Emotion Recognition from Variable-Length Speech Segments Using Deep Learning on Spectrograms

Xi Ma[1,3], *Zhiyong Wu*[1,2,3], *Jia Jia*[1,3], *Mingxing Xu*[1,3], *Helen Meng*[1,2], *Lianhong Cai*[1,3]

[1]Tsinghua-CUHK Joint Research Center for Media Sciences, Technologies and Systems, Graduate School at Shenzhen, Tsinghua University, Shenzhen 518055, China
[2]Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong SAR, China
[3]Tsinghua National Laboratory for Information Science and Technology (TNList), Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China

max15@mails.tsinghua.edu.cn, {zywu,hmmeng}@se.cuhk.edu.hk, {jjia,xumx,clh-dcs}@tsinghua.edu.cn

Abstract

In this work, a approach of emotion recognition is proposed for variable-length speech segments, which applied deep neutral network to spectrograms directly. The spectrogram contains more comprehensive para-lingual information than the traditional acoustic features, so it would be better to be the input for the emotion recognition system. We tried to extract such information from spectrograms and accomplish the emotion recognition task by combining Convolutional Neural Networks (CNNs) with Recurrent Neural Networks (RNNs). Meanwhile, in order to handle the variable-length speech segments, the neural network structure we constructed can deal with the variable-length input sequence. Compared to split the speech input into the smaller and fixed-length segments, our method can solve the loss of accuracy introduced in the speech segmentation process. We evaluated the emotion recognition model on the IEMOCAP dataset over four emotions. Experimental results demonstrate that the proposed method outperforms the fixed-length neural network on both weighted accuracy (WA) and unweighted accuracy (UA).

Index Terms: Speech Emotion Recognition, Variable-Length Speech Segments, Spectrogram, Deep Neural Network.

1. Introduction

Emotion recognition plays an important role in many applications, especially in human-computer interaction systems that are increasingly common today. As one of the main communication media between human beings, voice has attracted wide attentions from researchers [1]. Speech contains a wealth of emotional information. How to extract such information from speech signal is of great importance for automatic speech emotion recognition.

As an important part of speech emotion recognition, the extraction of the most relevant acoustic features has attracted lot of research interests. Most of these researches are devoted to designing some hand-crafted features that is most distinctive for emotion recognition. Recently, however, a trend in the machine learning community has emerged towards deriving a representation of the input signal directly from raw, unprocessed data. The reason behind this idea is that, ultimately, the network learns an intermediate representation of the raw input signal automatically that better suits the task at hand and hence leads to improve performance. Motivated by this, we tried to construct a emotion recognition system by leveraging spectrogram and deep neural network.

A spectrogram is a time-frequency decomposition of a signal that indicates its frequency content over time. In our work, Convolutional Neural Networks (CNNs) are first constructed to learn effectively spatial spectrogram patterns that represent the emotional information; Recurrent Neural Network (RNNs) are then used to model the temporal structure across the sentence being represented by the spectrogram; the final emotion categories are derived by a full-connected layer.

The idea of this work is similar to the previous Satt's work [2]. However, our neural network possesses the advantages of being able to handle the variable-length speech segments. Compared to split the speech input in the smaller and fixed-length segments, our method can solve the loss of accuracy introduced in the speech segmentation process. In the IEMOCAP dataset [3], we can achieve an weighted accuracy (WA) of 71.45% using 5-folds (leave-one-session out) cross validation, which is a 2.65% absolute (3.85% relative) improvement over the fixed-length approach. The unweighted accuracy (UA) on the same dataset is 64.22%, which also outperforms the fixed-length approach with 4.82% absolute (8.11% relative) improvement.

The reset of the paper is organized as follows. Section 2 summarizes the previous related work. The spectrogram based speech emotion recognition framework and the variable-length neural network structure is then detailed in Section 3. Experiments results are presented in Section 4. Section 5 concludes the paper.

2. Related Work

As a common issue for speech emotion recognition, feature extraction aims to design functions that map the raw speech signal to the most relevant representation to the emotions. There have been many studies on feature extraction for speech emotion recognition. In [4–7], prosody-based acoustic features, including pitch-related, energy-related and timing features have been widely used for recognizing speech emotion. Spectral-based acoustic features also play important role in emotion recognition, such as Linear Prediction Coefficients (LPC) [8], Linear Prediction Cepstral Coefficients (LPCC) [9] and Mel-frequency Cepstral Coefficients (MFCC) [10]. In [11], voice quality features have also been shown to be related to emotions.

