Speech Emotion Recognition Using Deep Learning on audio recordings

S. Suganya[#1] and E. Y. A. Charles[#2]

[#]*Department of Computer Science, University of Jaffna, Sri Lanka*

[1]suganyasuven@gmail.com

[2]charles.ey@univ.jfn.ac.lk

Abstract*— Speech emotion recognition plays a prominent role in human-centred computing. An intelligent agent should be able to extract the context of the speech of a person including the emotion underlying in that speech. This will enable the agent to provide sensible feedback. However, automatic recognition of emotions from the speech is a challenging task due to the complexity of emotional expressions. It is still unclear that, which features of a human speech are robust enough to distinguish emotions and research in this field continues. In most of the machine learning applications, the features representing the entities in concern should be identified by an expert and then needed to be extracted. The performance of most of the Machine Learning algorithm depends on how accurately the features are identified and extracted. Deep learning has emerged as an advancement of traditional machine learning. Deep learning algorithms are capable to learn high-level features from raw data itself. Due to these improvements, recent approaches to emotion recognition utilize various deep learning models. These models were built using advanced representations or pre-processed speech recordings such as spectrogram and time-frequency representation of speech. This work proposes an end-to-end deep learning approach which applies deep neural network on a raw audio recording of speech directly to learn high-level representations from the audio waveform. The performance of the proposed model was assessed on USC-IEMOCAP and EmoDB datasets. The CNN model with nine weight layers obtained an overall accuracy of 68.6% for IEMOCAP over four emotions and 85.62% for EmoDB over seven emotions. Obtained results show that the performance of the proposed model is comparably better than the performance achieved by traditional machine learning approaches as well as deep learning on spectrogram of audio recordings in the same datasets.

***Keywords*— Keywords: Speech Emotion Recognition, Raw audio recording, Deep learning, Convolutional Neutral Network**

I. INTRODUCTION

Speech Emotion Recognition (SER) plays an important role in the field of Human-Computer Interaction (HCI) due to the essential need of understanding the emotion of the interacting human. Nowadays many useful applications require human-machine interaction e.g spoken dialogue system, intelligent voice assistants and computer games. Even though there are advances in automatic speech emotion recognition, the accuracy of such systems is still low. The emotion detection remains as a challenging one due to the various complex factors such as speaking style, gender, variations of the speaker and contents. A typical SER system consists of two core components. The first component extracts the features from the speech and the second one performs the classification of speech to predict emotions. For evaluating the performances, two measures are used: overall accuracy and class accuracy. Overall accuracy is assessed on entire test set and class accuracy is the mean of accuracies achieved for each class.

Earlier approaches on speech emotion recognition used direct voice related features. Later, a different trend of computing global features by applying statistical function to the low-level features is used. In general, speech features can be categorized into four groups as Continuous features (pitch related features, energy-related features, and articulation features), Qualitative features (voice quality features, harsh, tense), Spectral features (Mel-frequency Cepstral Coefficient (MFCC), Linear Prediction Cepstral Coefficient (LPCC), Perceptual Linear Prediction(PLP)) and Teager Energy Operator (TEO) based features (TEO-decomposed FM variation, normalized TEO autocorrelation envelope area) [1, 2]. In conventional methods, most frequently used classification techniques are Hidden Markov Model (HMM), Gaussian Mixture Model (GMM), Naive Bayes classifier and Support Vector Machine (SVM).

In recent years, Deep learning is an emerging field in machine learning and has achieved tremendous success in various domains especially in computer vision, Speech recognition and natural language processing. Speech emotion recognition systems using these deep learning architectures have shown a significant improvement over the conventional approaches.

