Attention-based Region of Interest (ROI) Detection for Speech Emotion Recognition

Jay Desai
Department of Computer Science
New York Institute of Technology
New York, USA
jdesai09@nyit.edu

Houwei Cao
Department of Computer Science
New York Institute of Technology
New York, USA
hcao02@nyit.edu

Ravi Shah
Department of Computer Science
New York Institute of Technology
New York, USA
rshah79@nyit.edu

***Abstract*—Automatic emotion recognition for real-life applications is a challenging task. Human emotion expressions are subtle, and can be conveyed by a combination of several emotions. In most existing emotion recognition studies, each audio utterance/video clip is labelled/classified in its entirety. However, utterance/clip-level labelling and classification can be too coarse to capture the subtle intra-utterance/clip temporal dynamics. For example, an utterance/video clip usually contains only a few emotion-salient regions and many emotionless regions. In this study, we propose to use attention mechanism in deep recurrent neural networks to detection the Regions-of-Interest (ROI) that are more emotionally salient in human emotional speech/video, and further estimate the temporal emotion dynamics by aggregating those emotionally salient regions-of-interest. We compare the ROI from audio and video and analyse them. We compare the performance of the proposed attention networks with the state-of-the-art LSTM models on multi-class classification task of recognizing six basic human emotions, and the proposed attention models exhibit significantly better performance. Furthermore, the attention weight distribution can be used to interpret how an utterance can be expressed as a mixture of possible emotions.**

***Index Terms*—speech recognition, LSTM, ROI.**

I. Introduction

Human emotion plays an important role in everyday decision making. It impacts the way they communicate i.e. via body language, facial expressions, verbal communication, personality, etc [1], [2] and also shows the characteristics of the person like [3] leadership, honesty, teamwork. In fact, in a book Emotional Intelligence 2.0 by Travis Bradberry, Ph.D. it was found that 90% of all the people they studied at work had higher EQs. Speech is the key aspect of expressing emotion and most basic way of human-human and human-machine interactions. In this study, we use the acoustic features of speech to detect the part/region of speech which impacts the most in detecting the emotional state of the subject.

Speech emotion recognition (SER) is a trending topic in the field of research with the rapid use of artificial intelligence and machine learning in our lives. While there are many techniques and features that has been proposed [4], [5] but it is still not clear on which gives most information about emotions. There are two common types of feature extraction methods, Global features and Local features. Features extracted for whole audio file using full audio clip are usually known as global features while features extracted using multiple time frames at regular intervals usually 20-30ms with or without overlap are known as local features. Global features usually work well for general machine learning models whereas local features are better suited for deep neural networks. Features are the Low Level Descriptors (LLDs) which are believed to affect the most to emotions and High Level Statistical Functions (HSFs) which are applied to LLDs to extract variations and contours for temporal description of the data. While the majority of researchers have agreed that global features perform better than local features [6]–[9] but other researchers also believe that global features only show better performance for high arousal emotions like anger, fear, joy [10]. Here are some common LLDs and HSFs:

TABLE I
Common LLDs and HSFs [13]–[15]

LLDs	HSFs
pitch, MFCCS, voicing probability, energy, Zero-Crossing rate, formant locations/ bandwidths, harmonics-to-noise ratio, jitter, loudness, etc.	mean, variance, min, max, range, median, quartiles, skewness, kurtosis, linear regression coefficients, RMSenergy, SMA, PCM, Rfilters, etc.

In recent years, there has been advancements in the field of machine learning and artificial intelligence techniques and their applications. The authors in [16]–[20] used ensemble of SVM instead of just one to better classify emotions using MFCC, total energy and f0 features from audio signal. The work done in [21], [22] focused on using Deep Neural Networks and Recurrent Neural Network (RNNs) to learn short term acoustic features at frame level and then mapping them to sentence-level representation using extreme learning machines (ELM). The authors in [23] used various LSTM techniques and used logistic regression based general attention mechanism with weighted pooling for classification.

