Emotion Speech Recognition using Deep Learning

Othman O. Khalifa[1*], M. I. Alhamada, Aisha H. Abdalla
1- International Islamic University Malaysia, Electrical and Computer Engineering, Malaysia.
Email: khalifa@iium.edu.my (Corresponding author)

Received: April 2020 Revised: June 2020 Accepted: August 2020

ABSTRACT:
Emotion Speech Recognition (ESR) is recognizing the formation and change of speaker's emotional state from his/her speech signal. The main purpose of this field is to produce a convenient system that is able to effortlessly communicate and interact with humans. The reliability of the current speech emotion recognition systems is far from being achieved. However, this is a challenging task due to the gap between acoustic features and human emotions, which relies strongly on the discriminative acoustic features extracted for a given recognition task. Deep learning techniques have been recently proposed as an alternative to traditional techniques in ESR. In this paper, an overview of Deep Learning techniques that could be used in Emotional Speech recognition is presented. Different extracted features like MFCC as well as feature classifications methods including HMM, GMM, LTSTM and ANN have been discussed. In addition, the review covers databases used, emotions extracted, and contributions made toward ESR.

KEYWORDS: Speech Emotion Recognition, Deep Learning, Deep Neural Network, Deep Boltzmann Machine, Recurrent Neural Network, Deep Belief Network, Convolutional Neural Network.

1. INTRODUCTION

In recent years, more researches are being conducted to understand how human brain is built and how to build human brain-like systems that are comparable or more advance than human brain. Human brain has been the main inspiration in the making of most machines learning systems. Human brain contains neural networks that are capable of extracting high-level concepts by processing a very low-level data. Researchers have always aimed to create artificial intelligence systems that are able to think rationally and reach a level where human will be unable to differentiate between machine and human. Speech recognition is an essential part in developing machines that can naturally interact with human. That is, speech is the easiest tool in human communication. In early speech recognition system, the main focus was to extract the linguistic information rather than extracting paralinguistic information like emotion for speech. However, speech contains much more than just the language, a lot more information can be extracted from speech for example, gender, mood, identity and the emotional state.

Emotional information is traditionally extracted based on the relationship between the emotions and vocal features. Meaning, emotions were extracted using the speech signals features such as, energy, duration, fundamental frequency, and timber. Hidden Markova model and artificial neural networks are among the traditional models that use acoustic correlates off the speech signal to classify emotions [1],[2],[3],[4]. However, there are no agreement that specify the acoustic correlates of the all speech signals, this is because of the reason that there are many factors that affect the acoustic correlates of the emotions, for instance gender, personal cultural, spoken language. Moreover, the use of deep learning models is a possible solution of this problem. The study investigates the result of using one of the deep learning models which is Convolutional Neural Network (CNN) in classifying emotions from speech signals [5],[6],[7],[8],[9] However, the result obtained from the live demo are not highly accurate and that is because of a speech database used to train the CNN model, in which actors may exaggerate certain emotions.

2. EMOTION TAXONOMY

Most of human conversation and interactions highly depend on the emotional state of the speaker and the listener. Human emotions are usually hard to detect and that is because emotions are subjective, personal and cultural differences play an important role in defining and classifying emotions. Furthermore, Human emotions can be expressed in many ways, for instance facial expression, pitch characteristics, and or the actual language and the vocabularies used in conversations[10]. Unfortunately, in psychology there

Paper type: Research paper
DOI: https://doi.org/10.29252/mjee.14.4.39
How to cite this paper: O. O. Khalifa, M. I. Alhamada and A. H. Abdalla, **"Emotion Speech Recognition using Deep Learning"**, *Majlesi Journal of Electrical Engineering*, Vol.14, No. 4, pp. 39-55, 2020.

are no general scientific methods used to detect emotions, resulting emotion detection to be a challenging task for researches. Moreover, human performs emotion recognition automatically and subconsciously, they also use emotions to achieve better communications and interactions between one another. Nonetheless there are three noteworthy methodologies used to evaluate human emotions, which are appraisal-based, categorical and continues [11]. A famous emotional model is the circumflex model of affect [12]. The model describes emotions as a set of independent dimensions, such as excitement and pleasure. There also some emotional classification models with Robert Plutchik model being the most famous out of them. In his model, he divided emotion into 8 fundamental emotions showed in Table 1. They are grouped emotions into two groups of opposite emotions. Different emotions result in different intensity and combination of them may result in a secondary emotion.

Table. 1 The fundamental emotions by Robert Plutchik [13].

Joy	Sadness
Trust	Disgust
Fear	Anger
Anticipation	Surprise

Robert Plutchik graphically represented his model on a color wheel known as ’Plutchik’s wheel’ shown in Fig. 1. Plutchik’s wheel shows that there are 8 different fundamental emotions which are joy, trust, fear, anticipation, sadness, disgust, anger and surprise. These fundamental emotions are connected to a large number of emotions and the center of the wheel represents the highest level of intensity of these fundamental emotions

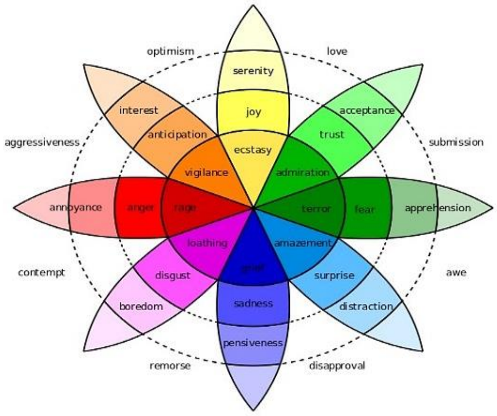


Fig. 1. Plutchik’s Wheel of Emotions.

3. SPEECH EMOTION RECOGNITION

In recent years speech emotion recognition has drawn the attention of researchers because of the increasing Human-computer interaction, SER has wide potential application such as call centers video games or automated cars[13]. Speech emotion recognitions system is a system that takes a raw audio waveform input, processes it and outputs one of the emotional states categories feed to the classification system as seen in Fig. 2. SER consists of two main stages which are : features extraction and feature/machine classification [14] [15] [16][17][18].

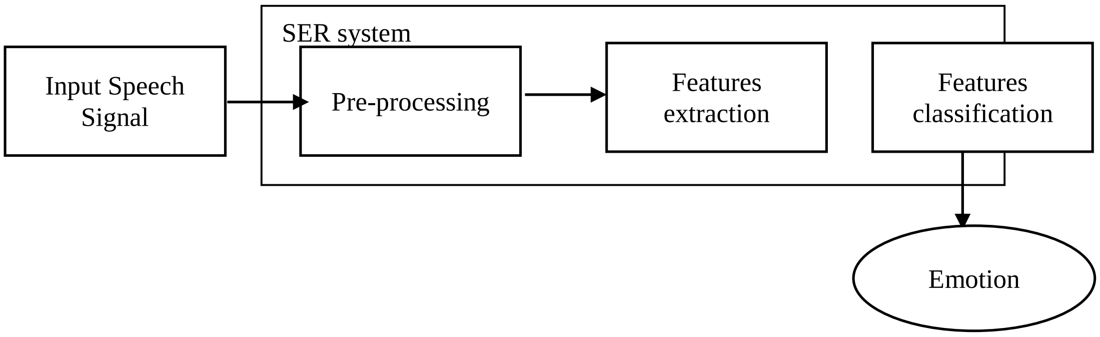


Fig. 2. Speech emotion recognition system block diagram.

3.1. Feature Extraction

Human emotions can be classified into many categories for instance happiness, anger, natural etc. For researchers to be able to accurately classify human emotion, the need of extracting specific human sound features becomes necessarily. Moreover, the most important stage in speech recognition is feature extraction. Many methods can be used in feature extractions for example Pitch, and MFCC [19]. In this section, we will briefly discuss some of the mentioned feature extraction methods.

3.1.1. Pitch Method

Pitch is the most important property of speech. It is a known fact that speech is nothing but a wave that is generated by vibrating objects in a medium such as air. Speech emotion recognition's researches have always considered the characteristics of pitch in order to have an accurate emotion classification. Many methods were invented to extract features from pitch for example the Cepstral method. In Cepstral method, the signal goes into several stages. First, the analog signal will be sampled and quantized to digitize the signal. The digital signal is then framed in a suitable size and this can be obtained by passing the signal into a hamming window and applying FFT (Fast Fourier Transform), the signal is then converted into a frequency domain. The next step is obtaining the absolute values and the signal logarithm. By applying the Inverse Fast Fourier Transform, the signal is finally transformed to the Cepstral domain where the pitch frequency is represented by the peak signal[20] as shown in Fig. 3.

3.1.2 Mel Frequency Cepstral Coefficients (Mfcc)

Mel frequency cepstral coefficients is a method that was invented in 1980 by Mermelstein and Davis. Implementing MFCC Requires following few steps shown in Fig. 4, which are Pre-emphasis, framing, Fast Fourier Transform, Mel Filter bank, computing DCT [19].

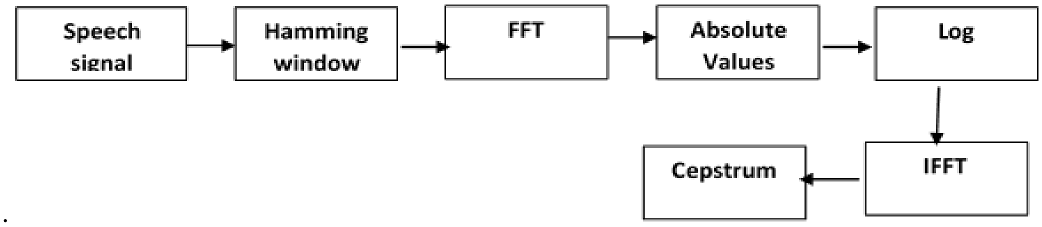


Fig. 3. Stages of Cepstral method [20].

