recurrent neural networks with local attention,” in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), New Orleans, U.S.A., Mar. 2017, IEEE, pp. 2227–2231.

[15] Neumann, Michael, and N. Thang Vu. "Attentive Convolutional Neural Network based Speech Emotion Recognition: A Study on the Impact of Input Features, Signal Length, and Acted Speech." *INTERSPEECH* 2017:1263-1267.

[16] William Chan and Ian Lane, “Deep convolutional neural networks for acoustic modeling in low resource languages,” in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2015, pp. 2056–2060.

[17] Keren, Gil, and B. Schuller. "Convolutional RNN: An enhanced model for extracting features from sequential data." *International Joint Conference on Neural Networks IEEE*, 2016:3412-3419.

[18] T. N. Sainath, V. Peddinti, B. Kingsbury, P. Fousek, D. Nahamoo, and B. Ramabhadhran, “Deep Scattering Spectra with Deep Neural Networks for LVCSR Tasks,” in *Proc. Interspeech*, 2014.

[19] Graves, Alex, et al. "Bidirectional LSTM Networks for Improved Phoneme Classification and Recognition." *Artificial Neural Networks: Formal MODELS and Their Applications - ICANN 2005, International Conference, Warsaw, Poland, September 11-15, 2005, Proceedings*DBLP, 2005:799-804.

[20] Vuckovic, Frano, G. Lauc, and Y. Aulchenko. "Normalization and batch correction methods for high-throughput glycomics." *Joint Meeting of the Society-For-Glycobiology* 2016:1160-1161.

[21] Busso, Carlos, et al. "IEMOCAP: interactive emotional dyadic motion capture database." *Language Resources & Evaluation* 42.4(2008):335.

[22] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss,“A database of German emotional speech.” in *Interspeech*, vol. 5, Lisbon, Portugal, 2005, pp. 1517–1520.

[23] F. Eyben, M. Wöllmer, and B. Schuller, “Open EAR—Introducing the Munich Open-source Emotion and Affect Recognition Toolkit,” in *Affective Computing and Intelligent Interaction and Workshops*, 2009, pp. 1–6.

[24] Abadi, Martın, et al. "TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems." (2016).

[25] D. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in *Proc. of International Conference on Learning Representations (ICLR)*, San Diego, CA, 2015.