The authors in [24] used deep convolutional recurrent neural networks where they used a combination of convolutional neural network (CNN) layers and bi-directional LSTM layers. They derived a generalized attention mechanism method from the methods proposed in [25], [26] and compare CNNs task-specific spectral decorrelation with that of the discrete cosine transformation (DCT) in clean and noisy conditions. One issue

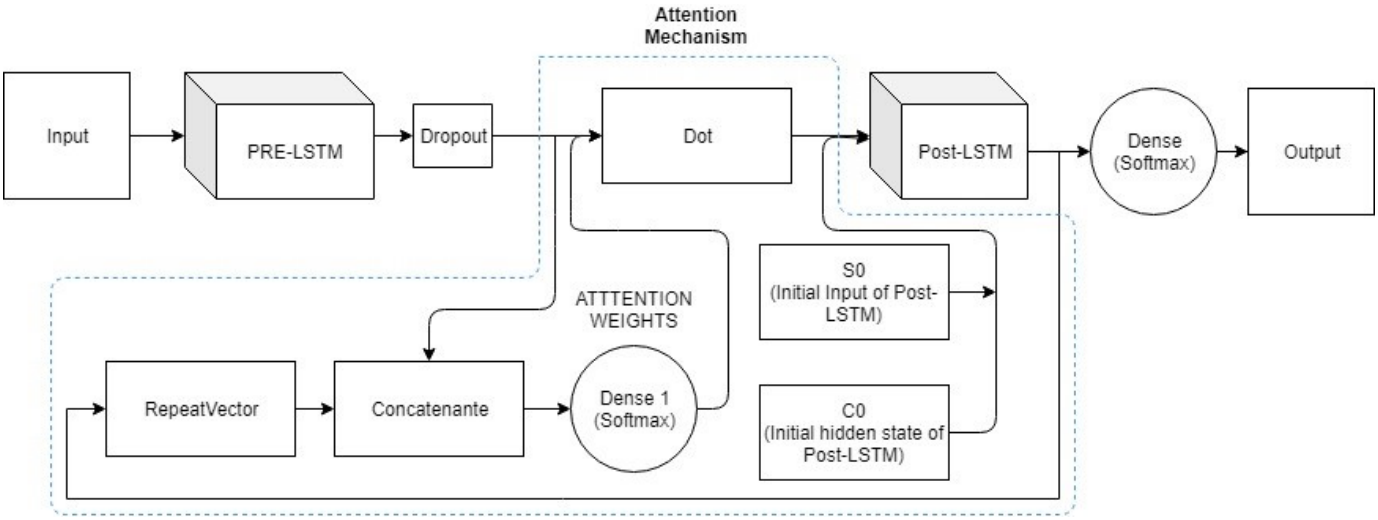

Fig. 1. Architecture for Model 1 LSTM with Attention (the dotted line shows the attention mechanism block).

that appears in most SER datasets is their labels are given at utterance level and in any speech, there are many silence periods and only a short time of utterance of few words that impact the overall emotion. The silent frames can be handled by labelling them as null, our approach handles it implicitly without any separate or explicit mechanisms to handle it.

In this paper, we used an LSTMs in an encoder - decoder based fashion proposed in [27], [28] using LSTM. Different from the attention technique used in [23] where they used logistic regression and mean pooling, in our approach for attention, we train a separate small Neural Network to learn how much attention to pay on each frame. Our method works similar to human mechanism for translating the input signals in brain and using only contextually important parts to decode the input and return the results as emotions. This has been proven better in many language translation tasks where to translate one word, only a part of the sentence is important rather than whole sentence. In the following we discuss various approaches we used for classification.

II. Neural Network Model

In the below subsections data description describing which dataset is used and details about it, features extraction techniques describing all the features extracted tools or libraries used, testing strategy, various models have been described.