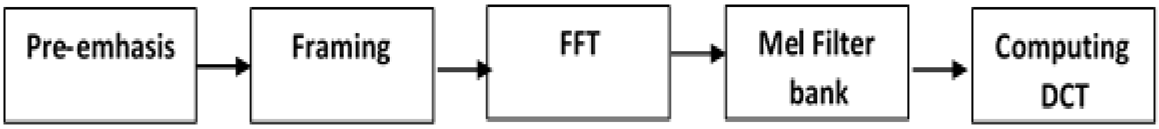


Fig. 4. Stages of MFCC [19].

Pre-emphasis: Pre-emphasizing includes filtering the sample word to increase the energy of the signal at higher frequencies. Frame blocking: this step involves segmenting the signal into small durations (frames) ranging between 20ms to 40ms. The speech is signal divided into N samples, the adjacent frames will be set rates by M(M<N). M and N values are 100 and 256 respectively. Fast Fourier Transform: FFT is converting signals from time domain into frequency domain. Mel Filter bank: voices signal in FFT uses triangular band pass Filter. To obtain smooth magnitude spectrum, and reduce the size of the involved features, a set of triangular filters are multiplied by the magnitude frequency response.

3.2. Features Classification

The second and final stage of speech emotion recognition is Feature classification, what comes after this step is the final categorized emotion. Various classification methods were used to categorize utterance-level features, some example of a famous classification methods are: neural network, hidden Markov models (HMM), and (GMM) Gaussian mixture model. However conventional classifying methods like HMM and GMM may not give a high and accurate classifying rate compared to deep learning and neural network methods. Moreover hybrid methods [21] may give better results compared to individual methods.

3.2.2 Hidden Markov Models (HMM)

HMM methods are contributed in many speech emotion recognition researches [22], because speech signals can be described as a transmission of states in time. It is essentially the base of all modern speech recognition methods; it was later replaced by RNN due to HMM being unable to handle non-linearity application. HMM is a statistical model that predicts series/sequence of events. It consists of a first order Markov chain with hidden state. Meaning that the observer will not be able to observe the states but the output is visible and depends on the states. The hidden states of HMM captures the structure of data.

3.2.3. Gaussian Mixture Model (GMM)

GMM is a probabilistic model and it is used for cluster observations, estimate densities or to specify a generative model. GMM is commonly used, as a classification tool in speech emotion recognition. In speech emotion application, GMM is implanted by firstly modeling the probability density function using multivariate Gaussian mixture model. By inputting data, GMM assigns a weight to each Gaussian distribution using expectation maximization algorithm.

4. DEEP LEARNING AND NEURAL NETWORK

Conventional neural networks are constructed using simple and connected neurons also known as processors, theses neurons produce a sequence of real-value activation, some input neurons get activated through an environmental sensing sensors while other get their activation through assigned weight. Fig. 5 shows a simple neural network with neuron inputs X_i,, the weight W_{ij} and the neuron which sums the weights multiplied by the input.

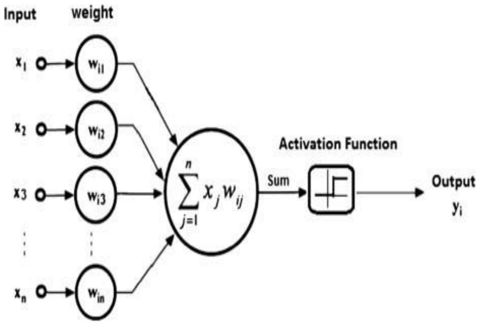


Fig. 5. Conventional neural networks [23].

However, deep learning in neural network uses more than just the simplified version of the neural network. Neural network can be defined as a an approach of composing networks into multiplexed layer of processing with the aim of learning multiple level of abstractions [16]. However, the Deep Learning (DL) concept appeared for the first time in 2006 as a new field of research within machine learning. It was first known as hierarchical learning at the [10], and it usually is involved in many research fields related to pattern recognition. Deep learning mainly considers two key factors: nonlinear processing in multiple layers or stages and supervised or unsupervised learning [12]. Nonlinear processing in multiple layers refers to an algorithm where the current layer takes the output of the previous layer as an input. Hierarchy is established among layers to organize the importance of the data to be considered as useful or not. On the other hand, supervised and unsupervised learning is related with the class target label, its availability means a supervised system, whereas its absence means an unsupervised system. Furthermore, Deep Neural Network is a feed-forward fully-connected multi-layer neural network. The different Deep learning architectures as well as their limitations are explained in the following section.

4.1. Convolutional Neural Networks

A Convolutional Neural Network (NNC) is a system of feeding forward neural networks in machine learning, where this system has a collection of small neurons in multiple layers which are called receptive field. Their main purpose is to process the input image in portions. Moreover, this process is repeated throughout all the network layers. A question to be raised is where CNNs is used? CNNs in most cases are used for natural language processing as well as video image processing. The reason to why the researchers could make the network with raw frames to extend the CNNS for videos, is that the ability of video to be separated into temporal components as well as spatial components. The temporal parts as movement (optical ow) over the frames capture the movement of the objects. The spatial parts in the type of frames capture the appearance data such as the objects present in the video [24]. The strengths of CNNs model can be seen when the use of shared weights in convolutional layers makes it ready to utilize a similar channel for every pixel in the layer. Furthermore, CNNs do not pre-processing technique a lot which implies that the CNN network oversees learning the channels where the classical calculations are hand-built. Thirdly, CNNs are anything but difficult to prepare and are less reliant on the human comprehension and exertion and as well on the past information in planning the highlights of the model. The main strength of CNNs is that when compared with the connected network, it has less parameters with a similar number of concealed units. The most distinctive element of the CNN is that it has the 3D volume of neurons in which the neurons are masterminded in three dimensions weight, height, and depth. One of the prominent limitations of CNNs is that it requires an enormous demand of memory requirement to hold all intermediate results of the convolutional layer for giving as input to the back-propagation layer [25]. Fig. 6 shows the layers of CNNs.

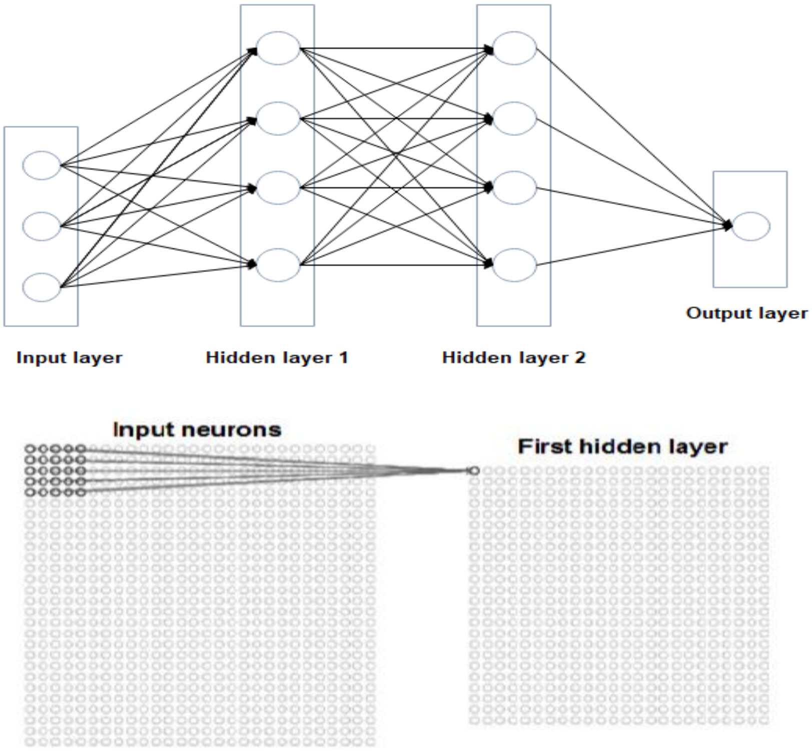


Fig. 6. Illustration of a full-connected model in an ordinary 3-layer convolutional neural system and an illustration of the local responsive.

4.2. Recurrent Neural Networks

CNNs model is unsuitable for learning sequences because learning patterns need a feedback mechanism as well as a memory of previous states. However, the Recurrent Neural Networks (RNNs) are neural nets which consist of at least one feedback connection. The RNNs is a stochastic multilayer model that is used in the past examinations to perceive objects in video scene, music, content and motion capture, this repetitive structure allows Recurrent neutral networks to learn temporal patterner and to have an internal memory [26]. The primary distinction between a RNNS and a multilayer perceptron is the nearness of recurrent associations. Along these lines, a RNNS can figure out how to delineate from the whole history of past contributions to each yield. However, they are exceptionally hard to prepare due to the vanishing slope issue. The Long Momentary Memory (LSTM) approach has been proposed to tackle these issues. RNNs needs training on datasets so that it generates new sequence by prediction. Since RNNs is unable to store the past input for so long, it avoids long sequences which makes one of the most optimized models. RNNs uses prediction from few inputs, though it might be needed for higher size of inputs, adding noise will be the next step so that RNNs can go all the way to past and do the process of learning. The architecture of RNNs model is long short-term memory (LSTM) [27]. Fig. 7 is an illustration of Multilayer perceptron.

4.3. Deep Belief Networks

Deep Belief Networks (DBNs) have been utilized effectively over the last decade for some recognition tasks, for example, written by hand digits recognition, object recognition or modeling human movement. Its networks are highly complex directed acyclic graph. RBM architectures are a sequence of restricted Boltzmann Machine as shown in Fig. 8. The main alteration between DBN and RBM model is that the top two layers in these networks are undirected whilst the lower layers are directed. RBMs consist of two layers one is called "*Visible*" which shows the input data. The other layer is called *"hidden"* where it learns how to represent features. All the hidden units which are under the hidden layer are connected to all visible units which are under visible layer. One of the strengths of DBNs is the capability to learn optimum set of parameters rapidly. DBNs utilize an unsupervised pre training procedure even for very large unlabeled databases. DBNs could likewise figure the yield estimations of the factors in the bottom layer utilizing rough derivation methodology. The impediments of DBNs incorporate the confinement of the rough derivation methodology to a single bottom-up pass [28]. Fig. 8 illustrates the connection between the layers and the visibility and the hiddenness. It is interesting to note that in DBNs, there is no connection between units if they are from the same layer. The greedy strategy adapts just the highlights of one layer at any given moment and it never straightens out with alternate layers or parameters of the system. The wake-sleep algorithm is commonly viewed as exceptionally moderate and inefficient however it tweaks all-inclusive [29].