In this study, an end-to-end deep learning model that considers voice recordings represented by time-series waveforms as input. The objective is to learn the high-level representations from the audio waveform automatically by the model and assess how well this representation distinguishes the emotions. Automated emotion recognition systems can be analysed using spontaneous speech and acted speech recordings. The proposed approach was evaluated using two data sets. First one is IEMOCAP (Interactive Emotional Motion Capture), a speech dataset released by the University of Southern California (USC) containing both improvised (spontaneous) and scripted(acted) speech sessions. The other one is the Berlin Database of Emotional Speech (EmoDB), which is a popular acted speech dataset for SER. Notably, automated emotion recognition for spontaneous speech is more complex than acted speech.

The rest of this paper is organized as follows: In the next section, relevant related works are summarised; the two data sets used for this study is introduced in section II; proposed Convolutional Neural Network (CNN) model is discussed in section III; section IV describes the experimental setup; results are presented with discussion in section V and finally conclusion is given in section VI.

978-1-7281-5156-4/19/$31.00 ©2019 IEEE

II." RELATED WORK

This section provides a summary of previous work on SER and related work. In speech emotion recognition literature, acoustic emotion features are considered as the most dominant features which distinguish speeches based on their underlying emotions. A study by Banse et. al. reported in [1], describes the emotion-specific profiles of acoustic parameters and analyses the correlation between the major groups of acoustic parameters.

Earlier studies used generative models, for instance, Gaussian Mixture models (GMMs) and Hidden Markov models (HMMs) to learn the distribution of the low-level features and then used various classifiers such as Bayesian classifiers, K-nearest neighbour (KNN) and decision trees [2]-[6]. Later, a different trend was introduced where statistical functions are applied to the low-level acoustic features to compute the global statistical features for classification for which SVM is commonly used [7]-[8]. As there are many difficulties in defining emotions as subjective and complex psychological and social phenomena, the main challenge faced by the researchers is choosing the optimal feature set for recognizing each emotion. Researchers started to use deep learning techniques for SER after the tremendous success achieved by the deep learning in various domains [9]-[12]. One of the first deep learning end-to-end approaches was proposed in which Deep Neural Network (DNN) was trained with segment level features to obtain emotion state probability distributions and used Extreme Learning Machine (ELM) for classification. This DNN-ELM model obtained 20% relative accuracy improvement compared to state-of-the-art approaches [9]. In the continuation of the DNN-ELM model proposed in [9], a Recurrent Neural Network (RNN) RNN-ELM model was proposed to account long-range context effect and obtained absolute improvements of 12% in overall accuracy and 5% in class accuracy compared to DNN-ELM baseline [11]. Another study tried a feature fusion method that combines the feature representations from both acoustic and lexical levels and achieved recognition accuracy of 62.8% on IEMOCAP dataset [10]. Performance using different types of acoustic features and different types of emotion speech (improvised / scripted) was analysed and revealed that logMel, MFCC, and eGeMAPS features achieved better performance than prosodic features [12]. For the first time, a convolutional recurrent neural network that operates on the raw signal was proposed to perform an end-to-end spontaneous emotion prediction task from speech data in [13].

Most of the recent studies presented deep learning approach using spectrograms [14]-[18]. A deep learning model was presented in [16], which used CNNs to extract high level features. To capture the temporal structure LSTMs were used in this model and achieved better accuracy than conventional classification methods for the EmoDB database [16]. Similarly, a strategy of combining a bi-directional LSTM with a novel pooling strategy was proposed in [15]. It was revealed that using deep RNNs, both frame-level characterizations, as well as temporal aggregation into longer time spans, can be learned well [15]. Furthermore, Mel-scale spectrograms were fed into two types of network topologies, convolution-only and convolution with LSTM. Experiments showed that a combination of CNN and LSTM helped in achieving better accuracy on IEMOCAP dataset [16]. Another neural network was proposed similar to the CNN-LSTM model [16] that was able to handle the variable-length speech segments [17]. In a most recent study, phoneme and spectrograms based CNN models were combined. This combined model achieved better than the state-of-the-art accuracy on the IEMOCAP [18]. A feature fusion method that combines CNN-based features and heuristic-based discriminative features based on ELM for SER is presented in [19].

