2018

Exploring Spatio-Temporal Representations by Integrating Attention-based Bidirectional-LSTM-RNNs and FCNs for Speech Emotion Recognition

Ziping Zhao

Yu Zheng

Zixing Zhang

Haishuai Wang
Fairfield University, hwang@fairfield.edu

Yiqin Zhao

Published Citation

Ziping Zhao, Yu Zheng, Zixing Zhang, Haishuai Wang, Yiqin Zhao and Chao Li. Exploring Spatio-Temporal Representations by Integrating Attention-based Bidirectional-LSTM-RNNs and FCNs for Speech Emotion Recognition. Interspeech, 272-276, 2018. DOI: 10.21437/Interspeech.2018-1477

Interspeech 2018
2-6 September 2018, Hyderabad

Exploring Spatio-Temporal Representations by Integrating Attention-based Bidirectional-LSTM-RNNs and FCNs for Speech Emotion Recognition

Ziping Zhao[1], *Yu Zheng*[1], *Zixing Zhang*[2], *Haishuai Wang*[3], *Yiqin Zhao*[1], *Chao Li*[1]

[1] College of Computer and Information Engineering, Tianjin Normal University
[2] Group on Language, Audio & Music, Imperial College London, UK
[3] Department of Biomedical Informatics, Harvard University

zhaoziping@tjnu.edu.cn,tjnuzhengyu@126.com,zixing.zhang@imperial.ac.uk,
haishuai_wang@hms.harvard.edu, hawkinszhao@outlook.com, superlee@tjnu.edu.cn

Abstract

Automatic emotion recognition from speech, which is an important and challenging task in the field of affective computing, heavily relies on the effectiveness of the speech features for classification. Previous approaches to emotion recognition have mostly focused on the extraction of carefully hand-crafted features. How to model spatio-temporal dynamics for speech emotion recognition effectively is still under active investigation. In this paper, we propose a method to tackle the problem of emotional relevant feature extraction from speech by leveraging Attention-based Bidirectional Long Short-Term Memory Recurrent Neural Networks with fully convolutional networks in order to automatically learn the best spatio-temporal representations of speech signals. The learned high-level features are then fed into a deep neural network (DNN) to predict the final emotion. The experimental results on the Chinese Natural Audio-Visual Emotion Database (CHEAVD) and the Interactive Emotional Dyadic Motion Capture (IEMOCAP) corpora show that our method provides more accurate predictions compared with other existing emotion recognition algorithms.

Index Terms: speech emotion recognition, bidirectional long short-term memory, fully convolutional networks, attention mechanism

1. Introduction

Speech emotion recognition, referring to the process of detecting the emotional state of a speaker, has become a very active research topic in the affective computing field and has had a wide range of applications recently. Automatic emotion recognition from speech largely depends on the effectiveness of the speech features used for classification. Due to the subtleties of human emotion, how to extract discriminative and affect-salient features from the speech signals is still one of the important research topics [1, 2].

Recently, there has been growing interest in employing deep learning to automatically discover emotionally relevant features from speech[3, 4, 5, 6, 7].

However, most of the previous methods consider only either spatial feature learning or temporal dependency construction[8]. In order to better recognize human emotions, the crucial spatial and temporal dependencies should be well modeled.

In the context of image processing, convolutional neural networks (CNNs) have been successfully used to learn salient features in a supervised setting [9, 10]. CNNs are exceptionally good at capturing high-level features in a spatial domain, which complements the attention model and enhances the spatial-temporal information, thereby compensating for each other. Recently, fully convolutional networks (FCNs) [11], as a variant of CNNs, which allow the input and output signals to have the same dimensions, have been shown to achieve state-of-the-art performance to solve time series sequences classification problem. However, FCNs are not good at learning temporal features. Therefore, LSTM-RNN, as a popular RNN architecture specialized in sequence learning, has been introduced to learn the temporal features from emotion speech sequences. LSTM-RNN has a built-in memory gate to retain long-term information. Given this context, LSTM is suitable for learning temporal features from input sequences.