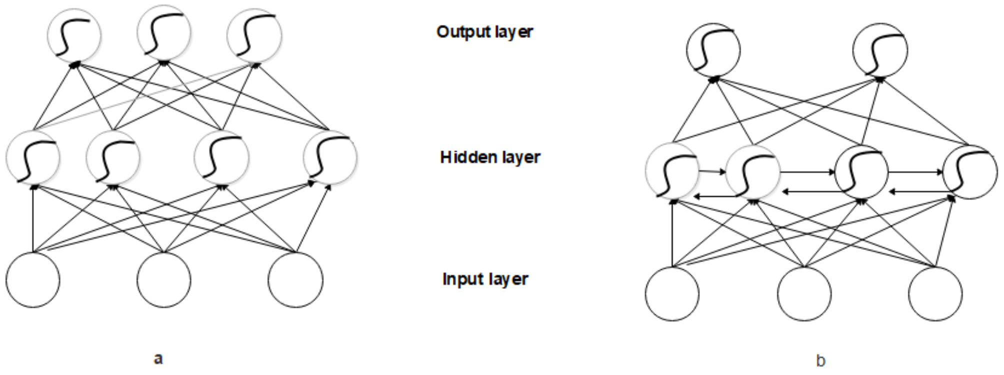


Fig. 7. Illustration of: (a) a Multilayer Perceptron and (b) a Recurrent Neural Network.

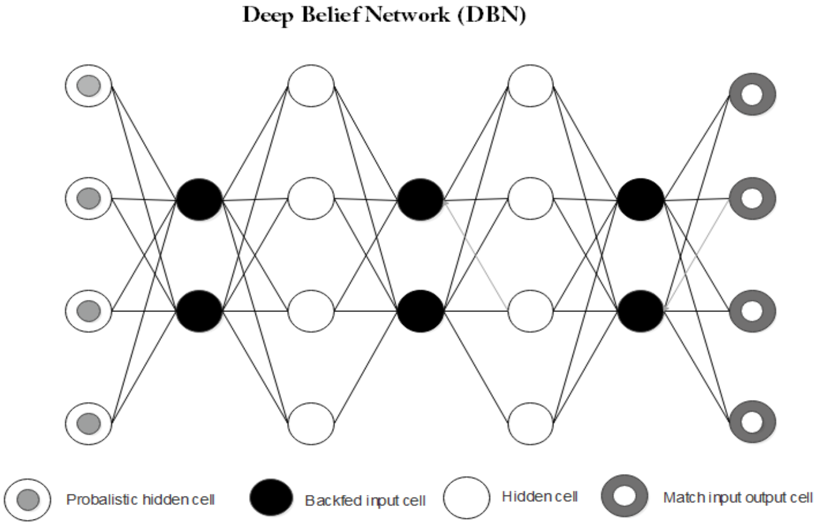


Fig. 8. The connection between the layers.

4.4. Deep Boltzmannn Machine

Deep Boltzmann Machine (DBMs) is considered as one of the most famous models when it comes to deep algorithms which contains many hidden layers. One of the promising capabilities of DMB's is solving object recognition as well as speech recognition due to its ability of learning internal representation that becomes increasingly complex, which is considered one of its strengths. Moreover, what can be built from a large supply of unlabeled sensory inputs and very limited labeled data is the high-level representation and what is got can then tweak all-inclusive the model for a pre-determined assignment for it [30]. To compare DBNs to DBMs, here in DBMs, top down algorithm is used for feedback and also it uses bottom up to forward data, permitting profound Boltzmann machines to more readily spread unsure about, furthermore, henceforth deal more heartily with uncertain inputs as shown in Fig. 9. One of DBMs limitations is the time required for exact stochastic settling and also examining is restrictive in generally big and complexed network [31],[32].

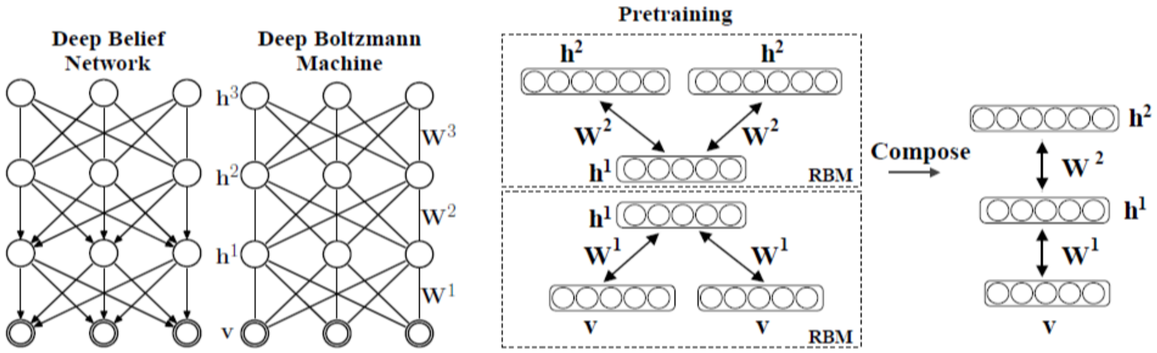


Fig. 9. Left: A three-layer Deep Belief Network and a three-layer Deep Boltzmann Machine. Right: Pretraining comprises of learning a heap of adjusted RBM's, that are then formed to make a profound Boltzmann machine [30].

4.5. Stacked Denoising Autoencoders

Stacked Denoising Autoencoders (SDAs) is one of DL architectures that was first introduced in 2008 by Vincet et al [33],[34]. SDAs is an extension of a classical autoencoder as shown in Fig. 10. What makes SDAs a *"deep"* architecture is that it is constructed by stacking multiple autoencoders together. The unsupervised pre training of each autoencoder is performed in a greedy layer by layer method. Once a SDAs is learnt, its yield will be utilized as the input portrayal of a regulated learning calculation for recognition tasks. So, the main idea is that the created or reconstructed input will have as minimum as possible of errors, so the input value closes to the real value. Notedly, in this method the output, which is in fact corrupted, comes out clean. The idea of an autoencoder can be quickly portrayed here: Given a set of data points $x = \{x_1,\ x_1\ \dots.,x_m\}$, map x to another set of data points $Y = \{Y_{1,}\ ,\{Y_{2,},\dots.,\{Y_{m,}\}$ where $n < m$. From the compacted set y, reproducing a lot of ~x, which approximates the first information x. The mapping x →y is called "*encoding*" and the mapping y → ~x is called *"decoding"*.

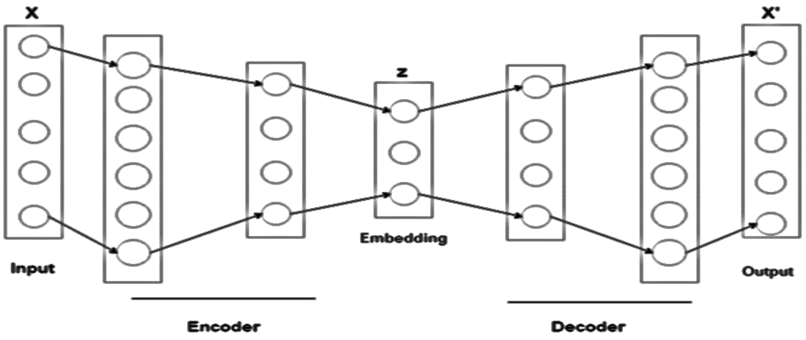


Fig. 10. Autoencoder of the SDAs model.

5. DEEP LEARNING TECHNIQUES FOR SPEECH EMOTION RECOGNITION

Recently Speech Emotion Recognition (SER) has caught the attention of many rehearses, the reason is the increasing interactions between human and machine. SER system is a system that takes a raw audio waveform input, processes it and outputs one of the emotional states categories feed to the classification system. Number of studies have discussed deep learning in speech emotion recognition systems, convolution networks as well as DNN and RNN have been used in these studies and the results vary between them. In 2017 alone more than 150 papers were published discussing topics related to SER.[1-7] investigated different used emotion recognition systems and proposed a system using Deep Feedforward Neural Network (DNN) with MFFC as a subject of feature extraction. The authors utilized the network to come up with the best recognition rate compared to other studies, the utilized system performance was relatively high with a recognition rate of 97.1%. However, it did not perform well for the custom database with a recognition rate of 27% only. The author explained the result obtained in the customer database to be a result of the language difference and the type of compression. Authors in [7] used Convolution recurrent neural network stacked on top of 2-layers of LTSM long short-term memory and concluded that their model has clearly outperformed other models in arousal and valence. Other studies preferred to use RNNs and were able to bypass what other studies could not, which is making the model be able to ignore the silence frames of the waveform by assigning very small weight to them which causes them to be ignored in the pooling operation [10]. In [35], a convolutional neural network is proposed as a deep learning model, his design consisted one convolutional layer followed by max pooling layer, his method showed a relatively higher performance than LDA yet a lower performance than other recognition methods namely RDA, KDA, SVM, KNN and CNN [21]. The two methods were later combined. The first used method was GMM-HMM were the system inputs are the MFCC features of the audio signal. The second method was the conventional SVM classifier with LLD (Low Lever Descriptors) as the input of the model. The research proved that combining these two methods that is a new classifier was able to outperform the traditional SVM, HMM and GMM-based models. However, the study did not provide the exact recognition rate obtained in the experiment. Another back draw of the method used is that it cannot handle big range gap in features dimensions. Authors in [20] used Deep Neural Networks (DNN) and Extreme Learning machine (ELM) in their system. The research provided a comparison between the methods used and other conventional methods like the HMM in term of weighted accuracy (classification accuracy of the whole test samples) and unweighted accuracy (classification accuracy for a specific emotion).The model proposed outperformed the conventional methods by 20% for both weighted and unweighted accuracy. Table 2 below provides a summary of some related work. Despite the facts that many researches discussed SER-related topics, there is still a room for improvement.

