Classifying Emotional Utterances by Employing Multi-modal Speech Emotion Recognition

Dipankar Das

Computer Sc. & Engineering Department,
Jadavpur University, West Bengal, India

dipankar.dipnil2005@gmail.com

Abstract

Deep learning methods are being applied to several speech processing problems in recent years. In the present work, we have explored different deep learning models for speech emotion recognition. We have employed normal deep feed-forward neural network (FFNN) and convolutional neural network (CNN) to classify audio files according to their emotional content. Comparative study indicates that CNN model outperforms FFNN in case of emotions as well as gender classification. It was observed that the sole audio based models can capture the emotions up to a certain limit. Thus, we attempted a multi-modal framework by combining the benefits of the audio and text features and employed them into a recurrent encoder. Finally, the audio and text encoders are merged to provide the desired impact on various datasets. In addition, a database consists of emotional utterances of several words has also been developed as a part of this work. It contains same word in different emotional utterances. Though the size of the database is not that large but this database is ideally supposed to contain all the English words that exist in an English dictionary.

1 Introduction

Human Computer Interaction (HCI) researches the way we humans interact with a computer in order to improve the existing technologies. Thus, Automatic Speech Recognition (ASR) has been an active field of AI research aiming to generate machines that communicate with people via speech [1] [2]. In recent trends, simple text based chatbot systems are adding extra flavor of personalized experiences to their users through speech interactions. However, emotions always play the important roles in our interactions with people and computers. Fundamental publications of Rosalind Picard on affective computing increased the awareness in HCI community regarding important roles of emotion [3] [4] [5] [6]. Since then, researchers have also become increasingly aware of the importance of emotion in the design process [7].

Speech Emotion Recognition (SER) is mostly beneficial for commercial HCI applications, such as speech synthesis, customer service, education, forensics and medical analysis. Emotion recognition is used in call center for classifying calls according to emotions [8] and it serves as the performance parameter for conversational analysis [9], customer satisfaction and so on. SER is also used in automotive industry especially in car board system based on mental state of the driver to initiate his/her safety by preventing accidents to happen [10].

Affective computing and HCI research used to target in reducing user frustration, building tools to support development of socio-emotional skills [11]. Without information about emotions, it is difficult to achieve a harmonic and natural man-machine interface for applications such as patient care, geriatric nursing, call centers, psychological consultation, and human communication [12]. Therefore, health care industry is becoming prominent because it leverages emotion recognition techniques to solve complex patient related problems.

Speech is an information-rich signal that contains paralinguistic information as well as linguistic information. As a result of this, speech conveys more emotional information than text. This reality motivates many researchers to

Proceedings of the Workshop on Speech and Music Processing (SMP-2021), pages 1–13
co-located with ICON-2021, December 16-19, 2021, Silchar, India.

consider speech signal as a quick, effective and natural process to identify interaction mysteries between computer and human. Although, there is a significant improvement in speech recognition but still researchers are away from natural interplay between computer and human, since computer is not capable of understanding human emotional state. The recognition of emotional speech aims to recognize the emotional condition of individual utterances by applying his/her voice automatically. Recognizing of emotional conditions in speech signals are so challenging area for several reasons.

Majority of the speech emotional methods used to select the best features that are powerful enough to distinguish between different emotions.

The presence of various languages, accents, sentences, speaking styles, speakers also adds another difficulty because these characteristics directly change most of the extracted features including pitch and energy [13].

Furthermore, it is possible to have a more than one specific emotion at a time in the same speech signal and each emotion may correlate with a different part of speech signals. Therefore, defining the boundaries between parts of emotion is very challenging task.

In the present task with respect to speech emotion recognition, we have proposed two systems based on deep learning method to classify a speech signal according to its emotional content.

1. The first model is based on simple deep Feed-Forward Neural Network (FFNN). As it is a very basic model, it was unable to recognize enough important features from speech signal to classify it accurately. The overall accuracy that we achieved from this system is only 40%.