Since deep learning with attention mechanism is very flexible in the decoding phase, it has been widely used in natural language processing and speech recognition. The attention mechanism maximizes the contribution of the relevant encoding context vectors and minimizes the influence of the irrelevant ones for the construction of the decoding context. In [12, 13], bidirectional LSTM is combined with a novel pooling strategy using an attention mechanism which enables the network to focus on the emotionally salient parts of a sentence.

Consequently, in this paper, we propose the augmentation of Attention-based Bidirectional LSTM-RNNs (Attention-BLSTM-RNNs) with FCNs to extract high-level spatial-temporal features for speech emotion recognition. A major advantage of the Attention-BLSTM-FCN model is that it does not need heavy data preprocessing or feature engineering. Meanwhile, we also investigate the usage of attention-based architectures to improve speech emotion recognition. The utilization of the attention mechanism allows the network to focus on the emotionally salient parts of a sentence. The experiment results show that the combination of BLSTM and FCN as well as the use of an attention mechanism can improve the results previously reported on the CHEAVD and IEMOCAP dataset.

The major contributions of this paper include: i) we present a novel method to involve both spatial and temporal features for speech emotion recognition by leveraging FCNs with attention-based BLSTM-RNNs, which is capable of automatically learning feature representations and modeling the temporal dependencies between their activation; ii) the proposed method can be easily adapted to enhance the existing state-of-the art methods to improve their performance. To the best of our knowledge, this is the first work in the literature that applies Attention-BLSTM-FCN model to the speech emotion recognition task.

10.21437/Interspeech.2018-1477

2. Related Work

There is a growing trend to combine CNN and RNN into one architecture and to train the entire model in an end-to-end fashion. For example, Sainath et al. proposed the Convolutional Long Short-Term Memory Deep Neural Networks (CLDNN) model, made up of a few convolutional layers, LSTM gated recurrent layers, and fully connected (FC) layers in the respective order. The CLDNNs model is trained on the log-Mel filter bank energies [14] and on the raw waveform speech signal [15] for speech recognition, and the results showed that both CLDNN models outperform CNN and LSTM alone or combined. Similarly, in [16] and [17] CLDNN-based speech emotion recognition experiments are conducted on log-Mels and spectrograms respectively. In [18], a network architecture of convolutional recurrent neural network (CRNN) is proposed for large vocabulary speech recognition by combining the CNN and LSTM-RNN. Experimental results show that the model can exceed state-of-the-art speech recognition performance. Recently, the authors used both convolutional and recurrent layers to improve the performance of speech emotion recognition task based on time domain speech signals or raw signal in [6, 19].

3. Integration of Attention-based BLSTM and FCNs

3.1. Long Short-Term Memory RNNs

LSTM is usually adopted as the basic unit in RNN because it is able to tackle the problem of vanishing and exploding gradients in RNN training [20]. LSTM-RNN is capable of learning long-term dynamic dependencies so the problem of vanishing or exploding gradients can be avoided during training. Since the operations of writing, reading, and resetting are performed by these gates respectively within a memory block, they allow the network to store and retrieve information over long periods of time. In this paper, we use BLSTM, which contains two sub-networks for the left and right sequence context, which are forward and backward pass respectively.

While LSTMs have the ability to learn temporal dependencies in sequences, they have difficulty in learning long-term dependencies in long sequences. With the help of the attention mechanism [21], the LSTM-RNN can learn these dependencies.

3.2. Attention Mechanism

Attention-based models have been successfully used in plenty of sequence-to-sequence learning tasks, e.g., speech recognition [22], part-of-speech tagging [23] and machine translation [21]. Basically, the attention mechanism is to select relevant encoded hidden vectors via attention weights (an informative sequence of weights) during the decoding phase. The architecture affords the possibility to construct an end-to-end system.