A. Data Description

The dataset used here is Crowd-sourced Emotional Multimodal Actors Dataset (CREMA-D) [29]. It is a dataset of 7442 audio files from 91 different subjects - 48 males and 43 females of various races such as African, Asian, American, Hispanic, Caucasian between the ages of 20 and 74. Subjects spoke 12 sentences in 6 different emotions i.e. Anger, Disgust, Fear, Happy, Neutral and Sad in four emotion levels - low, medium, high and unspecified. Categorical emotion labels and real-value intensity values for the perceived emotion were collected using crowd-sourcing from 2,443 raters. All 91 actors read 12 sentences in three to four emotion levels making total number of files to be 7442 clips. Each actor has an average of 82 audio clips. The problems with other datasets such as low recognition rates of human subjects [10], bad quality of recorded utterances is not found in this dataset. The human recognition rates for intended emotion for audio only data was 40%. Recognition rates were highest for neutral followed by happy, anger, disgust, fear and sad.

B. Feature Extraction

Every audio clip has been padded with zeros till maximum audio clip length to remove the inconsistencies between audio clip length. The first 13 Mel-frequency cepstral coefficients are extracted for every 20 milliseconds (frame length) with 50% or 10 milliseconds (frame step) of overlap between each frame for the data used in model 1,2,3 and 4.

C. Testing Strategy

We have used Leave One Subject Out (LOSO) strategy to train-test-split the data. In that for every 91 subjects that a model gets trained on, it gets tested on one subject. The total training files for one model comprised of 7360 training files and testing on one subject having 82 files. In total of 92 models were trained and tested and for the final result, mean from all the results from 91 models was taken and a confusion matrix was created as shown in Fig. 3. It gives a better information on how the model will perform on real world scenario because it has never seen the subject before than other methods such as k-fold cross validation or random train-test split where the model has already seen and trained on the audio clips of subject and can perform better on them in testing phase.

D. Model 1 - LSTM with Attention

We have used a Sequence to Sequence type model which has following parts:

1) Encoder (Pre-LSTM)
2) Intermediate Attention Layer environment

3) Decoder (Post-LSTM)
4) Dense (SoftMax) Layer

The Pre-LSTM and Post-LSTM layers are CuDNNLSTM and are unidirectional. The output of Pre-LSTM is given to the attention mechanism layer explained below which returns attention weights. These attention weights are dotted with the output of Pre-LSTM and fed to Post-LSTM followed by dense or fully connected layer with softmax activation function which calculates the probability for 6 classes. Dropout is used to reduce overfitting for regularization.

In general, LSTMs have ability to learn from the sequence of data and also learn the dependencies between sequences and store them in the memory cell for future use. It stores the relevant information and forgets the irrelevant information using the forget gate and passes on this learnt context to next time steps. The context learnt here is the result of backpropagation and loss minimization to optimize the overall accuracy. The attention mechanism used in model 1 and model 2 is described and explained below.

Attention Mechanism: The attention block shown in figure 1 shows the architecture of attention mechanism in the dotted line. Models without attention passes the output from dropout layer directly to post-LSTM layer. The similar attention mechanism is used in [11], [12] for generating image captions while focusing on only parts of the image.

There are two types of LSTM layers used for pre-LSTM, one is uni-directional and other is bi-directional. The post-LSTM passes outputs $o < t >$ and hidden cell state $h < t >$ from one time step to next. The inputs to post-LSTM are $s < t >$, context and $h < t >$. The outputs of pre-LSTM are represented as $p < t >$ for unidirectional and for bidirectional LSTM, the forward and backward direction, the outputs are concatenated $p < t >= [p < t > (forward), p < t > (backward)]$. The repeatVector copies $o < t - 1 >$ for x times where x is the number of time frames used to extract the data and then Concatenate layer concatenates it with $p < t >$ to compute $e < t, t' >$ which is then passed to dense layer and then softmax layer to output $a < t, t' >$ where the t represents the post-LSTM's time step and t' represents the pre-LSTM's time step.