Table 2. A summary of related work.

Author	Methodology	Strength	Limitations
M.F. Alghifari, T.S. Gunawan, and M. Kartiwi [34]	Classifier: MFCC-DNN Database:(Emo-DB) – Custom database. Emotions classified: happy, angry, sad and neutral.	Achieved 97.1% recognition rate, outperforming some other models	The model Performed poorly in a custom database with a recognition rate of 23.3% only
P. Tzirakis, J. Zhang, and W. Schuller [36]	Classifier : (CNN) -(LSTM) Database: RECOLA database. Emotions classified: No specific emotions.	The model outperforms deep convolutional recurrent network in both the arousal and valence dimensions for both the validation and test sets.	Lacks the information about the number and the types of classified emotion. Valence percentage can be improved when using DNN.
J. Engelbart [16]	Classifier : CNN Database: IEMOCAP - MSP-IMPROV databases. Emotions classified : Neutral, happy, sadness and anger.	Well explained methodology. Outperformed the state of art approaches "mostly used RNN" in term of unweighted accuracy.	The proposed model is Slower than RNN-based approaches in term of Real-time emotion classification.
X. Ke, Y. Zhu, L. Wen, and W. Zhang [35]	Classifier: SVM-GMM-HMM Database: IS2009 database. Emotion classified: Anger (angry, touchy, and reprimanding), Emphatic, Neutral Positive (joyful).	Performed better than conventional HMM-GMM and the SVM models.	Classification task was exacerbated because of the big gap in range of the features dimensions.
	Classifier : RNN Database : IEMOCAP database Emotions classified : happy, sad, neutral, angry.	Outperformed some other training methods and a achieved a 5.7% absolute improvement.	Joint learning provided low performance. Need more sufficient training samples.
W. Lim, D. Jang, and T. Lee [5]	Classifier: CNN-RNN-LTSM Database: German Berlin database Emotions classified: Neutral, Anger, Fear, Disgust, Sadness, Boredom, Happy.	High recognition performance using CNN and time distributed CNN. better results than conventional methods.	Poor organization of the paper. The proposed LTSM model performed Relatively Low compared to CNN.

6. PROPOSED METHODOLOGY

This section explains the proposed methodology, emotion database used for research and the inception model. We present a CNN-based framework for SER. The main objectives of the proposed framework are to implement a reliable and accurate emotion speech recognition system that computes and outputs one detected emotion. To achieve that the proposed architecture takes into consideration three main steps: first step is the itch preprocessing, this step aims to achieve a consistent sampling rate, second step is Training data set to be inputted to the next step. third step is the use of the deep learning model (convolutional neural network) to extract the features and classify the input sound. Fig. 11 demonstrates the overall system architecture.

6.1. Preprocessing

Taking into consideration that different databases might be tested on this project, it is important that all the audio files go through an antialiasing low pass filter in order to be resampled. The resultant frequency rate should be 16Khz before any processing is implemented. The utterances will then be converted into their respected spectrogram. the spectrogram consists of two axis time and frequency axis, where time is shown as a horizontal axis and the frequency is shown as the vertical axis. The spectrogram is simply a picture that shows the different variation of energy with different frequency at different times. The intensity of the energy will be represented by either a darkness or different colors. The fact that every period of vocal vibration is associated with glottal pulse gives a special importance to the vocal fold vibration. Furthermore, in the preprocessing stage of the proposed model, all the audio files will be converted in a wide-band spectrogram. The hamming window will be fixed along with the overlap. DFT will also be fixed. e.g. (hamming windows is 4ms with 4ms overlap, and 512 DFT points). Notice that any frequency that is greater than 4000hz will be discarded because 4kHz is cindered to be sufficient for speech emotion recognition according to many studies.

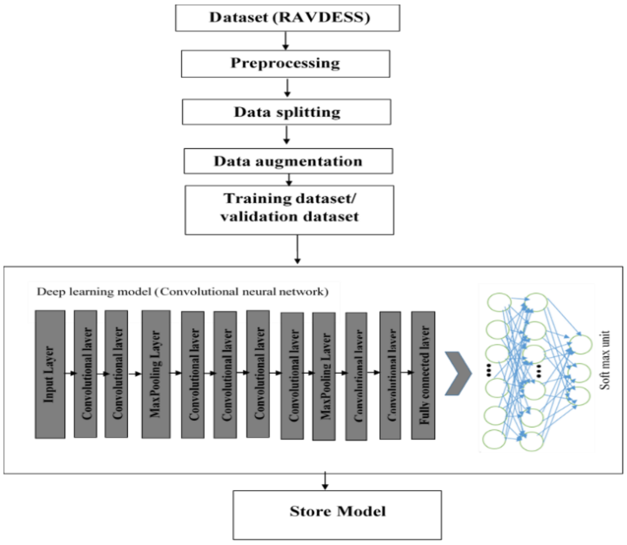


Fig. 11. The proposed architecture of the system (training stage).

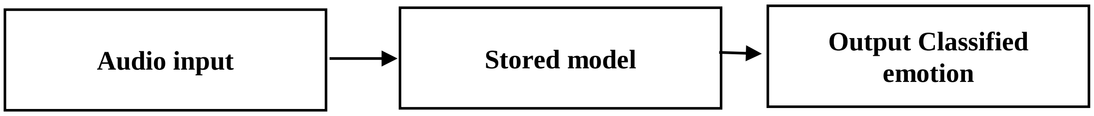
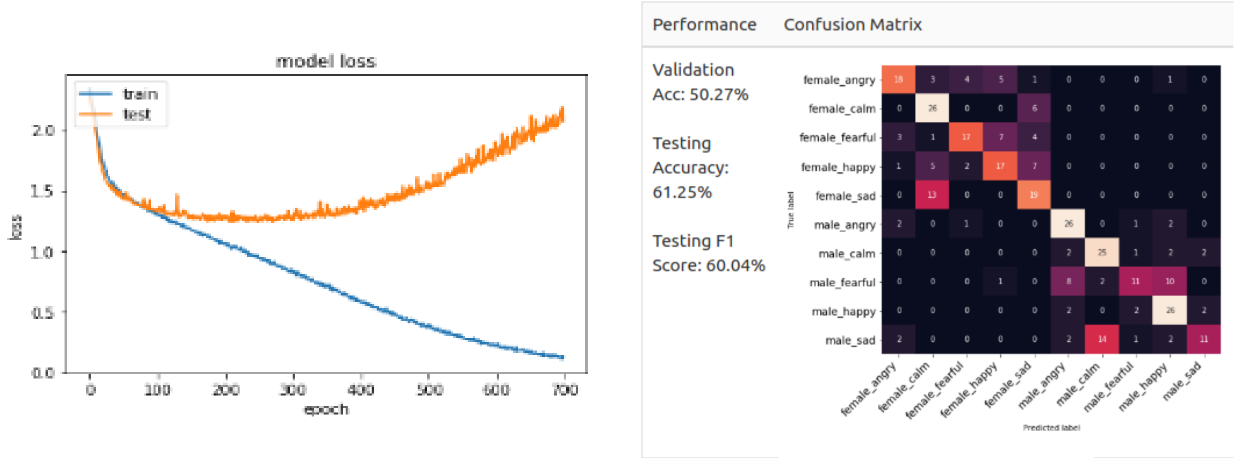


Fig. 12. The proposed model (testing stage).

6.2. Proposed Deep Learning Models

In this work, an initial model was constructed using 7 layers in which 6 of them are convolutional layers and the 7th being a fully connected layer (dense layers). In the first experiment, the whole dataset (male and female recordings) fed to the model, the target emotional classes were 10 set of emotions (angry female, calm female, fearful female, happy female, sad female, angry male, calm male, fearful male, happy male, sad male). The results were not very promising. The model came up with an accuracy of 50% in detecting the validation dataset, and 60% in detecting the test set. The learning curve that model has an overfitting problem as shown in Fig. 13, this is clear from the way the validation plot behaves as it decreases at some point and starts to increase again.



Performance Confusion Matrix

Validation Acc: 50.27%

Testing Accuracy: 61.25%

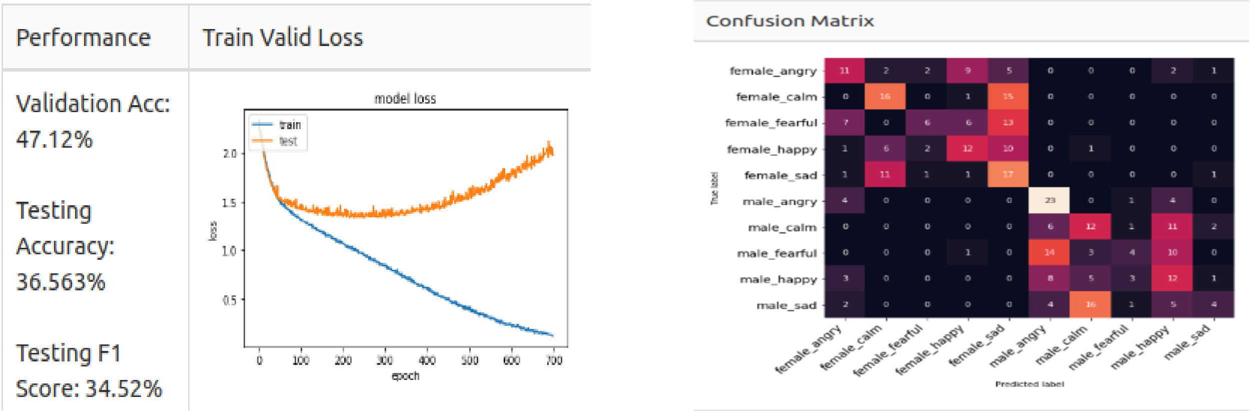
Testing F1 Score: 60.04%

	female_angry	female_calm	female_fearful	female_happy	female_sad	male_angry	male_calm	male_fearful	male_happy	male_sad
female_angry	18	3	4	5	1	0	0	0	1	0
female_calm	0	26	0	0	6	0	0	0	0	0
female_fearful	3	1	17	7	4	0	0	0	0	0
female_happy	1	5	2	17	7	0	0	0	0	0
female_sad	0	13	0	0	19	0	0	0	0	0
male_angry	2	0	1	0	0	26	0	1	2	0
male_calm	0	0	0	0	0	2	25	1	2	2
male_fearful	0	0	0	1	0	8	2	11	10	0
male_happy	0	0	0	0	0	2	0	2	26	2
male_sad	2	0	0	0	0	2	14	1	2	11

Fig. 13. (a) Learning curve and overfitting (b) Results of initial model in detecting 10 targeted classes.