We calculate the attention weights α_i for each vector $\mathbf{x_i}$ in a sequence of inputs $\mathbf{x}$, as follows:

$$\alpha_i = \frac{exp(f(\mathbf{x}_i))}{\sum_j exp(f(\mathbf{x}_j))} \tag{1}$$

where $f(\mathbf{x})$ is the scoring function. Here, we use a linear function $f(\mathbf{x}) = W^T\mathbf{x}$ for $f(\mathbf{x})$. The W in the linear function is a trainable parameter. The output of the attention layer is the weighted sum of the input sequence, which is defined as $attentive_{\mathbf{x}}$:

$$attentive_{\mathbf{x}} = \sum_i \alpha_i \mathbf{x}_i \tag{2}$$

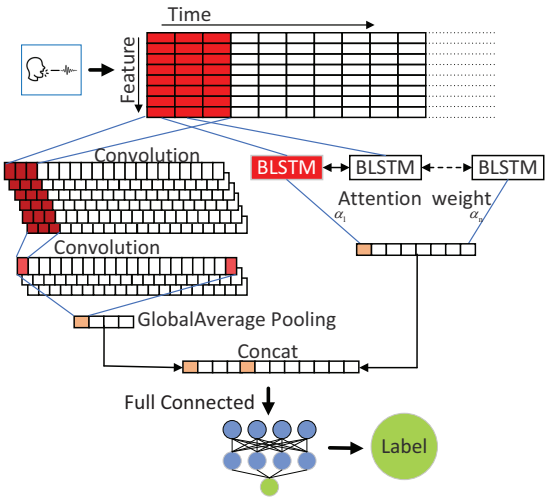

Figure 1: *The Attention-BLSTM-FCN architecture*

3.3. Temporal Convolutions

As a learning model, temporal convolutions have shown their efficiency in solving time series classification problems [24]. Moreover, temporal convolutions can extract spectral features from raw wave signals and capture long-term dependencies [6]. FCNs comprised of temporal convolutions are typically used for feature extraction, and global average pooling [25] can greatly reduce the number of parameters of the network.

Temporal Convolutional Networks are employed as a feature extraction module in a FCN branch in this paper. A basic convolution block comprises of a convolution layer, which is accompanied by batch normalization [26] then an activation function, which can be either a ReLU (rectified linear unit) or a PReLU(parametric rectified linear unit)[27, 28].

3.4. Attention BLSTM Fully Convolutional Network

In the presented architecture, the fully convolutional block is augmented by an Attention-BLSTM block followed by dropout [29], as shown in Fig. 1. The motivation to introduce this model is illustrated by two requirements of speech emotion recognition: firstly, the information in each sequence contains different proportions of emotion. Different speech streams can generate important time information through weighted speech sequences that determine the final mood of speech through contextual relationships. Secondly, speech emotion information is distributed in different feature parts of the sequence. Speech streams can be obtained through convolution of the spatial information that can represent emotions. In our proposed model, fully convolution blocks and BLSTM blocks perceive the same speech sequence input in two different perspectives. The BLSTM block in the proposed model uses the input speech sequence as a multiple time series with a single time step. In contrast, full convolution blocks treat a speech sequence as a univariate time series with multiple time steps. The fusion of these two kinds of information provides us with an emotional space that contains space-time information, and finally a DNN uses the emotional space to predict the emotion.

The fully convolutional block includes two stacked temporal convolutional blocks with the filter sizes of 128 and 64 respectively, which is different from [28]. Each convolutional block is the same with the convolution block in the CNN archi-

Table 1: *Final number of instances of the eight emotion classes in the CHEAVD Dataset*

	Train	Val	Test	Total
Neutral	1400	200	400	2000
Angry	884	128	252	1264
Happy	828	119	236	1183
Sad	462	67	132	661
Worried	567	81	162	810
Anxious	457	66	131	654
Surprise	175	25	51	251
Disgust	144	21	42	207
Sum	4917	707	1406	7030

tecture in [24]. Each block includes a temporal convolutional layer along with batch normalization [26], followed by a ReLU activation function. After that, global average pooling is used after the final convolution block. At the same time, the time series input is transformed through a dimension shuffle layer, and then passed into the Attention-BLSTM block. The Attention-BLSTM block consists of an BLSTM layer, with a cell size of 64 and an attention layer, with a dropout. The dropout rate is set to 0.5. Finally, we concatenate the output of the global pooling layer and the Attention-BLSTM block, and then fed the concatenated outputs into a DNN, which is trained to predict the final emotion.