In the recent years, deep learning methodologies and tools have been introduced to speech processing area, used for feature extraction, classification/regression, or both [12–17]. Researchers have shown that replacing hand-crafted low-level (frame-level) features with statistical learning by the raw signal with different layers of the deep network can significantly enhance the accuracy of classification and regression solutions. In the speech recognition, one of the first studies that suggested learning better features for automatic speech recognition (ASR) that used directly the speech waveform was Jaitly and Hinton [18]. Although they did not train the system in an end-to-end manner, they proposed learning an intermediate representation by training a Restricted Boltzmann Machine directly on the speech time signal. Bhargava and Rose [19] used stacked bottleneck deep neural networks (DNNs) trained on windowed speech waveforms and obtained results only slight worse than corresponding MFCC on the same architecture. Sainath et al. match the performance of a large-vocabulary speech recognition (LVCSR) system based on log-Mel filterbank energies by using a Convolutional, LSTM-DNN [20, 21]. In [22, 23], a recently published state of the art robust speech recognition system is described based on linearly-spaced spectrograms. Besides, direct use of Mel-scale spectrograms for speaker recognition was proved successful as well [24].

In the field of speech emotion recognition, several studies have been carried out using deep neural network for feature learning. Recently, George et al. [25] proposed a convolutional recurrent neural network that operates on the raw signal, to perform an end-to-end spontaneous emotion prediction task from speech data. Satt et al. [2] also combined CNN with LSTM to classify emotions from linearly-spaced spectrograms, which achieves beyond the state-of-the-art accuracy on the common benchmarking dataset IEMOACP. However, all of these methods split the speech input into the smaller and fixed-length segments, which introduce the loss of accuracy in the training and predicting stage. Our method proposed to use a variable-length neural network to solve this problem.

3. The Proposed Recognition System

In [2], they first split each sentence longer than 3 seconds to shorter sub-sentences of equal lengths. The part of no longer than 3 second is padded to 3 second with zero after extracting spectrograms. Each sub-sentence is assigned the emotion labeling of the corresponding whole sentence. These shorter sentences are used throughout the proposed system, where only during the testing phase they evaluate the prediction for the whole sentences by averaging the posterior probabilities of the respective sub-sentences. Although this method can reduce the difficulty for constructing neural network (ensure that the lengths of inputs are equal), some errors will also be introduced. First, it is not a good treatment to assign each sub-sentence to the emotion label of the corresponding whole sentence. The observation indicates that only a part of voice contains obvious non-neutral emotional information in the non-neutral emotional sentence. In other word, some shorter sub-sentences, which belong to a sentence with non-neutral emotion, may not contains any non-neutral emotional information. Such sub-sentences are used as the inputs for training neural network, which may lead to reduce accuracy of the model. Second, in the predicting stage, the sentence also need to split into sub-sentences and average the posterior probabilities to get the final result. This would also cause the loss of some accuracy. Our study aims to solve this problem. We first extract the spectrogram for the whole sentence, and then the spectrogram, as the input of deep neural network, is classified to the emotion categories, as shown in Figure 1. For comparison, we use the similar spectrogram extraction setting and neural network as those used in [2].