Another problem with this model is that the testing set scored a higher accuracy than the validation dataset, this indicated that there is a data leakage problem in which the data used in the validation dataset are the same as the one used in the testing dataset. To validate this theory of a data leakage problem, the RAVDESS dataset was divided into 3 different sections: training, validation, and testing, to ensure that data are not used in validation and testing. Although a better practice would use a cross validation method to come up with the best possible training and testing set. The dataset was split as follow:

- Training and validation data set (actor 1 to actor 20)
- Testing dataset (actor 21 to actor 24).



Performance Train Valid Loss

Validation Acc: 47.12%

Testing Accuracy: 36.563%

Testing F1 Score: 34.52%

Confusion Matrix

	female_angry	female_calm	female_fearful	female_happy	female_sad	male_angry	male_calm	male_fearful	male_happy	male_sad
female_angry	11	2	2	9	5	0	0	0	2	1
female_calm	0	16	0	1	15	0	0	0	0	0
female_fearful	7	0	6	6	13	0	0	0	0	0
female_happy	1	6	2	12	10	0	1	0	0	0
female_sad	1	11	1	1	17	0	0	0	0	1
male_angry	4	0	0	0	0	23	0	1	4	0
male_calm	0	0	0	0	0	6	12	1	11	2
male_fearful	0	0	0	1	0	14	3	4	10	0
male_happy	3	0	0	0	0	8	5	3	12	1
male_sad	2	0	0	0	0	4	16	1	5	4

Fig. 14. Learning curve after splitting the data into training, validation and testing.

Fig. 14 shows the resulting confusion matrix and the train valid loss plot. We can still see the model suffering from the overfitting problem, however, data leakage problem should be solved and this can be shown in the substantial drop in the testing accuracy of the model. The next step is to split the recording into two different parts, male and female, this is because it can be shown in the confusion matric that the model is confusing male and female emotions, especially in the case of anger and happiness. This could be because of the fact that male actors express their (anger) emotion the same way female actors expresses their (happiness) emotions look section (DATA OBSERVATION), hence the results of splitting the data was as follows:

Male dataset:

Training dataset: consists of 640 samples gathered from actors (1-10).

Validation dataset: consists of 160 samples gathered from actors (1-10).

Testing dataset: consists of 160 samples gathered from actors (11-12).

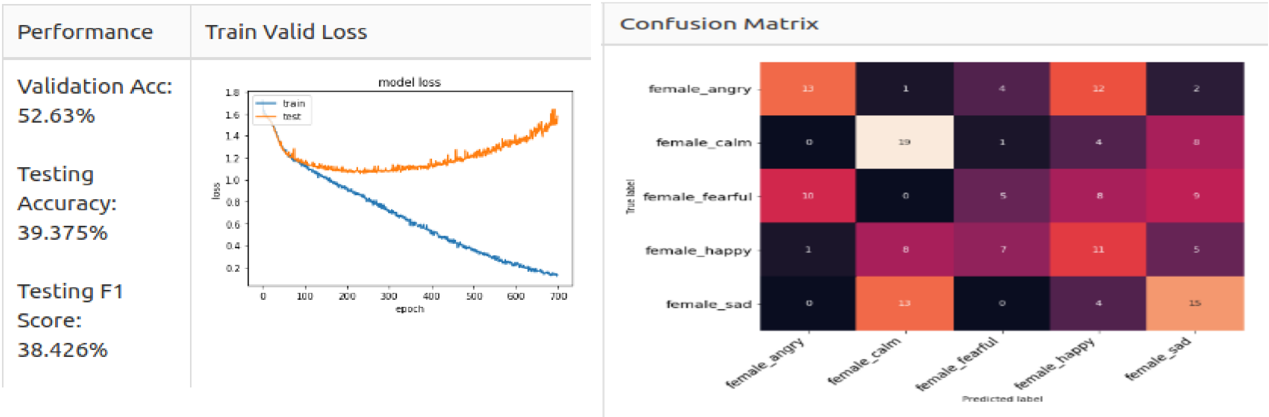
Female dataset

Training dataset: consists of 608 samples gathered from actors (1-10).

Validation dataset: consists of 152 samples gathered from actors (1-10).

Testing dataset: consists of 160 samples gathered from actors (11-12).

After splitting the dataset, the number of targeted classes reduced to 5 and, then the data fed into the model in two different times and the results came as follows:



Performance

Validation Acc: 52.63%

Testing Accuracy: 39.375%

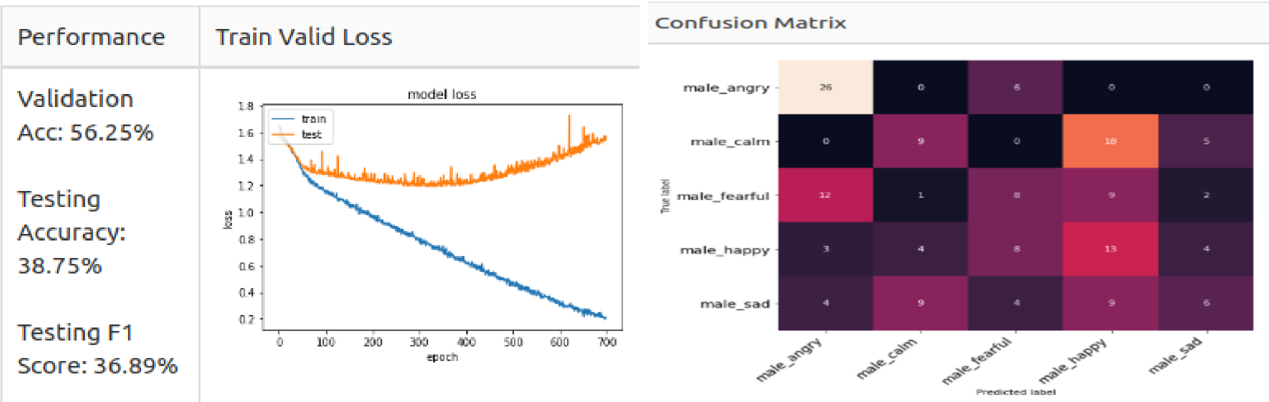
Testing F1 Score: 38.426%

Train Valid Loss

Confusion Matrix

	female_angry	female_calm	female_fearful	female_happy	female_sad
female_angry	13	1	4	12	2
female_calm	0	19	1	4	8
female_fearful	10	0	5	8	9
female_happy	1	8	7	11	5
female_sad	0	13	0	4	15

Fig. 15. (a) Learning curve and overfitting (b) Results of initial model in detecting lasses of female emotions.



Performance

Validation Acc: 56.25%

Testing Accuracy: 38.75%

Testing F1 Score: 36.89%

Train Valid Loss

Confusion Matrix

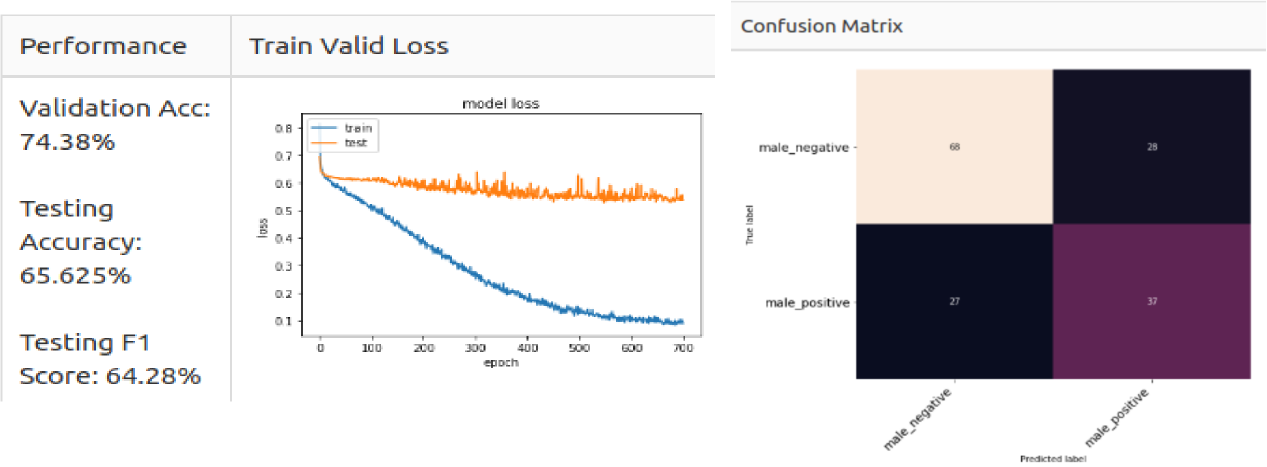
	male_angry	male_calm	male_fearful	male_happy	male_sad
male_angry	26	0	6	0	0
male_calm	0	9	0	18	5
male_fearful	12	1	8	9	2
male_happy	3	4	8	13	4
male_sad	4	9	4	9	6

Fig. 16. (a) Learning curve and overfitting (b) Results of initial model in detecting classes of male data.