It is apparent from the literature that a large number of features and methods are tried in the past for SER. Unlike these approaches of feature extraction or feature computation, the approach of this study is to apply CNN on raw waveforms as it provides a more thorough end-to-end process.

III."DATA SET

For this study, the Berlin Database of Emotional Speech (EmoDB) database [21] and the Interactive Emotional Motion Capture dataset (IEMOCAP) [22] published by the University of Southern California, are used to train and evaluate the proposed CNN model.

1) IEMOCAP: IEMOCAP corpus is a well- known dataset for emotion recognition consists of around 12 hours of speech from 10 actors including improvised and scripted dialogues. This corpus comprises of 5 sessions where each session contains the conversation between a male and a female speaker. In improvised speech, each utterance is annotated by at least three evaluators and labelled to an emotion which is agreed by at least two evaluators. Fig 1. shows the distribution of emotions of the recordings in this data set.

2) EmoDB: EmoDb is another popular dataset consist of 535 utterances in which 233 of 535 are uttered by male speakers, whereas the remaining 302 are uttered by female speakers. The database comprises of approximately 30 minutes of speech which comprised of a total of 10 German utterances, 5 short utterances and 5 longer ones. All these utterances were recorded from everyday communication. It covers seven classes of emotions namely Anger, Happiness, Sadness, Neutral, Fear, Disgust and Boredom. Fig 2. shows the distribution of emotions of the recordings in the EmoDB data set.

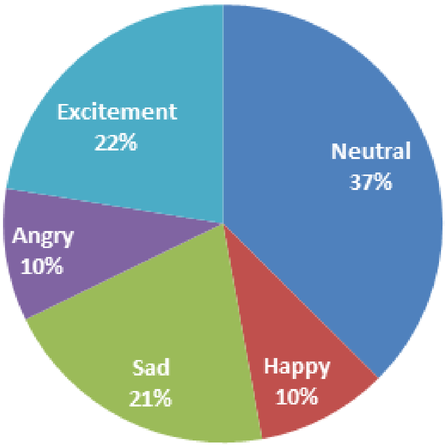


Fig.1 Emotion Distribution of IEMOCAP

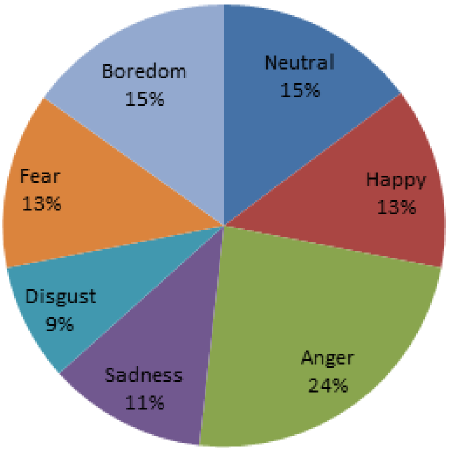


Fig. 2 Emotions distribution of EmoDB data set

IV.'PROPOSED MODEL

Dai et. al. [20] demonstrated the capability of very deep convolutional neural networks trained using raw waveforms to recognise environmental sounds. Hence, this research study tries to evaluate the possibility of the idea proposed in [20] for emotion recognition. As mentioned above, this study aims to develop a deep neural network which takes raw waveforms that represented as a long vector of values as input, instead of handcrafted features or spectrograms. Fig 3 shows a block diagram of the Proposed model.