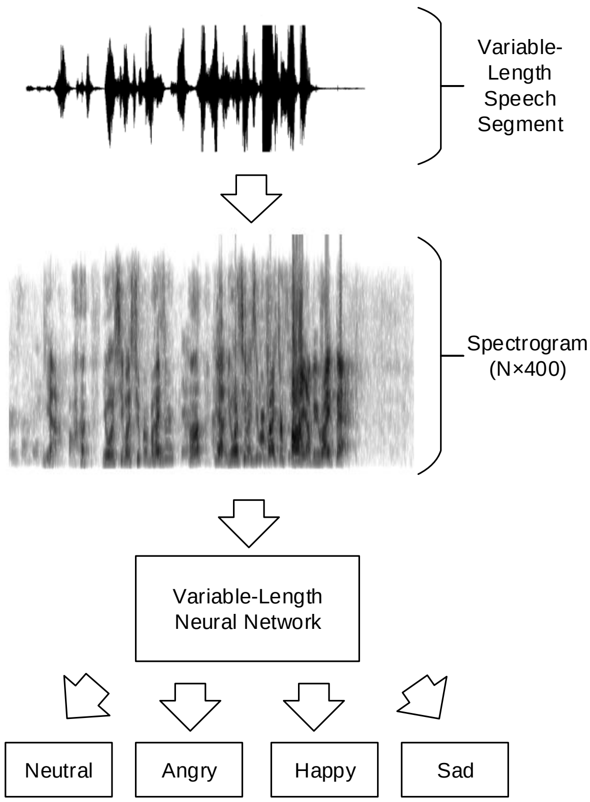

Figure 1: *Flow chart of our emotion recognition system.*

3.1. Spectrogram Extraction

The speech signal in the IEMOCAP dataset is sampled at 16KHz and organized as single sentences with duration from less than a second to about 20 second. Each sentence is labeled with one emotion. A sequence of overlapping Hamming windows is applied the speech signal, with frame step (window shift) of 10msec, and frame length (window size) of 40 msec. For each frame we calculate a DFT of length 1600 (for 10Hz grid resolution). We use the frequency range of 0-4KHz, ignoring the rest. Following aggregation of the short-time spectra, we obtain a matrix of size NxM, where N according to the selected timing grid resolution, which is variable for different sentences, and M=400 according to the selected frequency grid resolution.

Next, we implement a normalization step. First, the DFT data is converted to log-power-spectrum, expressed in dB; and then we normalized the energy of frequency domain using z-normalization with mean and standard deviation of training dataset.

At last, for improving the computing efficiency, the spectrograms, whose timing length have little difference, need to be putted in the same batch and padded to the max size of spectrograms in the corresponding batch data with zero. The is can be achieved by sorting sentence lengths. Parallel computing can be achieved for a batch of samples during training stage.

3.2. Deep Neural Network

As we all known, convolutional neural networks (CNNs) can be thought of as a kind of neural network that uses many identical copies of the same neuron. This allows the network to have lots of neurons and express computationally large models while keeping the number of actual parameters − the values describing how neurons behave − that need to be learned fairly small. A lot of works have been carried out using CNNs for feature learning. The common used means, especially in the computer vision, is to process the inputs of the same size. This is convenient to connect other neural network, such as full connected layer. But in fact, the convolutional neural network just need to train convolutional kernels, so it can also be trained, even if the input size is different.

Recurrent Neural Networks (RNNs) are popular models that have shown great promise in many sequence modeling tasks. They perform the same task for every element of a sequence, with the output being depended on the previous computations. For the computing efficiency, the input sequence is usually fixed-length. The variable-length sequence is usually padded into the same length. But we can ignore the outputs of invalid padding time-steps, so that the variable-length sequence can be processed correctly.

In our work, the input sequence are padded into the same length with zero in the same batch during training stage, and will have no padding during predicting stage, so our neural network need to possess the capacity to avoid the interference of padding value on the output. Let $S = [x_1, x_2, ..., x_V, ..., x_T]$ is a input sequence, where $S1 = [x_1, x_2, ..., x_V]$ is the valid part and $S2 = [x_{V+1}, x_{V+2}, ..., x_T]$ is the padding part.

First, for the convolution neutral network, we can use masking to reserve the outputs from $S1$ and ignore the outputs from $S2$, which can be represented as follows:

$$S_{conv} = Conv(S) \bullet Mask(S) \tag{1}$$

where $Conv(S)$ is the output of convolution layer for S, $Mask(S)$ is a masking matrix, and $S_{conv} = [y_1, y_2, ..., y_V, ..., y_T]$ is the output sequence with the same length as S. The values of masking matrix are ones for the valid part, and zeros for the padding part. The valid output can be achieved by element-wise multiply between $Conv(S)$ and $Mask(S)$. Besides, convolutional layers are often interweaved with pooling layers. We need to take care of the border value between the valid part and the padding part, which could introduce the invalid information. For example, assuming S_{conv} is the input of the max-pooling layer. If the pooling kernel size is 2 and the input path contains y_V and y_{V+1}, the output will be y_{V+1} when $y_V < 0$ and $y_{V+1} = 0$. But the expected value should be y_V because y_{V+1} is a padding value. In our experiment, this problem will lead to that the neural network does not converge. Hence, the y_V will be masked as zero before inputting to the max-pooling layer in our design. In this way, padding or no padding, the same input will get the same output after convolution layer and pooling layer. It ensures the consistency in training stage and predicting stage, because there is no padding in the predicting stage.