2. The second model is based on Convolutional Neural Network (CNN) model. Our main contribution lies in the way we applied the CNN model to our dataset. In several studies, it is observed that CNN have been used to classify speech emotion but the CNN model was applied on the spectrogram image which is a visual representation of the spectrum of frequencies of an audio signal. In contrast, we have applied our CNN model on the array of low-level MFCC features, extracted from the spectrogram image of an audio signal. Due to this fact, we used 1- Dimensional Convolutional layers in our CNN and not 2-Dimensional ones, which are generally used on image data. The overall accuracy we achieved from this model is 65%.

Now apart from these two systems, we have also developed an emotional lexicon that contains utterances of words along with their emotional class. Moreover, the lexicon also contains the utterances of a particular word when belongs to one or more emotion categories (same word can belong to multiple categories of emotion).

Finally, we introduce a deep recurrent encoder model that exploits text data and audio signals both simultaneously to obtain a better understanding of the emotional aspects in speech signals. In real world, a multi-modal dialogue system is composed of sound and spoken content. This actually motivated us to build a system which can encode the information from audio and text sequences and then can combine the information from these sources to predict the emotion class. Our system reported accuracies ranging from 62.7% to 70.8% when it was applied to the IEMOCAP dataset.

The rest of the paper is organized as follows. Section 2 describes the related attempts carried out under speech emotion recognition. The details on two types of emotional speech datasets along with two different models for speech emotion classification are discussed in Section 3. Section 4 describes a deep learning based multi-modal framework that takes into account the roles of speech and text in order to develop an improved system. Experiments and associated results with respect to all the models and framework are explained in Section 5. Finally, Section 6 briefs the process of developing speech emotion lexicon as an outcome whereas the concluding remarks are made in Section 7.

2 Related Work

If we observe a comprehensive review of speech emotion recognition systems targeting pattern recognition researchers who do not necessarily have a deep background in speech analysis, we notice three main aspects of this research field: (1) important design criteria of emotional speech corpora, (2) impact of speech features on the classification performance of SER and (3) classification systems employed in SER.

L. Chen et al. [15] used multi-level SVM classifier and ANN to reduce dimensionality by employing several parmeters (e.g., energy, ZCR, pitch, SC, spectrum cut-off frequency, correlation density (Cd), fractal dimension, MFF etc.) and obtained 86.5%, 68.5% and 50.2% recognition rates at different levels on Beihang University Database of Emotional Speech (BHUDES). Similarly, the authors in [16] used binary classifier and QDC with prosodic and contour features to obtain 75.8% rate of recognition on SEMAINE functional data. In recent trends, H. Cao et al. [14] used SVM with prosodic and spectral features and obtained 44.4% recognition rate on Berlin & LDC & FAU Aibo dataset.

In addition to the above mentioned works, in [17], a novel Modulation Spectral Features (MSFs) for the recognition of human emotions in speech is presented. An auditory-inspired ST representation is acquired by deploying an auditory filter bank as well as a modulation filter bank, to perform spectral decomposition in the conventional acoustic frequency domain and in the modulation frequency domain, respectively

This authors in [18] focused on the data pre-processing techniques which aim to extract the most effective acoustic features to improve the performance of the emotion recognition. The technique can be applied on a small sized data set with a high number of features. The presented algorithm integrates the advantages from a decision tree method and the random forest ensemble. Experiment results on a series of Chinese emotional speech data sets indicate that the presented algorithm can achieve improved results on emotional recognition, and outperform the commonly used Principle Component Analysis (PCA) / Multi-Dimensional Scaling (MDS) methods, and the more recently developed ISO-Map dimensionality reduction method.

In [19], a fusion-based approach to emotion recognition of affective speech using multiple classifiers with acoustic-prosodic information (AP) and semantic labels (SLs) is presented. The acoustic-prosodic information was adopted for emotion recognition using multiple classifiers and the MDT was used to select an appropriate classifier to output the recognition confidence.