At any time step t, given the outputs of pre-LSTM $[p < 1 >, p < 2 >, \ldots, p < x >]$ and the previous output of post-LSTM $o < t - 1 >$, the attention mechanism will compute attention vector or attention weights $[a < t, 1 >, a < t, 2 > \ldots a < t, x >]$ and output the context vector shown in (1)

$$context < t > = \sum_{t'}^{x} a < t, t' > p < t' > \qquad (1)$$

E. Model 2 - Bidirectional LSTM with Attention

A basic LSTM cell has memory cell which preserves information for learning Long Term Dependencies. It learns the dependencies of the inputs that are passed through it and uses that for the next output. A Uni-Directional LSTM only preserves past information where the inputs are passed forward as they come whereas the bidirectional LSTM passes the inputs in both forward and backward directions. For instance, let's say on a high level the LSTM predicts the emotion for various time steps with inputs given in forward and inputs given in backward direction as shown below,
Forward LSTM: Angry, Sad, Angry, Fear, Angry....
Backward LSTM: .... Fear, Fear, Fear, Fear, Fear, Fear, Fear.

Here one can see that the past and future information both can make it easier for model to predict the emotion for next time step.

The hidden state in bi-directional LSTM is the concatenation of hidden states from forward run and backward run of the inputs. Bi-directional LSTM generally outperforms the uni-directional LSTM and it is becoming a new standard where the input sequence doesn't matter. In this model we have changed the type from uni-directional to bi-directional in Pre-LSTM layer of model 1.

F. Model 3 - LSTM without attention

The model consists of 4 layers: Pre-LSTM (uni-directional), Dropout, Post-LSTM, Dense. The model is almost like model 1 but without the attention mechanism. We removed the attention mechanism in this model to compare the results with model 1 and check the impact attention makes on the results. The model architecture can be seen in Fig. 1 by removing the attention block. The output of Pre-LSTM is fed to the dropout layer followed by Post-LSTM. Then the final layer is dense or fully connected layer with softmax as activation function.

G. Model 4 - Bidirectional LSTM without attention

This model is same as model 3 except for the Pre-LSTM layer is replaced by Bidirectional Pre-LSTM layer whose outputs are combined and fed to next layer.

III. Evaluation and Result Analysis

We compare the results of model 1 and model 2 to analyse the attention values for unidirectional LSTM and bidirectional LSTM. For example, in a statement 'DFA' recorded by subject 1015 in 'ANG' anger emotion with unspecified emotion level, attention weights from model 1 and model 2 are shown with spectrogram and raw waveform above them. It can be observed that both the models can handle the silent areas very well and the values tend to zero. It can also be seen that Bi-Directional LSTM helps the model predict the correct emotion compared to Uni-Directional LSTM at the word 'JACKET' which spans from 30000th sample to 40000th sample which can be seen in Fig. 2. Note that the values for attention are the output of a softmax function which gives the probability for every time frame and they have been expanded to fit the samples in the raw waveform and spectrogram shown in figure 3.

We used confusion matrix to describe the performance of the classification model on test data for each model. For 91 models, 91 confusion matrices were generated, aggregated and normalized to generate one final confusion matrix which represents the overall model performance. Each item(I,j) in confusion matrix tells the number of items belonging to emotion I classified as emotion J.