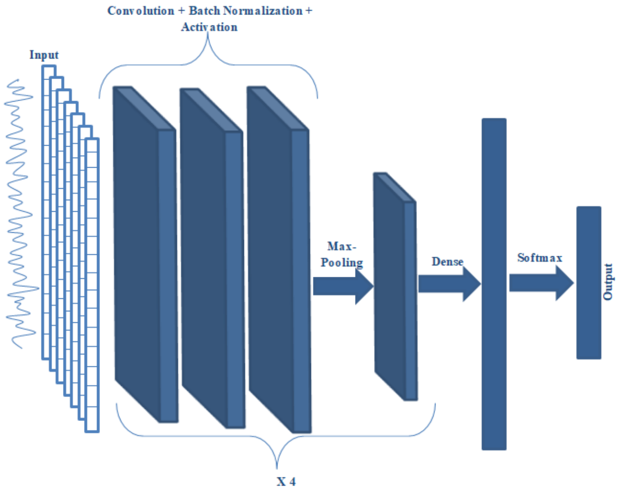


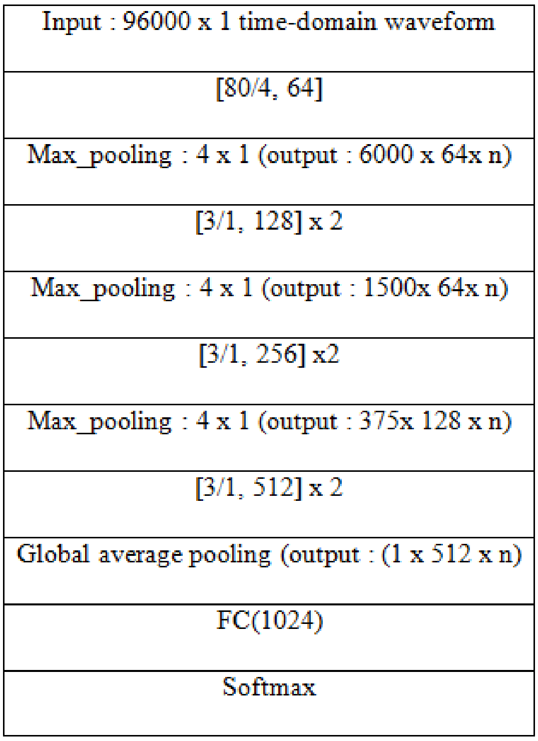
Fig. 3 Block diagram of the proposed CNN model

According to the Nyquist sampling theorem, the sampling frequency of a signal should be at least twice the highest frequency contained in the signal. The highest frequency of normal human speech is around 8 kHz, thus a sampling rate of 16 kHz is enough for any processing. Hence the speech recordings in both IEMOCAP and EmoDB datasets were sampled at 16 kHz for this study. Based on the analysis performed on the dataset the input length for the EmoDB data set was set as 6400 (16000Hz x 4s) and for the IEMOCAP as 96000(16000Hz x 6s).

Several models with different number of layers, number of hidden units were developed and analysed. The one which performed better among all attempts is selected. The proposed network topology, shown in Fig. 4, contains seven convolutional layers, one fully connected layer and a softmax layer.

Kernel size of the first layer is 80 to cover 5-milli second duration of one frame and used very small receptive field value of 3 for all other layers to enable deep learning. The role of this first layer is to extract the features from each frame and evaluate differences among the overlapping frames. Subsequent layers help to learn good representation to get better generalization. Stacked convolutional layer outputs the combinations of signal components in some frequency bands. An essential component of the CNN block is the max-pooling layer, which helps to reduce the dimensions of the feature maps. In this task, it is expected to pick the speech section which contains the information of the speech while ignoring the silent sections. Experimental results have shown that the proposed network effectively chooses the significant frames, ignoring the silent frames which do not contain any information.

Moreover, batch normalization layer is adopted after each convolutional layer and before applying the non-linearity function. This is done to reduce the problem of vanishing gradients as batch normalization layer normalizes the output of the previous layer. In addition to that, L2-Regularization is also used for all CNN layers to reduce overfitting. Stacked convolution layers are then followed by one fully connected layer of size 1024 and final softmax layer is to perform the classification task



Input : 96000 x 1 time-domain waveform
[80/4, 64]
Max_pooling : 4 x 1 (output : 6000 x 64x n)
[3/1, 128] x 2
Max_pooling : 4 x 1 (output : 1500x 64x n)
[3/1, 256] x2
Max_pooling : 4 x 1 (output : 375x 128 x n)
[3/1, 512] x 2
Global average pooling (output : (1 x 512 x n)
FC(1024)
Softmax