Second, for the recurrent neural network, because speech emotion recognition is a sequence classification problem, we just need the output in the last valid time-step. Assuming S is the input of recurrent neural network, the expected result should be the output at $t = V$. Besides, in the bi-directional recurrent neural network, the output of backward recurrent neural network should be at $t = 0$. The final output is the concatenation of the outputs in forward and backward recurrent neural network.

We hypothesized that the convolutional neural networks, capable of learning spatial patterns, will learn effectively spatial spectrogram patterns that represent the emotional information. We also hypothesized that adding an LSTM layer will help learning the temporal behavior across the sentence being represented by the spectrogram. Dozens of combinations of topologies and parameters were tested. The best deep neural network topology is depicted in Figure 2.

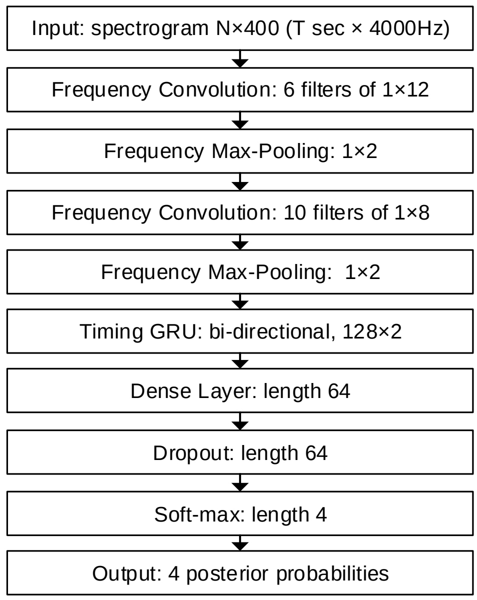

Figure 2: *Deep neutral network topology*

Since the length of sentence is different, we will assign different weights to the loss, in inverse proportion to the length. Besides, the IEMOCAP corpus is significantly unbalanced, the number of some emotional categories is obviously more than other emotional categories. The weight, in inverse proportion to the class size, will also be assigned to the loss.

4. Experiments

4.1. Experimental Setup

In this work, the IEMOCAP dataset [3] is used for conducting the experiments. The dataset was designed for studying multimodal expressive dyadic interactions. It was collected using motion capture and audio/video recording (approximately a total of 12h) over 5 dyadic sessions with 10 subjects. Each session consists of a different dyad where one male and one female actor perform scripted plays and engage in spontaneous improvised dialogs elicited through affective scenario prompts. At least three evaluators annotated each utterance in the dataset with the categorical emotion labels chosen from the set: happy, sad, neutral, angry, surprised, exited, frustration, disgust, fear and other. We consider only the utterances with majority agreement (at least two out of three evaluators gave the same emotion label) over the emotion classes of: Angry, Happy, Sad and Neutral. The dataset contains scripted and improvised dialogs. As the script text exhibits strong correlation with the labeled emotions, it may give rise to lingual content learning, at least partially, which is an undesired side effect. Therefore we used the improvised data only. Such configuration is the same as [2], which makes the experimental results comparable between our work and [2].

Our work focuses on speaker independent emotion recognition, hence the 5-folds cross-validation method is used to conduct the experiments. In each fold, the data from four sessions is used for training the deep neural network, and the data from the remaining session is split – one speaker for validation and the other for the accuracy testing.

The experimental results can be divided into two parts. In the first part, we show the spectrogram of different speech segments belonging to the same sentence, whose difference can be used to verify the our assumption about only a part of voice contains obvious non-neutral emotional information in the non-neutral sentence. In the second part, we compare and analyze the performance of the variable-length neural network and the fixed-length neural network.

4.2. Experimental Results

It is informative to examine the difference between two speech segments from the same sentence, by looking at spectrograms. The Figure 3 below show the spectrograms of two different speech segments (length is 3s), which from the same sentence labeled as Angry.