4. Datasets and Feature Extraction

4.1. Datasets

To demonstrate the effectiveness of the model, the experiments are performed on the Interactive Emotional Dyadic Motion Capture (IEMOCAP) database [30] and the Chinese Natural Audio-Visual Emotion Database (CHEAVD) 2.0, which is utilized as the MEC2017 (Multimodal Emotion Recognition Challenge) dataset [31]. The IEMOCAP, which is a well-known corpus, is made up of audio-visual data with transcriptions from recordings of dialogues between two persons.

The CHEAVD 2.0 dataset is extracted from Chinese movies and TV programs, which contains 2852 spontaneous emotional segments (140 min). Discrete emotion (Angry(Ang), Disgust(Dis), Happiness(Hap), Sad(Sad), Surprise(Sup), Anxious(Anx), Worried(Wor) and Neutral(Neu)) labels are provided. To measure the emotion recognition performance, the samples in the CHEAVD dataset are split into three sets: the training set, the validation set and the testing set, which contain 4 917, 707 and 1 406 samples, respectively. Since the CHEAVD corpora is utilized as the MEC 2017 dataset, the labels of the test set will be unknown. So we only use the training set for training and choose the hyper-parameters based on the validation set to find the best emotion recognition model with the highest performance. The emotion class distribution of the CHEAVD dataset is shown in Table 1.

4.2. Feature Extraction

The YAAFE [32] toolbox is employed to extract acoustic features and we extract all the 27 features of the toolbox. The audio data is resampled to 16KHz and the default parameters of each feature are utilized. Finally, 743 dimension features are extracted for each frame and the length of each frame is 1024. The acoustic features are whitened with PCA [33].

5. Experiments and Results

5.1. Experiment Setup

In order to evaluate the performance, a leave-one-speaker-out strategy is applied in the experiments. Our models were implemented with the Tensorflow library. We used stochastic gradient descent with an adaptive learning rate. The batch size is set to 50.

The utterances in the IEMOCAP corpus are divided into five sessions, where in each session, a pair of actors (male-female) talk to each other. We use the 5-fold cross-validation technique for the five sessions to ensure the cross-validation is speaker-independent. For each evaluation, four sessions of the dataset are employed for training, and we divide the remaining session into two sub-sessions depending on gender. One actor is used for parameter validation, and the other for measurement. Four emotion categories are used: angry, happy, sad, and neutral. The numbers of utterances in each category are 1 103, 595, 1 084 and 1 708, respectively, making a total of 4 490, which is referred to as IEMO_dataset1. Then, similar to [34], we merge the happy and excited utterances into the happy class since they are close in emotion, making a total of 1636 utterances. The merged are referred to as IEMO_dataset2 in this paper.

We first evaluate the performance of the proposed Attention-BLSTM-FCN model on the CHEAVD database. Then two experiments are conducted on the IEMOCAP dataset. In the first experiment on the IEMOCAP dataset, we compare three pooling strategies, namely global mean-pooling, global max-pooling and attention-pooling. Then, we compare our proposed approach to the following baselines on IEMOCAP:

- The DNN-ELM approach in [4] where the authors train a DNN with an ELM.
- The Attention-BLSTM approach in [13] where the authors train the Attention-based BLSTM.
- The DBN-ivector approach in [34] where the authors train i-vector space and a deep belief network (DBN).
- The ACNN approach in [35] where the authors train an attentive convolutional neural network.

As evaluation measures, we employ two metrics in this paper: unweighted (UA, thus better reflecting imbalance among classes) and weighted (WA, i. e., accuracy) average recall which are standard measurements employed in several previous emotion challenges.

5.2. Results on the CHEAVD Dataset

A comparison of the results of the MEC2017 baseline, FCNs, Attention-LSTM, Attention-BLSTM, Attention-LSTM-FCN and our proposed model is shown in Table 2.

It can be seen that the proposed approach outperforms the others in terms of UA and WA and yields a UA of 31.8%, WA of 46.3% respectively under the cross validation condition, which is a relative increase of 5.1% for UA and 4.2% for WA over the other baseline models.

5.3. Results on the IEMOCAP Dataset

Firstly, a comparison of the results of the three pooling strategies on IEMO_dataset1 and IEMO_dataset2 is shown in Table 3 in terms of UA and WA. We can observe that attention-pooling consistently achieves the best performance on both IEMOCAP datasets.