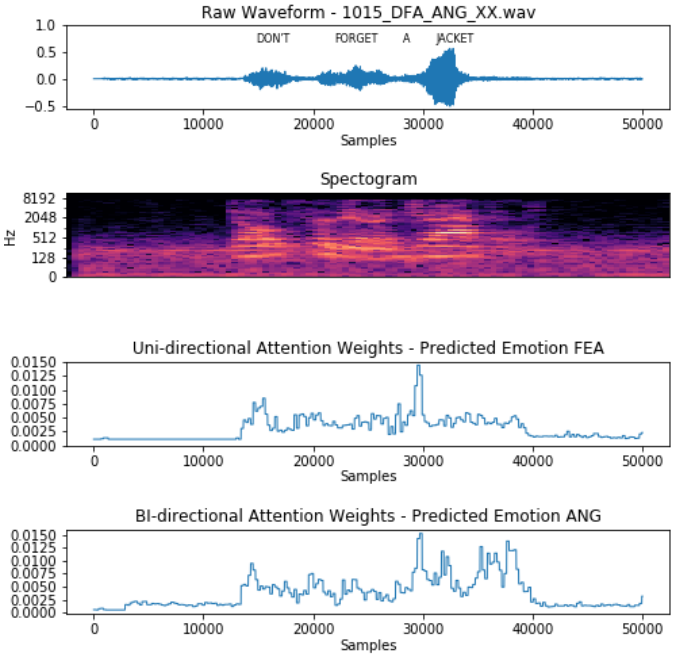

Fig. 2. For audio file 1015_DFA_ANG_XX.wav, RAW Waveform, Spectrogram, Uni-Directional Attention Weights (Model 1) and Bi-directional Attention weights in order are shown.

TABLE II
THE BEST AND WORST PERFORMING MODELS ON VARIOUS EMOTIONS SHOWING THE (%) OF TRUE POSITIVES.

Emotion	Best Classified (%)	Worst Classified (%)
Anger	Model 2 – 75.6	Model 1 – 63.25
Disgust	Model 2 – 48.46	Model 3 – 36.18
Fear	Model 2 – 48.62	Model 3 – 6.89
Happy	Model 1 – 62.7	Model 3 – 43.71
Neutral	Model 1 – 63.10	Model 3 – 52.48
Sad	Model 2 – 70.57	Model 4 – 63.25

As seen in table II, model 2 best classifies the anger, disgust, fear and sad emotions of all the models. Model 1 best performs on happy and neutral emotions and worst performs on anger emotion of all the models. Model 3 worst performs on disgust, fear, happy and neutral emotions of all the model. Model 4 worst performs on sad emotion of all the models. From this it can be deduced that model 3 is the worst performer and model 4 is the best performer. Overall performances of different models are shown in table III.

A. Uni-directional Vs Bi-directional Results

Comparing the results of model 1 and model 2, we can see that model 2 classifies every emotion better than model 1 except for "happy" emotion where model 1 has 62.6% true positives vs 55.62% of model 2. Comparing the results of model 3 and model 4, we can see that model 4 outperforms model 3 for all emotions except for emotion "sad" having 68.67% true positives vs 63.25% true positives. This shows that the inputs passed to LSTM in both the directions help the model to perform better at classifying compared to the inputs passed only in one direction.

TABLE III
OVERALL ACCURACY OF NEURAL NETWORK MODELS

Model	Accuracy (%)
Model 1 LSTM with attention	51.93
Model 2 Bidirectional LSTM with attention	**60.11**
Model 3 LSTM without attention	46.39
Model 4 Bidirectional LSTM without attention	52.91

B. Attention vs Non-Attention

Comparing the unidirectional results - model 1 and model 3, model 1 with attention performs better than model 3 for emotions: disgust, neutral, happy and significantly better in classifying fear. Comparing the bidirectional results – model 2 and model 4, model 2 with attention outperforms model 4 in every emotion. Although model 4 was able to classify the fear emotion better due to bidirectional LSTM than model 3, adding attention to it model 2 performs better than all of the models in classifying fear emotion.

Finally, out of all the four models, model 2 with Bidirectional LSTM and Attention mechanism outperforms every other model.

IV. Discussion

From the results of all the experiments on the Crema-D dataset and using the LOSO strategy and extracting 13 mfcc features for 20ms time-frame and 10ms time-step, we conclude that using a bi-directional LSTM instead of uni-directional LSTM helps the model classify the emotions better and using attention mechanism improves the performance and helps handle the noisy, silent and other non-useful parts of speech. However, increasing the number of input features along with 13 mfccs such as pitch, total energy, mean, variance, etc. may contribute to better performance but may significantly increase the training time and use far more resources in terms of gpu, memory bandwidth, CPU, storage, etc. Furthermore, multiple classifiers as seen in [30] can be combined as shown in [**?**], [31] in hierarchical, serial and parallel fashion.

V. Conclusion

In future, the more complex feature sets can be used such as LLDs and HSFs which contains hundreds and thousands of features can be used to get more better results.