Fig. 4 Model Architecture ([80/4, 64] denotes a convolutional layer with 64 filters and kernel size 80 with stride 4)

V. EXPERIMENTAL SETUP

The proposed CNN model was implemented using Keras. Keras is a deep learning neural networks API which provides high-level methods for developing deep-learning models. It is a model-level open source library and is capable of running on top of Tensorflow. Implementation also used Adam, one of the most common optimization algorithm currently used. Adam is an extension to stochastic gradient descent that has recently seen broader adoption for deep learning applications in computer vision and natural language processing with good performance improvements

As the first step, all audio recordings were sampled at 16kHz. Sampled values were standardized to 0 mean and variance 1 and trimmed to fixed-length to have a length of 64000 (16000Hz x 4s) and 96000 (16000Hz x 6s) for EmoDB and IEMOCAP respectively.

Proposed CNN model was trained over 30 epochs using Adam optimizer with an initial learning rate 0.0001. Whenever the accuracy did not increase over 5 epochs, the rate was reduced by half. Besides, L2 regularization was used with coefficient 0.0001. The activation function used is Rectified Linear Unit (ReLU) and for objective function optimization ‘categorical cross-entrophy’ criterion was used.

For the model trained with IEMOCAP dataset leave-one-session-out, cross-validation method was used. In each fold, data from 4 sessions were used for training and remaining session for testing. Following the previous studies, utterances comprised of 4 classes named Neutral, Happiness, Anger, and Sadness were considered. Along with that, two different types of emotion sets were used for evaluating the performance of the proposed model. The Excitement class was included instead of class Happy in one type and in the other type only three emotion classes were included (Neutral, Anger, and Sadness). For the model trained with EmoDB dataset, seventy per cent of the data is used for training and remaining used for testing.

VI. RESULTS AND DISCUSSION

The proposed model was trained and its performance was evaluated for each data set. Various trials were conducted to analyse the performance of the model as well as to understand the correlations of emotions in the datasets. It can be noticed that there is an imbalance of classes in the datasets. Thus, the confusion matrix is presented to provide a better picture of the performance. Tables I, II and III show the confusion matrix produced by the trained model for different test sets of emotions in IEMOCAP while Table IV shows the confusion matrix for EmoDB.

TABLE I
CONFUSION MATRIX FOR EMOTION SET OF IEMOCAP

Class labels	Prediction			
	Anger	**Happiness**	**Neutral**	**Sadness**
Anger	59.2	3.1	36.0	1.7
Happiness	11.2	14.3	69.2	5.2
Neutral	4.7	4.1	79.9	11.3
Sadness	1.8	1.6	20.2	76.4

Table I shows the confusion matrix for the emotion set containing four emotions, namely, Anger, Happiness, Neutral and Sadness. Proposed model achieved an overall accuracy of 68.6% with this emotion set. From the confusion matrix [Table I], it can be observed that the accuracy for Neutral and Sadness classes are high compared to other classes whereas class Happiness recorded low classification accuracy. On the other hand, Happiness is heavily misclassified as Neutral. Although, Happiness performed poorly, the proposed model was able to distinguish other emotions well for this emotion set.

In some previous studies, Excitement was also considered as Happy. Since classification accuracy for Happiness class was very low, another experiment was carried out where the Excitement was considered as one of the emotions instead of Happy, without considering both emotions as Happy. For this arrangement, the proposed model obtained an overall accuracy of 64.3% and class accuracy of 59.1%.

From Table II, it can be observed that the Neutral and Sadness classes showed better performance than other classes while classes Anger and Excitement were mostly misclassified as Neutral. Besides, it can be noticed that class accuracies of all emotions, especially Anger are reduced for this emotion set compared to the previous one. Further, Anger and Excitement classes also confused with each other.