It is observed that all the above mentioned approaches are either tried to deal with signals, acoustic features or to use machine learning classifiers and feature reduction techniques to improve the performance of SER. In contrast, our proposed method is based on deep learning and applied on three different datasets to show the effectiveness. In addition, the multi-modal framework deals with both the texts and speech together to capture the insights under the deep learning umbrella. The development of speech emotion lexicon directs us the utilization of the proposed models.

3 Speech Emotion Recognition

A single word can be associated with multiple emotions [20]. Based on this hypothesis, we have built our emotion classifier and chosen datasets carefully. Although there are several other modalities such as facial expression, body language, through which emotions can be expressed but we limited our present study to speech modality only. Speech emotion corpora that were prepared by actors have been used in the current study because the emotions expressed with exaggeration potentially compensate the lack of information provided by other modalities. This also allows us to explore the effectiveness of deep learning models with greater control compared with daily-life utterances. However, we limited our model to classify emotions for 'English' language only.

3.1 Speech Emotion Corpora

SAVEE: British English Database: The Surrey Audio-Visual Expressed Emotion (SAVEE) database was recorded from four native English male speakers (identified as DC, JE, JK and KL), postgraduate students and researchers at the University of Surrey aged from 27 to 31 years. Emotion has been described psychologically in discrete categories: anger, disgust, fear, happiness, sadness and surprise [21]. This is supported by the cross-cultural studies of Ekman [22] and studies of automatic emotion recognition tended to focus on recognizing these [23]. We added the class neutral to provide recordings of 7 emotion categories. The text material consisted of 15 sentences per emotion: 3 common among all emotions, 2 emotion-specific and 10 generic sentences that were different for each emotion and phonetically-balanced. The sampling rate of all recordings was 44.1 kHz. The 3 common and $2 \times 6 = 12$ emotion-specific sentences were recorded as neutral to give 30 neutral sentences. This

resulted in a total of 120 utterances per speaker, for example:

***Common**: She had your dark suit in greasy wash water all year.*

***Anger**: Who authorized the unlimited expense account?*

***Disgust**: Please take this dirty table cloth to the cleaners for me.*

***Fear**: Call an ambulance for medical assistance.*

***Happiness**: Those musicians harmonize marvelously.*

***Sadness**: The prospect of cutting back spending is an unpleasant one for any governor.*

***Surprise**: The carpet cleaners shampooed our oriental rug.*

***Neutral**: The best way to learn is to solve extra problems.*

RAVDESS: Emotional Speech and Song Database: The corpus, Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [24] contains 7356 files (total size: 24.8 GB). The database contains 24 professional actors (12 male, 12 female), vocalizing two lexically-matched statements in a neutral North American accent. Speech includes calm, happy, sad, angry, fearful, surprise and disgust expressions whereas song contains calm, happy, sad, angry and fearful emotions. The statements are "Kids are talking by the door" and "Dogs are sitting by the door". Each expression is produced at two levels of emotional intensity (normal and strong) with an additional neutral expression. All conditions are available in three modality formats: Audio-only (16bit, 48kHz .wav), Audio-Video (720p H.264, AAC 48kHz, .mp4), and Video-only (no sound). We used only the audio modality as our focus was on the recognition of emotion from speech. Speech file (size 215 MB) contains 1440 files: 60 trials per actor x 24 actors = 1440.