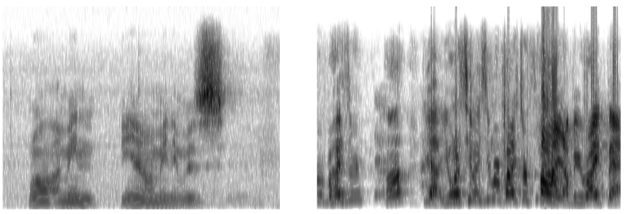

Figure 3: *Left: non-emotional spectrogram; Right: emotional spectrogram*

In the Figure 3, the left side – shows the non-emotional spectrogram; the right side – shows the emotional spectrogram. The horizontal axis denotes the time, and the vertical denotes the frequency. The gray scale represents the energy strength – darker colors designate higher energy. It is clearly seen that the right spectrogram shows higher energy than the left spectrogram in most of frequency range. In general, the loudness of angry voice is higher, so the difference of two spectrograms can reflect that the right speech segment looks more like a angry voice than the left speech segment. Besides, through artificial listening, we also find that the right speech segment contains more intense anger, but the left speech segment is more like a neutral speech segment. If all of two speech segments are labeled as Angry, it may cause confusion between Neutral and Angry in the model training and lead to the actual neutral speech segment is misclassified. However, when we listen to the whole sentence, the front neutral speech segment can enhance angry feeling of the back angry speech segment, the so-called angry ending without angry beginning. This problem also occurs in the other non-neutral emotions. Hence, this proves that the whole sentence as the input is more reasonable.

Next, we conduct emotion recognition experiments by reporting the weighed accuracies (WA) and the unweighted accuracy (UA) of different methods, where WA is the accuracy of all samples in the test set and UA is the average value of the accuracy values of all emotions. Both metrics are standard measurements used in several previous emotion recognition challenge and is adopted in [2]. Also, owing to the parameters of our network is different from [2], we also show the result about the fixed-length speech segment as the input for our network. This is to ensure that the improvement of performance don't come from parameters changing.

In Table 1, we present weighted accuracy (WA) and unweighted accuracy (UA) in the IEMOCAP dataset, where "Baseline" represents the result reported in [2], "Fixed-Length" represents the fixed-length inputs for our neural network, "Variable-Length" represents the variable-length inputs for our neural network. As can be seen, the WA of "Variable-Length" is 71.45%, which is a 2.65% absolute (3.85% relative) improvement over "Baseline"; And the UA also reaches 64.22%, which is a 4.82% absolute (8.11% relative) improvement compared to "Baseline". Besides, the WA and UA of "Fixed-Length" is also lower than "Variable-Length", so this proves that the improvement of performance don't come from parameters changing.

To further analyze the experimental result, we present the confusion matrices of both the fixed-length input and the variable-length input for our network in Table 2 and Table 3. We can find that the recognition accuracies of Neutral for our variable-length network are improved. As we have made the analysis for Figure 3, this is because the confusion between Neutral and other emotions can be alleviated by inputting the whole sentence to the network. Besides, the accuracy of Happy is improved significantly. This is caused by the neutral speech segment in the Happy sentence may be misclassified to other emotions, because these neutral speech segments are very similar by using fixed-length model. Meanwhile, the accuracy of Sad is decreased. It may be that the sample num of Sad is more than other non-neutral emotions, so the increase recall of other non-neutral emotion will lead to the decrease recall of Sad. These results indicate that the variable-length neural network really provides improvement in the recognition accuracy of some emotions.

Table 1: *Comparison of weighted accuracy (WA) and unweighted accuracy (UA) on IEMOCAP dataset.*

	WA	UA
Baseline	68.8%	59.4%
Fixed-Length	68.86%	57.45%
Variable-Length	71.45%	64.22%

Table 2: *Confusion matrix by using the fixed-length neutral network.*

Predict / Actual	Neutral	Angry	Happy	Sad
Neutral	**71.75%**	8.88%	5.93%	13.45%
Angry	30.84%	**58.79%**	7.05%	3.34%
Happy	55.92%	31.42%	**11.72%**	0.95%
Sad	11.41%	0%	1.02%	**87.57%**

Table 3: *Confusion matrix by using the variable-length neutral network.*

Actual \ Predict	Neutral	Angry	Happy	Sad
Neutral	**73.64%**	2.74%	12.41%	11.21%
Angry	11.44%	**59.55%**	26.52%	2.5%
Happy	45.2%	13.81%	**40.05%**	0.95%
Sad	15.89%	0%	0.48%	**83.64%**

5. Conclusion

In this paper, we propose a variable-length neutral network that operates on the spectrogram, to perform an end-to-end emotion classification task from variable-length speech segments. Through inputting the whole sentence into the model, our method can effectively alleviate the confusion between Neutral and other emotions introduced in the speech segmentation process. The weighed accuracies (WA) and the unweighted accuracy (UA) achieves beyond the state of the art accuracy on the common benchmarking dataset IEMOACP, comparing with the previous work of fixed-length neural network. In the future, we will continue to explore how to use other deep neutral network structure to handle variable-length speech emotion recognition.