Table 2: *Performance comparison between the proposed Attention-BLSTM-FCN With the other baselines on the validation set of CHEAVD dataset*

Approaches [%]	Validation UA	WA
MEC2017 Baseline[36]	27.2	39.9
FCNs	25.1	41.5
Attention-LSTM	25.2	41.7
Attention-BLSTM	29.4	45.3
Attention-BLSTM-FCN	**31.8**	**46.3**

Table 3: *Performance comparison between the three pooling strategies on the IEMOCAP dataset*

Strategies [%]	IEMO_dataset1 UA	WA	IEMO_dataset2 UA	WA
Global max-pooling	54.1	62.8	58.8	56.8
Global average-pooling	53.7	62.5	58.4	57.4
Attention-pooling	**56.0**	**64.0**	**60.1**	**59.7**

Table 4: *Performance comparison between the proposed model and the other baselines on the IEMOCAP dataset for recognizing four emotional classes*

Dataset	Approaches [%]	UA	WA
IEMO_dataset1	DNN-ELM [4]	52.1	57.9
	Attention-BLSTM [13]	49.9	59.3
	Attention-BLSTM-FCN	**56.0**	**64.0**
IEMO_dataset2	Attentive CNN [35]	55.1	56.1
	DBN-ivector [34]	59.6	58.1
	Attention-BLSTM-FCN	**60.1**	**59.7**

Then a comparison of the results between our model and the other existing methods on the IEMOCAP dataset is summarized in Table 4 in terms of UA and WA.

Overall, it can be seen from Table 4 that the proposed approach outperforms the other existing methods for UA and WA on the IEMOCAP dataset, which yields the best performance of 56.0%(UA), 64.0%(WA) on IEMO_dataset1 and 60.1%(UA), 59.7%(WA) on IEMO_dataset2. When comparing the IEMO_dataset1 to the IEMO_dataset2, the improvement for UA and WA is largely noticeable in the IEMO_dataset1. In this study, we concentrate more on the weighted accuracy, as it represents the classification accuracy over the entire test set and it is self-weighted from the usability standpoint.[37]

5.4. Discussions

From an overall experimental point of view, our approach consistently has a notable performance improvement over the other existing methods on the CHEAVAD and IEMOCAP dataset in terms of UA and WA. Overall, the proposed approach outperforms the other baseline models with a 5.1%, 4.2% relative improvement on the CHEAVD dataset, 5.0%, 5.4% on the IEMO_dataset1 and 2.8%, 2.6% on the IEMO_dataset2 respectively for unweighted accuracy and weighted accuracy.

In terms of improved performance, it is clear that the FCNs and the Attention-BLSTM models play a role improving the performance. Both the FCNs embedding and attention mechanism can effectively encode temporal dependencies and improve prediction accuracy. The attention mechanism provides a context-informative selection of frames, and FCNs can also complement the performance of the Attention-BLSTM modules for speech emotion recognition.

However, the weighted accuracy improvement margin is not very significant on the two datasets, with only a 4.2% and 4.0% relative improvement over the other baseline models on the CHEAVD dataset and IEMOCAP respectively. In relation to this aspect, the reason is twofold. First, the Attention-BLSTM approach might cause a sub-optimal region where the attention mechanism has very little contribution. Second, even though the FCNs complements the attention mechanism, there is room for further improvement. Currently, the combination of features is linear. A non-linear combination of temporal-spatial feature weights and a non-linear combination of attention weights would be a prospective direction for future study.

In terms of the databases, the overall performance on IEMOCAP is better than CHEAVD for UA and WA. The reason for this seems to be that the IEMOCAP dataset includes four basic emotion classes, while the CHEAVD dataset has eight emotion classes. It can be seen from Table 1 that the instance distribution of the eight emotion classes in the CHEAVD dataset is normally highly imbalanced. The problem of class imbalance or the issue of the scarcity of certain classes in the CHEAVD dataset result in low accuracy. How to develop a model which is more robust to data imbalance for emotion classification requires further study.