The results of the experiment are shown in Fig. 16, the accuracy of the model increased a tiny bit, however it also showed that the model still has an overfitting problem, thus we discovered that it was needed for the model to be modified in order to come up with better results.

6.3. Improvement of the Model

The new model consisted of 8 convolutional layers, followed by a Dense layer. The Dropout value was set to 0.25 and the SGD optimizer was used with a learning rate of 0.0001. Also, the model was set to reduce learning rate when a metric has stopped improving using Keras 'ReduceLROnPlateau' function. Below are the details of the mentioned model. Moreover, the best resulting accuracy that is based on the new model have been tested on the custom database as well as being implanted on the real live demo. Because of the limited computational resources and for a better accuracy, the data were divided into 3 groups of emotional classes, first group contains 2 emotional classes, the second group contains 3 emotional classes and the third group countians 5 emotional classes. The 2 emotional classes group contain a positive class and a negative class. The positive class includes the following emotions: (happy, calm), while the negative class includes the following emotions (fear, anger, sadness). The 3 classes emotion group contain a positive, neutral and a negative, the positive includes (happiness), the neutral contains (calm or neutral), and lastly the negative class containing (anger, fear, and sadness). The third group contains the following emotions: angry, calm fearful happy, and sad. Then emotion distribution of each group was plotted and then the male date was used to test the model. The 2 class and 3 class emotion groups were used in the test. The results are shown in Figs. 17-19.:



Performance

Validation Acc: 74.38%

Testing Accuracy: 65.625%

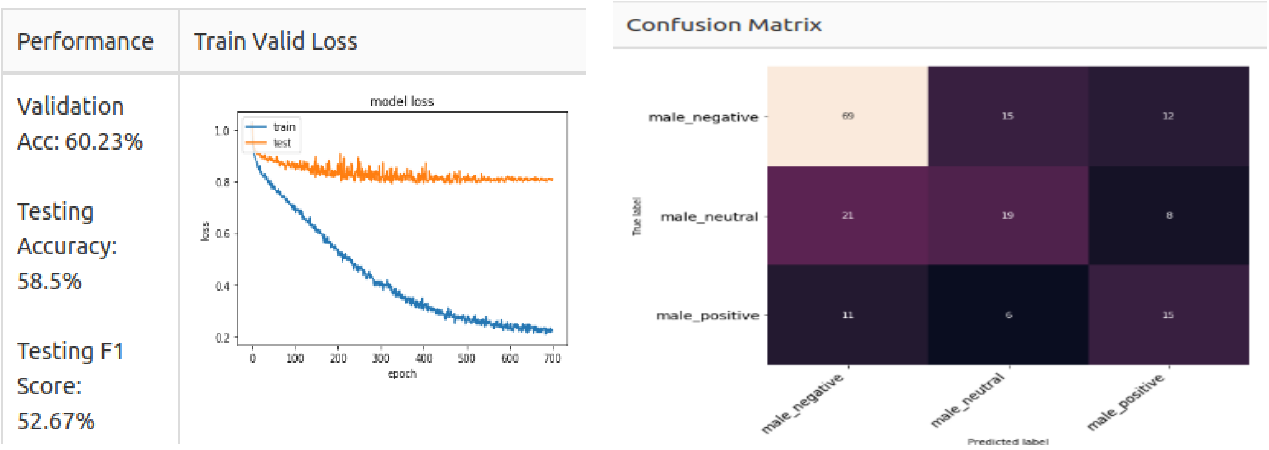
Testing F1 Score: 64.28%

Train Valid Loss

Confusion Matrix

	male_negative	male_positive
male_negative	68	28
male_positive	27	37

Fig. 17. (a) Learning curve and overfitting (b) Results of improved model in detecting two emotional classes (Male samples).



Performance

Validation Acc: 60.23%

Testing Accuracy: 58.5%

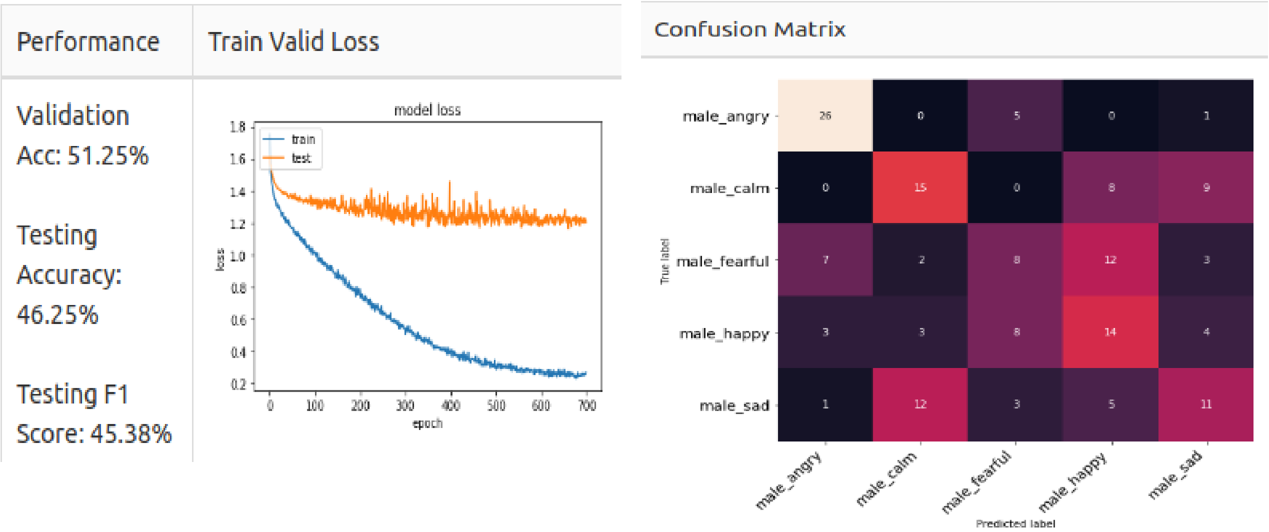
Testing F1 Score: 52.67%

Train Valid Loss

Confusion Matrix

	male_negative	male_neutral	male_positive
male_negative	69	15	12
male_neutral	21	19	8
male_positive	11	6	15

Fig. 18. (a) Learning curve and overfitting (b) Results of improved model in detecting three emotional classes (Male samples).



Performance

Validation Acc: 51.25%

Testing Accuracy: 46.25%

Testing F1 Score: 45.38%

Train Valid Loss

Confusion Matrix

	male_angry	male_calm	male_fearful	male_happy	male_sad
male_angry	26	0	5	0	1
male_calm	0	15	0	8	9
male_fearful	7	2	8	12	3
male_happy	3	3	8	14	4
male_sad	1	12	3	5	11

Fig. 19. (a) Learning curve and overfitting (b) Results of improved model in detecting five emotional classes (Male samples).

The result in the figure above shows a dramatic increase in the validation accuracy. Nevertheless, it is clear from the results obtained that the training valid loss shows a relatively high loss value which is a clear sign of an underfitting problem in the model. We figured that this problem occurred because the model was not fed enough data. The reason that we did not face this problem in the first experiment is that we did not split the dataset into female and male set, the splitting process caused the number of training data to drop significantly.

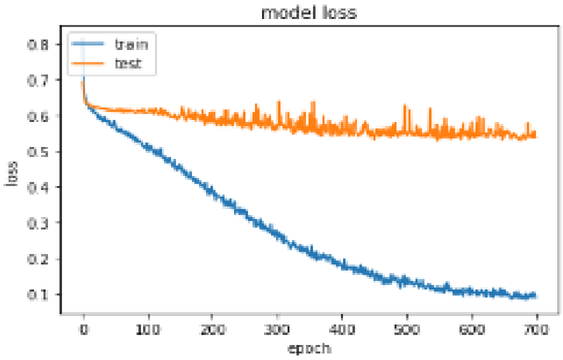


Fig. 20. learning curve shows a high loss value for validation dataset indicating an under fitting problem.

7. DATA AUGMENTATION

This step took place to make sure that the model will not have an underfitting problem. Initially we implemented the augmentation method on the model to classify 2 targeted classes, the goal was to determine the best data augmentation method. The next step was implementing the chosen augmentation method on the model to classify 3 and 5 targeted classes.

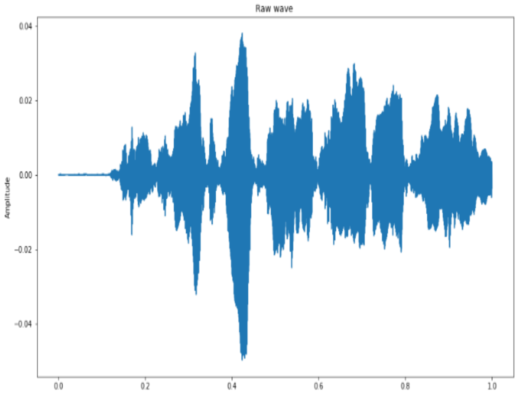


Fig. 21. Before data augmentation.