TABLE II
CONFUSION MATRIX FOR EMOTION SET OF IEMOCAP

Class labels	Prediction			
	Anger	**Excitement**	**Neutral**	**Sadness**
Anger	40.1	24.9	34.3	0.7
Excitement	11.2	45.8	39.8	3.2
Neutral	2.4	11.0	75.8	10.8
Sadness	0.9	1.6	22.8	74.7

Furthermore, another experiment was conducted with three emotions for which the proposed model showed high class accuracy as well as showed the correlation between audio recordings on emotions. When considered Anger, Neutral and Sadness classes, the proposed model recorded an overall accuracy of 79.3% and class accuracy of 75.3%. Confusion matrix for this experiment is shown in Table III. According to the results, it can be concluded that the correlation between these three emotions are less compared to Happiness and Excitement emotion classes.

TABLE III
CONFUSION MATRIX FOR EMOTION SET OF IEMOCAP INCLUDES ANGER, HAPPINESS, NEUTRAL AND SADNESS

Class labels	Prediction		
	Anger	**Neutral**	**Sadness**
Anger	62.6	35.3	2.1
Neutral	3.7	82.2	14.1
Sadness	1.8	17.1	81.1

Recognizing emotions is not a straight-forward task as even for human it is not easy to define and distinguish emotions. In a recent study, human evaluation for the emotion recognition using this IEMOCAP dataset was investigated and found out that the assigned labels for the utterances are inconsistent with the experts [23]. Moreover, the study revealed that both overall accuracy and class accuracy are about 70% and it will be very difficult for any model to achieve even this accuracy. Thus, it can be concluded that the proposed model achieved a commendable performance in comparison to the state-of-the-art model as well as in view of the findings reported in [23].

As mentioned before, EmoDB was also used to assess the performance of the proposed model and for which the proposed model demonstrated an overall accuracy of 85.62%. From Table IV, it can be observed that the model was able to perform with high class accuracy for all emotions except happiness. For emotion class Sadness, the model achieved 100% accuracy. Class Happiness was heavily confused with anger emotion. 42.9% percentage of data items belonging to Happiness were classified as Anger. Notably, as both happiness and anger emotions have high values of energy and arousal.

Sadness emotion is better perceived from audio while the emotions like Anger and Excitement can be better recognized from video [22]. From the obtained results also, Sadness and Neutral showed high true positive whilst Sadness got a better precision than other emotions. Specifically, Sadness achieved 100% class accuracy with high precision for EmoDB.

TABLE IV

CONFUSION MATRIX FOR EMODB

Class label	**Prediction**						
	Neutral	**Happiness**	**Anger**	**Sadness**	**Fear**	**Disgust**	**Boredom**
Neutral	91.7	0.0	4.2	4.2	0.0	0.0	0.0
Happiness	4.8	47.6	42.9	0.0	4.8	0.0	0.0
Anger	2.6	5.3	92.1	0.0	0.0	0.0	0.0
Sadness	0.0	0.0	0.0	100	0.0	0.0	0.0
Fear	0.0	5.3	0.0	0.0	89.5	5.3	0.0
Disgust	0.0	7.1	7.1	0.0	0.0	78.6	7.1
Boredom	4.0	0.0	0.0	0.0	4.0	0.0	92

VII." CONCLUSIONS

This research study proposed a deep learning model for speech emotion detection which can be trained and used using the waveform of audio recordings directly. The proposed model was evaluated using two datasets, IEMOCAP and EmoDB. Experimental results showed that the proposed model was able to achieve an overall accuracy of 68.6% for IEMOCAP and 85.62% for EmoDB. Evaluations revealed that the obtained results are close to the accuracy of the CNNs with spectrogram features. Although class accuracy for happiness was low, other classes showed better performance in IEMOCAP dataset. With the EmoDB, the proposed model was able to achieve good results with a mean class accuracy of 84.5%. Hence, it can be concluded that the proposed approach is highly feasible for the emotion recognition task. Since it is still unclear about the features which are good enough to distinguish emotions, deep neural network is a better choice to learn and identify optimal features for speech emotion recognition. The proposed approach can be further evaluated with other emotion datasets. In addition, the proposed model can be improved further by using large datasets to recognize all emotions effectively.