6. Conclusion

The hybrid model proposed in the paper is motivated by the existing progress on deep models, and takes advantage of FCNs, Bidirectional-LSTMs, the attention mechanism and their fusions to achieve speech emotion recognition. With the proposed model, we achieved a potent improvement in the current state-of-the-art for the task of speech emotion recognition on the CHEAVD and IEMOCAP datasets. The increase in performance in comparison to other existing models shows that combining Attention-based BLSTM networks with fully connected convolutional neural networks can improve the performance of FCNs modules for speech emotion recognition. An overall analysis of the performance of the Attention-BLSTM-FCN model is provided and compared to other techniques.

In future, we want to investigate the performance of the presented Attention-BLSTM-FCN model on a different database. Furthermore, end-to-end learning strategies and multi-model feature descriptors will also be investigated to further boost the performance of the speech emotion recognition task.

7. Acknowledgements

The work presented in this paper was substantially supported by the National Science Foundation of China(Grant No: 61702370), the technology plan of Tianjin (Grant No: 14RCGFGX00847) and the Open Projects Program of National Laboratory of Pattern Recognition.

8. References

[1] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, no. Mar., pp. 572–587, 3 2011.

[2] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Recognising realistic emotions and affect in speech: State of the art and lessons

learnt from the first challenge," *Speech Communication*, vol. 53, no. 910, pp. 1062–1087, Nov. 2011.

[3] A. Stuhlsatz, C. Meyer, F. Eyben, T. Zielke, G. Meier, and B. Schuller, "Deep neural networks for acoustic emotion recognition: raising the benchmarks," in *2011 IEEE international conference on Acoustics, speech and signal processing (ICASSP)*. Prague, Czech Republic: IEEE, May 2011, pp. 5688–5691.

[4] K. Han, D. Yu, and I. Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," in *Fifteenth Annual Conference of the International Speech Communication Association*, Singapore, Sep. 2014, pp. 223–227.

[5] J. Lee and I. Tashev, "High-level feature representation using recurrent neural network for speech emotion recognition," in *INTERSPEECH*, Dresden,Germany, Sep. 2015, pp. 1537–1540.

[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 *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. Shanghai, China: IEEE, Mar. 2016, pp. 5200–5204.

[7] A. Metallinou, M. Wollmer, A. Katsamanis, F. Eyben, B. Schuller, and S. Narayanan, "Context-sensitive learning for enhanced audiovisual emotion classification," *IEEE Transactions on Affective Computing*, vol. 3, no. 2, pp. 184–198, Jan. 2012.

[8] T. Zhang, W. Zheng, Z. Cui, Y. Zong, and Y. Li, "Spatial-temporal recurrent neural network for emotion recognition," *IEEE Transactions on Cybernetics*, vol. PP, pp. 1–9, Jan. 2018.

[9] 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, Sep. 2014.

[10] H. Wang, Z. Cui, Y. Chen, M. Avidan, A. B. Abdallah, and A. Kronzer, "Predicting hospital readmission via cost-sensitive deep learning," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, pp. 1–10, Apr. 2018.

[11] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, Boston, Massachusetts,USA, June 2015, pp. 3431–3440.

[12] S. Mirsamadi, E. Barsoum, and C. Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, New Orleans, LA, USA, Mar. 2017, pp. 2227–2231.

[13] C. W. Huang and S. S. Narayanan, "Attention assisted discovery of sub-utterance structure in speech emotion recognition," in *17th Annual Conference of the International Speech Communication Association*, San Francisco, California, USA, Sep. 2016, pp. 1387–1391.

[14] T. N. Sainath, O. Vinyals, A. Senior, and H. Sak, "Convolutional, long short-term memory, fully connected deep neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Brisbane, QLD, Australia, Apr. 2015, pp. 4580–4584.

[15] T. N. Sainath, R. J. Weiss, A. Senior, K. W. Wilson, and O. Vinyals, "Learning the speech front-end with raw waveform cldnns," in *Sixteenth Annual Conference of the International Speech Communication Association*, Dresden, Germany, Sep. 2015, pp. 1–5.