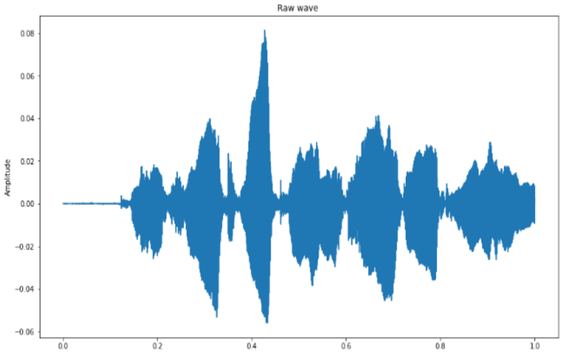
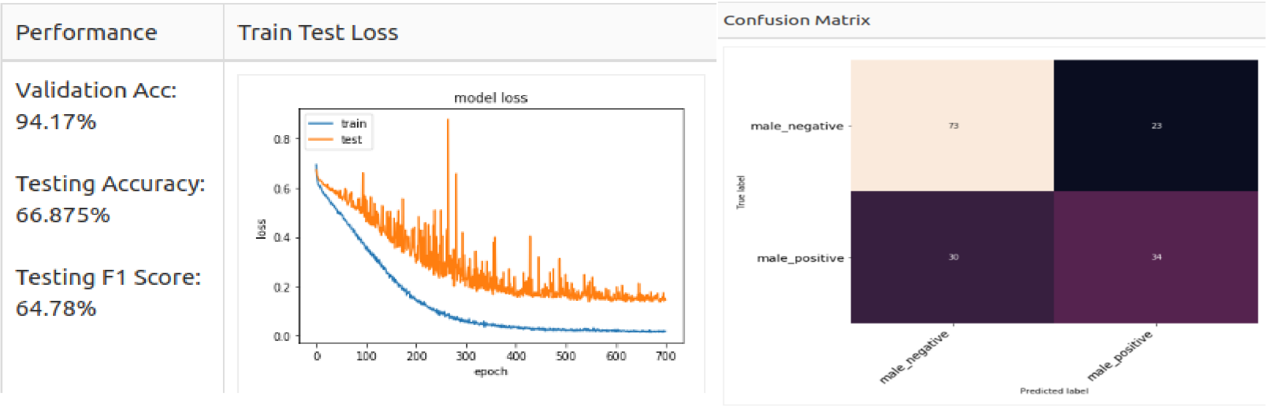


Fig. 22. After data augmentation.

The two methods used were Noise Adding combined with Shifting, Noise Adding combined with Pitch Tuning. The results came as follow:



Performance

Validation Acc: 94.17%

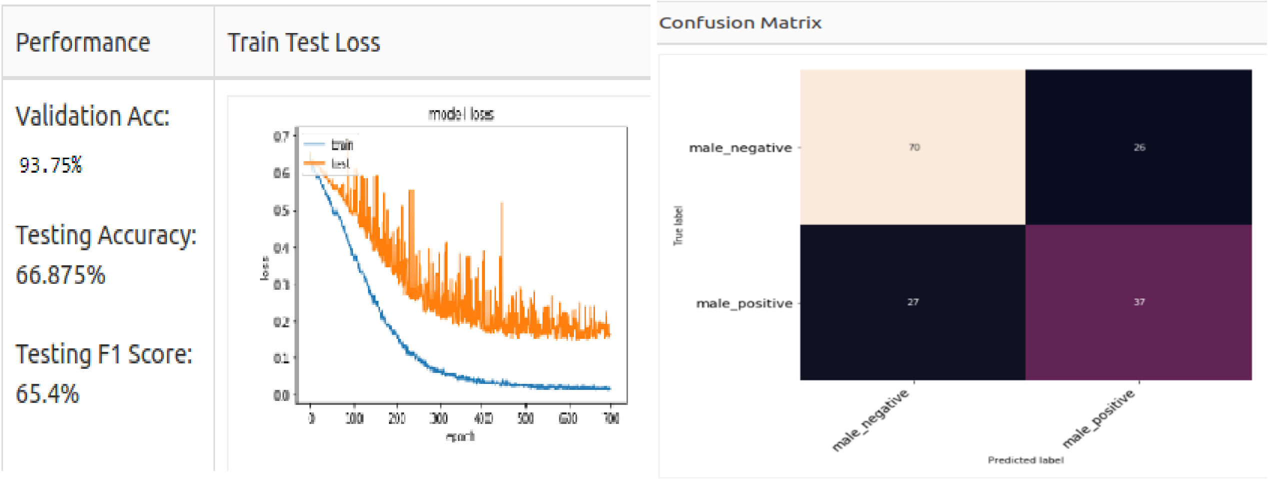
Testing Accuracy: 66.875%

Testing F1 Score: 64.78%

Train Test Loss

Confusion Matrix

Fig. 23. Two classes emotional classifier with (Noise Adding +Time Shifting).

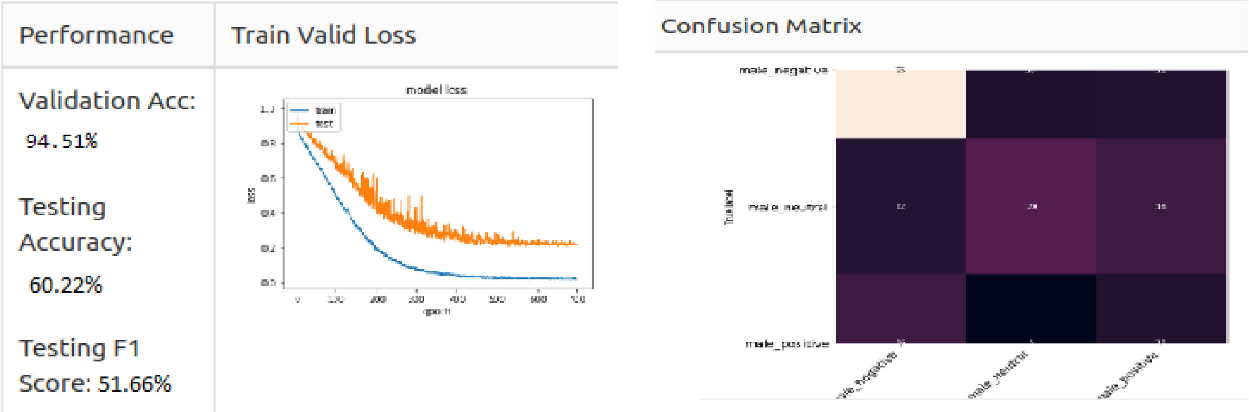


Performance	Train Test Loss
Validation Acc: 93.75% Testing Accuracy: 66.875% Testing F1 Score: 65.4%	

Confusion Matrix

Fig. 24. Two classes emotional classifier with (Pitch Shifting + Noise Adding).

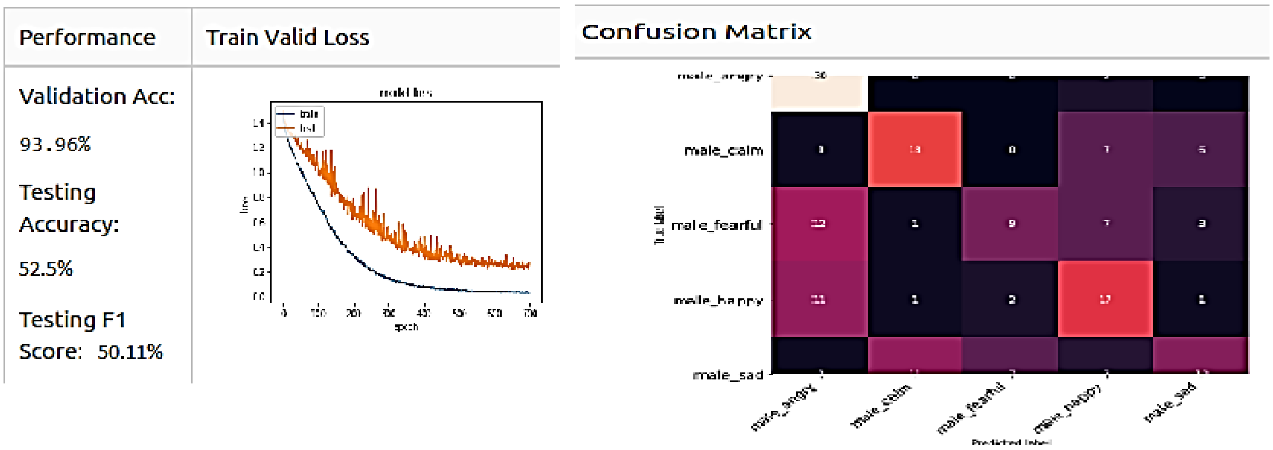
As we can see from the results above, Noise Adding + Time Shifting method produced the best accuracy, hence I carried the method forward and tested on the two remaining groups (3 classes emotions, 5 classes emotions) and the results came as follow:



Performance	Train Valid Loss
Validation Acc: 94.51% Testing Accuracy: 60.22% Testing F1 Score: 51.66%	

Confusion Matrix

Fig. 25. Three classes emotional classifier with (Noise Adding + Time Shifting).



Performance	Train Valid Loss
Validation Acc: 93.96% Testing Accuracy: 52.5% Testing F1 Score: 50.11%	

Confusion Matrix

Fig. 26. Five classes emotional classifier with (Noise Adding + time Shifting).

8. LIVE DEMO

To demonstrate the effectiveness of the trained model in real life, a simple live demo was produced. The demo uses the trained 5 emotional classes model to try to classify speech sample from real life. The samples were collected from 4 IIUM students in which every student has to speak one sentences in 5 different emotions. The given sentence was “kids are walking by the door”. The table below shows the results of the experiment.

Table 3. Results from life experiment.

Actor No.	Actual value	Predicted value	Result
1	Calm	Calm	True
1	Happy	Calm	False
1	Sad	Scared	False
1	Scared	Calm	False
1	Angry	Angry	True
2	Calm	calm	True
2	Happy	Calm	False
2	Sad	Calm	False
2	Scared	Angry	False
2	Angry	Angry	True
3	Calm	calm	True
3	Happy	Calm	False
3	Sad	Calm	False
3	Scared	Angry	False
3	Angry	Angry	True
4	Calm	calm	True
4	Happy	Calm	False
4	Sad	calm	False
4	Scared	Calm	False
4	Angry	Sad	False

Table 3 shows the accuracy of the model scored 35% only. The poor result is a due to many factors. First is that the training and testing data samples are from different distributions. Ideally the training and testing data should be from the same dataset as the trained. Model will be able to detect familiar data easily. Another reason is that the accent of actors in which the model was trained on is far different from the test actors. Person’s accent makes a huge different of how the speech wave form generated will look like. However, it is noticeable that the model did well in detecting the Anger emotions. This is because anger can mostly be expressed by increasing the volume of the speaker. Looking at the waveform coming from different actor expressing anger shows a lot of similarities.