References

[1] M. Schubiger, English intonation: its form and function, Niemeyer, Tubingen, Germany, 1958.
[2] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, S. Kollias, W. Fellenz, J. Taylor, Emotion recognition in human–computer interaction, IEEE Signal Process. Mag. 18 (2001) 32–80.
[3] C. Williams, K. Stevens, Vocal correlates of emotional states, Speech Evaluation in Psychiatry, Grune and Stratton, 1981, pp. 189–220.
[4] Marie Tahon and Laurence Devillers, "Towards a small set of robust acoustic features for emotion recognition: challenges," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 24, no. 1, pp. 16–28, 2016.

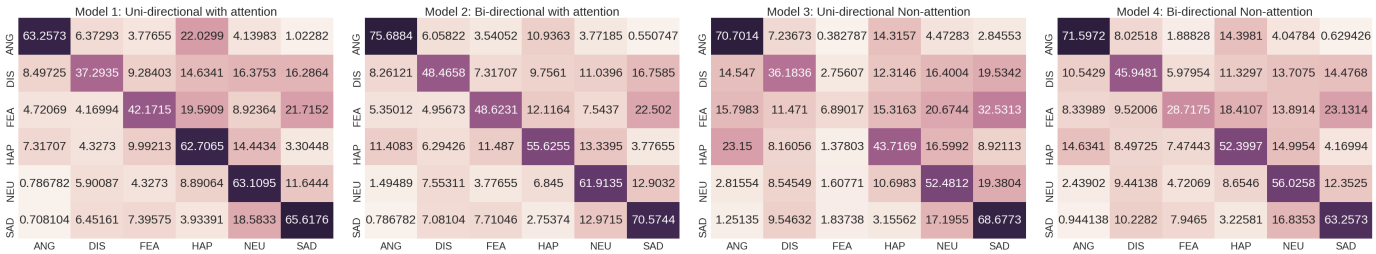

Fig. 3. Mean normalized confusion matrices of 91 subjects tested on 91 models trained using LOSO strategy. UNI: Uni-directional LSTM. BI: Bi-directional LSTM. NA: Non (without) -Attention model. AT: Attention.

[5] Bj¨orn Schuller, Anton Batliner, Dino Seppi, Stefan Steidl, Thurid Vogt, Johannes Wagner, Laurence Devillers, Laurence Vidrascu, Noam Amir, Loic Kessous, et al., "The relevance of feature type for the automatic classification of emotional user states: low level descriptors and functionals.," in INTERSPEECH, 2007, pp. 2253–2256.

[6] T. Nwe, S. Foo, L. De Silva, Speech emotion recognition using hidden Markov models, Speech Commun. 41 (2003) 603–623.

[7] D. Ververidis, C. Kotropoulos, Emotional speech classification using Gaussian mixture models and the sequential floating forward selection algorithm, in: IEEE International Conference on Multimedia and Expo, 2005. ICME 2005, July 2005, pp. 1500–1503.

[8] H. Hu, M.-X. Xu, W. Wu, Fusion of global statistical and segmental spectral features for speech emotion recognition, in: International Speech Communication Association—8th Annual Conference of the International Speech Communication Association, Interspeech 2007, vol. 2, 2007, pp. 1013–1016.

[9] M.T. Shami, M.S. Kamel, Segment-based approach to the recognition of emotions in speech, in: IEEE International Conference on Multimedia and Expo, 2005. ICME 2005, 2005, 4pp.

[10] R.W. Picard, E. Vyzas, J. Healey, Toward machine emotional intelligence: analysis of affective physiological state, IEEE Trans. Pattern Anal. Mach. Intell. 23 (10) (2001) 1175–1191.

[11] Xu, Kelvin & Ba, Jimmy & Kiros, Ryan & Cho, Kyunghyun & Courville, Aaron & Salakhutdinov, Ruslan & Zemel, Richard & Bengio, Y. (2015). Show, Attend and Tell: Neural Image Caption Generation with Visual Attention.

[12] Bahdanau, Dzmitry & Cho, Kyunghyun & Bengio, Y. (2014). Neural Machine Translation by Jointly Learning to Align and Translate. ArXiv. 1409.