[16] C. W. Huang and S. S. Narayanan, "Deep convolutional recurrent neural network with attention mechanism for robust speech emotion recognition," in *IEEE International Conference on Multimedia and Expo*, Hong Kong, China, July 2017, pp. 583–588.

[17] W. Lim, D. Jang, and T. Lee, "Speech emotion recognition using convolutional and recurrent neural networks," in *Signal and information processing association annual summit and conference (APSIPA), 2016 Asia-Pacific*. Jeju, South Korea: IEEE, Dec. 2016, pp. 1–4.

[18] X. Li and X. Wu, "Long short-term memory based convolutional recurrent neural networks for large vocabulary speech recognition," *arXiv preprint arXiv:1610.03165*, Sep. 2016.

[19] G. Keren and B. Schuller, "Convolutional rnn: an enhanced model for extracting features from sequential data," in *Neural Networks (IJCNN), 2016 International Joint Conference on*. Vancouver, BC, Canada: IEEE, July 2016, pp. 3412–3419.

[20] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. Aug., pp. 1735–1780, Nov. 1997.

[21] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[22] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *28th International Conference on Neural Information Processing Systems*, Montreal, Canada, Dec. 2015, pp. 577–585.

[23] O. Vinyals, Ł. Kaiser, T. Koo, S. Petrov, I. Sutskever, and G. Hinton, "Grammar as a foreign language," in *28th International Conference on Neural Information Processing Systems*, Dec. 2015, pp. 2773–2781.

[24] Z. Wang, W. Yan, and T. Oates, "Time series classification from scratch with deep neural networks: A strong baseline," in *2017 International Joint Conference on Neural Networks (IJCNN)*. Anchorage, Alaska: IEEE, May 2017, pp. 1578–1585.

[25] M. Lin, Q. Chen, and S. Yan, "Network in network," *arXiv preprint arXiv:1312.4400*, Dec. 2013.

[26] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning*, Lille, France, July 2015, pp. 448–456.

[27] L. Trottier, P. Giguere, and B. Chaib-draa, "Parametric exponential linear unit for deep convolutional neural networks," *arXiv preprint arXiv:1605.09332*, pp. 1–16, 2016.

[28] F. Karim, S. Majumdar, H. Darabi, and S. Chen, "Lstm fully convolutional networks for time series classification," *IEEE Access*, vol. 6, pp. 1662–1669, Dec. 2018.

[29] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, June 2014.

[30] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, p. 335, Nov. 2008.

[31] Y. Li, J. Tao, L. Chao, W. Bao, and Y. Liu, "Cheavd: a chinese natural emotional audio–visual database," *Journal of Ambient Intelligence and Humanized Computing*, vol. 8, no. 6, pp. 913–924, Sep. 2017.

[32] B. Mathieu, S. Essid, T. Fillon, J. Prado, and G. Richard, "Yaafe, an easy to use and efficient audio feature extraction software." in *ISMIR*, Utrecht, Netherlands, Aug. 2010, pp. 441–446.

[33] Y. Bengio, "Deep learning of representations for unsupervised and transfer learning," in *Proceedings of ICML Workshop on Unsupervised and Transfer Learning*, Washington, USA, June 2012, pp. 17–36.

[34] R. Xia and Y. Liu, "Dbn-ivector framework for acoustic emotion recognition," in *17th Annual Conference of the International Speech Communication Association*, San Francisco, USA, Sep. 2016, pp. 480–484.

[35] M. Neumann and N. T. Vu, "Attentive convolutional neural network based speech emotion recognition: A study on the impact of input features, signal length, and acted speech," in *proceedings of Interspeech 2017*, Aug. 2017, pp. 1263–1267.

[36] Y. Li, J. Tao, B. Schuller, S. Shan, D. Jiang, and J. Jia, "MEC 2017: the multimodal emotion recognition challenge," in *the first Asian Conference on Affective Computing and Intelligent Interaction*, Beijing, China, May 2018.

[37] I. J. Tashev, Z.-Q. Wang, and K. Godin, "Speech emotion recognition based on gaussian mixture models and deep neural networks," in *Information Theory and Applications Workshop (ITA), 2017*. San Diego, CA, USA: IEEE, Feb. 2017, pp. 1–4.