REFERENCES

[1]" R. Banse and K. R. Scherer, "Acoustic profiles in vocal emotion expression," *Journal of personality and social psychology*, vol. 70, no.3, pp. 614–634, 1996.

[2]" M. B. Mustafa, A. M. Yusoof, Z. M. Don, and M. Malekzadeh, "Speech emotion recognition research: An analysis of research focus," *International Journal of Speech Technology*, vol. 21, pp. 137–156, 2018.

[3]" B. Schuller, G. Rigoll, and M. Lang, "Hidden markov model-based speech emotion recognition," *2003 IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP)*, vol. 2, pp. II–1, 2003.

[4]" H. Hu, M.-X. Xu, and W. Wu, "GMM supervector based SVM with spectral features for speech emotion recognition," *2007 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 4, pp. IV–413, 2007.

[5]" E. Lee C. and Mower, Busso, S. Lee, and S. Narayanan, "Emotion recognition using a hierarchical binary decision tree approach," *INTERSPEECH*, vol. 4, pp. 320–323, 2009.

[6]" Y. Kim and E. Mower Provost, "Emotion classification via utterance level dynamics: A pattern-based approach to characterizing affective expressions," *2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* 2013.

[7]" F. Eyben, M. Wollmer, and B. Schuller, "Openear - introducing the munich open-source emotion and affect recognition toolkit," *2009 3rd International Conference on Affective Computing and Intelligent Interaction (ACII)* , pp. 1–6, 2013.

[8]" E. Mower, M. J. Mataric, and S. Narayanan, "A framework for automatic human emotion classification using emotion profiles," *IEEE Transactions on Audio Speech and Lang*uage *Processing*, vol. 19, no.5, pp. 1057–1070, 2011.

[9]" K. Han, D. Yu, and I. Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," *INTERSPEECH*, pp. 223–227, 2014.

[10]" Q. Jin, C. Li, S. Chen, and H. Wu, "Speech emotion recognition with acoustic and lexical features," *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4749–4753, 2015.

[11]" J. Lee and I. Tashev, "High-level feature representation using recurrent neural network for speech emotion recognition," *INTERSPEECH*, pp. 223–227, 2015.

[12]" 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," *INTERSPEECH*, 1263-1267, 2017.

[13]" G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, S. Zafeiriou, and B. Schuller, "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network," *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5200–5204, 2015.

[14]" W. Lim, D. Jang, and T. Lee, "Speech emotion recognition using convolutional and recurrent neural networks," *2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference(APSIPA)* , pp. 1–4, 2016.

[15]" S. Mirsamadi, E. Barsoum, and C. Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2227–2231, 2017.

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

[17]" X. Ma, Z. Wu, J. Jia, M. Xu, H. Meng, and L. Cai, "Emotion recognition from variable-length speech segments using deep learning on spectrograms," *INTERSPEECH*, pp. 3683–3687, 2018.

[18]" P. Yenigalla, A. Kumar, S. Tripathi, C. Singh, S. Kar, and P. Vepa, "Speech emotion recognition using spectrogram phoneme embedding," *INTERSPEECH*, pp. 3688–3692, 2018.

[19]" L. Guo, L. Wang, J. Dang, L. Zhang, and H. Guan, "A feature fusion method based on extreme learning machine for speech emotion recognition," *2018 IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP)*, 2666-2670, 2017.

[20]" W. Dai, C. Dai, S. Qu, J. Li, and S. Das, "Very deep convolutional neural networks for raw waveforms," *2017 IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP)*, pp. 421–425, 2017.

[21]" F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A database of German emotional speech," *INTERSPEECH*, pp.1517–1520, 2005.

[22]" C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

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