[13] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, S. Kollias, W. Fellenz, J. Taylor, Emotion recognition in human–computer interaction, IEEE Signal Process. Mag. 18 (2001) 32–80.

[14] C. Busso, S. Lee, S. Narayanan, Analysis of emotionally salient aspects of fundamental frequency for emotion detection, IEEE Trans. Audio Speech Language Process. 17 (4) (2009) 582–596.

[15] L. Bosch, Emotions, speech and the asr framework, Speech Commun. 40 (2003) 213–225.

[16] Danisman T., Alpkocak A. (2008) Emotion Classification of Audio Signals Using Ensemble of Support Vector Machines. In: André E., Dybkjær L., Minker W., Neumann H., Pieraccini R., Weber M. (eds) Perception in Multimodal Dialogue Systems. PIT 2008. Lecture Notes in Computer Science, vol 5078. Springer, Berlin, Heidelberg

[17] B. Schuller, G. Rigoll, M. Lang, Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture, in: Proceedings of the ICASSP 2004, vol. 1, 2004, pp. 577–580.

[18] C. Lee, S. Yildrim, M. Bulut, A. Kazemzadeh, C. Busso, Z. Deng, S. Lee, S. Narayanan, Emotion recognition based on phoneme classes, in: Proceedings of ICSLP, 2004, pp. 2193–2196.

[19] O. Kwon, K. Chan, J. Hao, T. Lee, Emotion recognition by speech signal, in: EUROSPEECH Geneva, 2003, pp. 125–128.

[20] O. Pierre-Yves, The production and recognition of emotions in speech: features and algorithms, Int. J. Human–Computer Stud. 59 (2003) 157–183.

[21] Kun Han, Dong Yu, and Ivan Tashev, "Speech emotion recognition using deep neural network and extreme learning machine.," in Interspeech, 2014, pp. 223–227.

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

[23] Seyedmahdad Mirsamadi, Emad Barsoum and Cha Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention" in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2017.

[24] Eyben, Florian & Wöllmer, Martin & Schuller, Björn. (2010). openSMILE – The Munich Versatile and Fast Open-Source Audio Feature Extractor. MM'10 - Proceedings of the ACM Multimedia 2010 International Conference. 1459-1462. 10.1145/1873951.1874246.

[25] Huang, C., Narayanan, S.S. (2016) Attention Assisted Discovery of Sub-Utterance Structure in Speech Emotion Recognition. Proc. Interspeech 2016, 1387-1391.

[26] Chao, Linlin & Tao, Jianhua & Yang, Minghao & Li, Ya & Wen, Zhengqi. (2016). Audio Visual Emotion Recognition with Temporal Alignment and Perception Attention.

[27] Cho, Kyunghyun & van Merriënboer, Bart & Gulcehre, Caglar & Bougares, Fethi & Schwenk, Holger & Bengio, Y. (2014). Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. 10.3115/v1/D14-1179.

[28] Sutskever, Ilya & Vinyals, Oriol & V. Le, Quoc. (2014). Sequence to Sequence Learning with Neural Networks. Advances in Neural Information Processing Systems. 4.

[29] Houwei Cao, David G. Cooper, Michael K. Keutmann, Ruben C. Gur, Ani Nenkova, Ragini Verma, "CREMA-D: Crowd-sourced Emotional Multimodal Actors Dataset" in IEEE Trans Affect Comput. 2014 Oct-Dec; 5(4): 377–390

[30] M. Lugger, B. Yang, Psychological motivated multi-stage emotion classification exploiting voice quality features, in: F. Mihelic, J. Zibert (Eds.), Speech Recognition, In-Tech, 2008.

[31] Ayadi, Moataz & Kamel, Mohamed S. & Karray, Fakhri. (2011). Survey on speech emotion recognition: Features, classification schemes, and databases. Pattern Recognition. 44. 572-587. 10.1016/j.patcog.2010.09.020.

[32] L.I. Kuncheva, Combining Pattern Classifiers: Methods and Algorithms, Wiley, 2004.