6. References

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

[2] A. Satt, S. Rozenberg, and R. Hoory, "Efficient emotion recognition from speech using deep learning on spectrograms," in *INTERSPEECH*, 2017, pp. 1089–1093.

[3] 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," *Journal of Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[4] X. Ma, Z. Wu, J. Jia, M. Xu, H. Meng, and L. Cai, "Speech emotion recognition with emotion-pair based framework considering emotion distribution information in dimensional emotion space," in *INTERSPEECH*, 2017, pp. 1238–1242.

[5] C. Busso, S. Lee, and S. Narayanan, "Analysis of emotionally salient aspects of fundamental frequency for emotion detection," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 4, pp. 582–596, 2009.

[6] R. Cowie, E. Douglascowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor, "Emotion recognition in human-computer interaction," *IEEE Signal Processing Magazine*, vol. 18, no. 1, pp. 32–80, 2001.

[7] E. Vayrynen, J. Kortelainen, and T. Seppanen, "Classifier-based learning of nonlinear feature manifold for visualization of emotional speech prosody," *IEEE Transactions on Affective Computing*, vol. 4, no. 1, pp. 47–56, 2013.

[8] M. G. Bellanger, *Digital processing of speech signals*. Prentice-Hall, 1978.

[9] B. S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *the Journal of the Acoustical Society of America*, vol. 55, no. 6, pp. 1304–1312, 1974.

[10] S. B. Davis and P. Mermelstein, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[11] G. Christer and A. N. Chasaide, "The role of voice quality in communicating emotion, mood and attitude," *Speech Communication*, vol. 40, no. 1-2, pp. 189–212, 2003.

[12] K. Han, D. Yu, and I. Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," in *INTERSPEECH*, 2014.

[13] J. Lee and I. Tashev, "High-level feature representation using recurrent neural network for speech emotion recognition," in *INTERSPEECH*, 2015.

[14] Z. Huang, M. Dong, Q. Mao, and Y. Zhan, "Speech emotion recognition using cnn," in *Acm International Conference on Multimedia*, 2014, pp. 801–804.

[15] D. Le and E. M. Provost, "Emotion recognition from spontaneous speech using hidden markov models with deep belief networks," in *Automatic Speech Recognition and Understanding*, 2013, pp. 216–221.

[16] R. Rana, "Emotion classification from noisy speech - a deep learning approach," *arXiv preprint arXiv:1603.05901*, 2016.

[17] V. Chernykh, G. Sterling, and P. Prihodko, "Emotion recognition from speech with recurrent neural networks," *arXiv preprint arXiv:1701.08071*, 2017.

[18] N. Jaitly and G. Hinton, "Learning a better representation of speech soundwaves using restricted boltzmann machines," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2012, pp. 5884–5887.

[19] M. Bhargava and R. Rose, "Architectures for deep neural network based acoustic models defined over windowed speech waveforms," in *INTERSPEECH*, 2015.

[20] 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*, 2015, pp. 4580–4584.

[21] T. N. Sainath, R. J. Weiss, A. W. Senior, K. W. Wilson, and O. Vinyals, "Learning the speech front-end with raw waveform cldnns," in *INTERSPEECH*, 2015, pp. 1–5.

[22] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, and A. Coates, "Deep speech: Scaling up end-to-end speech recognition," *Computer Science*, 2014.

[23] D. Amodei, R. Anubhai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, J. Chen, M. Chrzanowski, A. Coates, and G. Diamos, "Deep speech 2: End-to-end speech recognition in english and mandarin," *Computer Science*, 2015.

[24] E. Variani, X. Lei, E. Mcdermott, I. L. Moreno, and J. Gonzalez-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2014, pp. 4052–4056.

[25] 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.