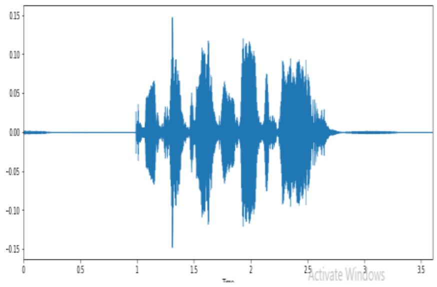


Fig. 27. Actor expressing anger (self-collected test data).

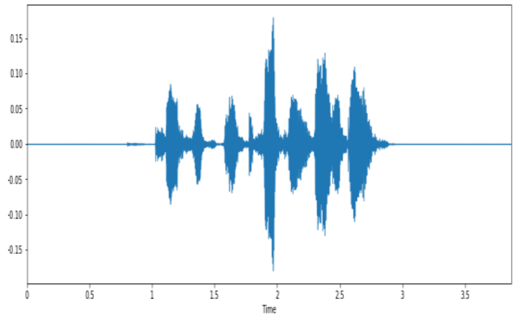


Fig. 28. Actor expressing anger (RAVESSD database).

9. CONCLUSION

In this paper, a full neural network was constructed and demonstrated, and all the results were analyzed. The database was successfully chosen based on the availability and the language used in it. The proposed deep learning architecture is a convolution neural network with 8 convolutional layers and the model consisted of only one fully connected layer. A mix data augmentation method was used which is Noise Adding and Time Shifting which the proposed CNN model achieved 93.96% accuracy rate in detecting 5 emotions, 94.51% when detecting 3 emotions and, 94.17% when detecting 2 emotions. The results were analyzed using confusion matrix.

REFERENCES

[1] M. A. Mahjoub, M. Mbarki, Y. Serrestou, L. Kerkeni, and K. Raoof, **“Speech Emotion Recognition: Methods and Cases Study,”** Vol. 2, Icaart, pp. 175–182, 2018.

[2] Po-Yuan Shih and Chia-Ping Chen, **“Speech Emotion Recognition with Skew-Robust Neural Networks**”, *Computer Science and Engineering Kaohsiung, Taiwan ROC Academia Sinica Institute of Information Science*, pp. 2751–2755, 2017.

[3] Li Zheng; Qiao Li; Hua Ban and Shuhua Liu, **“Speech emotion recognition based on convolution neural network combined with random forest”,** *2018 Chinese Control And Decision Conference (CCDC)*, 2018..

[4] Lin Yilin and Wei Gang, **"Speech Emotion Recognition Based on HMM and SVM",** *Proc of the 4th International Conference on Machine Learning and Cybernetics*, Vol. VIII, pp. 4898-4901, 2005.

[5] W Lim, D Jang and T. Lee, **"Speech emotion recognition using convolutional and Recurrent Neural Networks[C]",** *Signal and Information Processing Association Annual Summit and Conference (APSIPA) 2016 Asia-Pacific*, pp. 1-4, 2016.

[6] M. Fayek, M. Lech, and L. Cavedon, **“Evaluating deep learning architectures for Speech Emotion Recognition,”** *Neural Networks*, Vol. 92, pp. 60–68, 2017.

[7] J. Engelbart, **“A Real-Time Convolutional Approach To Speech Emotion Recognition,”** July, 2018.

[8] Y. Niu, D. Zou, Y. Niu, Z. He, and H. Tan, **“A breakthrough in Speech emotion recognition using Deep Retinal Convolution Neural Networks,”** pp. 1–7.

[9] J. Zhao, X. Mao, and L. Chen, **“Speech emotion recognition using deep 1D & 2D CNN LSTM networks,''** *Biomed. Signal Process. Control*, Vol. 47, pp. 312_323, Jan

[10] C. Z. Seyedmahdad Mirsamadi, Emad Barsoum, **“Automatic Speech Emotion Recognition Using Recurrent Neural Networks With Local Attention Center for Robust Speech Systems , The University of Texas at Dallas , Richardson , TX 75080 , USA Microsoft Research , One Microsoft Way , Redmond , WA 98052 , USA,”** *ICASSP 2017, Proc. 42nd IEEE Int. Conf. Acoust. Speech, Signal Process.*, pp. 2227–2231, 2017.

[11] P. Tzirakis, J. Zhang, and W. Schuller, **“END-TO-END SPEECH EMOTION RECOGNITION USING DEEP NEURAL NETWORKS”,** Department of Computing , Imperial College London , London , UK Chair of Embedded Intelligence for Health Care and Wellbeing , University of Augsburg , Germany, *Icassp 2018*, pp. 5089–5093, 2018.

[12] J. A. Russell, **“A Circumplex Model of Affect,”** December 1980, 2016.

[13] T. A. Burns, **“The Nature of Emotions,”** *Int. J. Philos. Stud.*, Vol. 27, No. 1, pp. 103–106, 2019.

[14] Lin Yilin and Wei Gang, **"Speech Emotion Recognition Based on HMM and SVM"**, *Proc of the 4th International Conference on Machine Learning and Cybernetics*, Vol. VIII, pp. 4898-4901, 2005.

[15] M. Fayek, M. Lech, and L. Cavedon, **“Evaluating deep learning architectures for Speech Emotion Recognition,”** *Neural Networks*, Vol. 92, pp. 60–68, 2017.

[16] J. Engelbart, **“A Real-Time Convolutional Approach To Speech Emotion Recognition,”** July, 2018.

[17] S. Demircan and H. Kahramanlı, **“Feature Extraction from Speech Data for Emotion Recognition,”** January, 2014.

[18] P. P. Singh and P. Rani, **“An Approach to Extract Feature using MFCC,”** Vol. 04, no. 08, pp. 21–25, 2014.

[19] S. Bhadra, U. Sharma, and A. Choudhury**, “Study on Feature Extraction of Speech Emotion Recognition,”** Vol. 4, No. 1, pp. 3–5, 2016.

[20] K. Han, D. Yu, and I. Tashev**, “Speech Emotion Recognition Using Deep Neural Network and Extreme Learning Machine,”** *Fifteenth Annu. Conf. …*, September, pp. 223–227, 2014.

[21] O. Ameena, **“Speech Emotion Recognition Using Hmm, Gmm and Svm Models B.Sujatha (Me),”** *Int. J. Prof. Eng. Stud.*, Vol. VI, No. 3, pp. 311–318, 2016.

[22] Y. Sun, **“Neural Networks for Emotion Classification,”** *Science (80-.).*, Vol. 20, No. 9, pp. 886–899, 2011.

[23] B. Xu, N. Wang, and T. Chen**, “Empirical evaluation of rectified activations in convolutional network,”** *arXiv preprint arXiv*:1505.00853, 2015.

[24] M. A. . Tanner and W. H. Wong , **“The Calculation of Posterior Distributions by Data Augmentation”**, *Journal of the American Statistical Association* , Vol . 82 , No . 398 (Jun ., 1987), pp. 528–540, 2009.

[25] N. Jaitly and G. E. Hinton, **“Vocal Tract Length Perturbation (VTLP) improves speech recognition.”,** *Proceedings of the 30 th International Conference on Machine Learning, Atlanta, Georgia*, USA, 2013.

[26] R. Y Rubinstein, D. P Kroese, **“Imulation and the Monte Carlo Method”,** *2nd edn. 2007 Wiley, New York.*

[27] Y. Q. K. Y. J. Niu, **"Acoustic emotion recognition using deep neural network,"** *Acoustic emotion recognition using deep neural network,* pp. pp. 128-132, 2014.

[28] X. Zhou, J. Guo, and R. Bie, “**Deep Learning Based Affective Model for Speech Emotion Recognition,”** pp. 841–846, 2016.

[29] R. R. Salakhutdinov and G. E. Hinton. **“Deep Boltzmann machines”,** *In Proceedings of the International Conference on Artificial Intelligence and Statistics*, Vol. 12, 2009.

[30] F. Liu, B. Liu, C. Sun, M. Liu, X. Wang, **“Deep belief network-based approaches for link prediction in signed social networks”,** *Entropy 17* pp. 2140–2169, 2015.

[31] G.E. Hinton, S. Osindero, Y.W. Teh, **“A fast learning algorithm for deep belief nets”,** *Neural Comput 18*, pp. 1527–1554, 2006.

[32] Yoshua Bengio. **“Learning deep architectures for AI. Foundations and Trends in Machine Learning”,** *Also published as a book. Now Publishers*, Vol. 2(1),1pp. –127, 2009.

[33] P. Vincent , **“Hugo Larochelle Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion”,** *Journal of Machine Learning Research* Vol. 11, pp. 3371-3408,2010.

[34] M. F. Alghifari, T. S. Gunawan, and M. Kartiwi, **"Speech emotion recognition using deep feedforward neural network,"** *Indones. J. Electr. Eng. Comput. Sci.*, Vol. 10, No. 2, pp. 554–561, 2018.

[35] X. Ke, Y. Zhu, L. Wen, and W. Zhang, **"Speech Emotion Recognition Based on SVM and ANN,"** *Int. J. Mach. Learn. Comput.*, Vol. 8, No. 3, pp. 198–202, 2018.

[36] P. Tzirakis, J. Zhang, and W. Schuller, **"End-To-End Speech Emotion Recognition Using Deep Neural Networks Department of Computing**", Imperial College London , London , UK Chair of Embedded Intelligence for Health Care and Wellbeing , University of Augsburg , Germany, *Icassp 2018*, pp. 5089–5093, 2018.