3.2 Data Cleaning and Pre-processing

In order to have a consistent sampling rate across all databases, all utterances were resampled and filtered by an antialiasing FIR low pass filter to have frequency rate of 44.1 kHz prior to any processing. All audio utterances were then converted into spectrograms. A spectrogram is an image that displays the variation of energy at different frequencies across time. There are two general types of spectrograms: wide-band and narrow-band spectrograms. Wide-band spectrograms have higher time resolution than narrow-band spectrograms. This property enables the wide-band spectrograms to show individual glottal pulses. In contrast, narrow-band spectrograms have higher frequency resolution than wide-band spectrograms. This feature enables the narrow-band spectrograms to resolve individual harmonics. Considering the importance of vocal fold vibration, along with the fact that glottal pulse is associated with one period of vocal fold vibration, we decided to convert all utterances into wide-band spectrograms.

3.3 Model 1: Feed Forward Neural Network (FFNN)

Deep feed-forward neural network constitutes several layers of hidden neurons, where each neuron is connected to every neuron in its previous layer. The first layer is called input layer. For our study, the input layer consists of 216 MFCC features extracted from the audio data and the batch size has been set to 16. Thus, the dimension of our input data is (16 X 216). We employed three hidden layers in our architecture as depicted in Figure 1. The number of neurons in the first, second and third hidden layer are 256, 512 and 256, respectively. We have used the Rectified Linear Unit (ReLU) activation function in all of the three hidden layers to achieve non-linearity. Only in the output layer, softmax activation function is used as it gives the probability distribution across 10 output classes. As the problem is a classification problem, we have used the cross-entropy loss. Adam optimizer is also employed to minimize the loss function across the training data. We have also employed a dropout rate of 20% after every hidden layer. The dropout layers are employed to reduce the over-fitting problem.

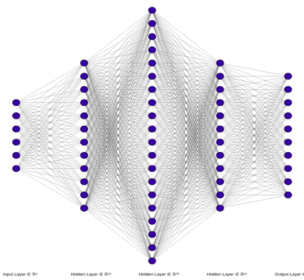

Figure 1: *Baseline architecture of FFNN*

3.4 Model 2: Convolutional Neural Network (CNN)

We have applied Convolutional Neural Network algorithm on audio data. As the data is of one-dimensional, we cannot use the conventional CNN architecture used for image data in general. As a result, we used 1-D convolutional layers instead of the most popular 2-D convolutional layers. All other layers like max-pooling and dense layers are used as it is. The convolutional neural network (CNN) architecture that has been implemented in the current study constitutes two convolutional layers and two fully connected layer, also known as dense layers. Among the two dense layers, the first one has 128 hidden neurons and the second one has 256 hidden neurons. For the current study, we tried to classify each audio file to a particular emotion class among 5 emotion classes and also to classify the gender of the voice. Thus, we have 10 (5 emotions X 2 genders) output classes. As a result we added 10 softmax units in our last (output) layer to estimate the probability distribution of the classes.

In our architecture, every convolutional layer is followed by a max-pooling layer. Each of the first and second convolutional layers is followed by a 1-D max-pooling layer with max-pooling window size of 7 and 4, respectively. The number of kernels (filters) is set to 64 and 128 for the first and second convolutional layers, respectively. The sizes of the kernels that have been applied to the first and second convolutional layers are 5 and 3, respectively. Batch size of 16 is applied throughout the training process. Rectified Linear Units (ReLU) were used in convolutional layers and fully connected layers, except in the last dense layer, as activation functions to introduce non-linearity to the model. Similar to Model 1, as the problem is a classification problem, we have used the cross-entropy loss and adam optimizer. The number of epochs is set to 100. The training procedure for this study was performed entirely on a CPU-based system, no GPU has been used for conducting any part of the training process. We have also used dropout and flatten function. Flatten function is used whenever we needed to reduce the dimension of the data which was output by a layer in the network whereas dropout layer is used to reduce the over-fitting issue during training process. Dropout layers reduce over-fitting by dropping out or ignoring some of the neurons. We have used two dropout layers in our network architecture with each of them residing right after each of the two dense layers. A dropout rate of 20% has been used in both of the two cases. Figure 2 gives a detailed overview of the network with the input and output dimensions of data in each of the layer in the network.

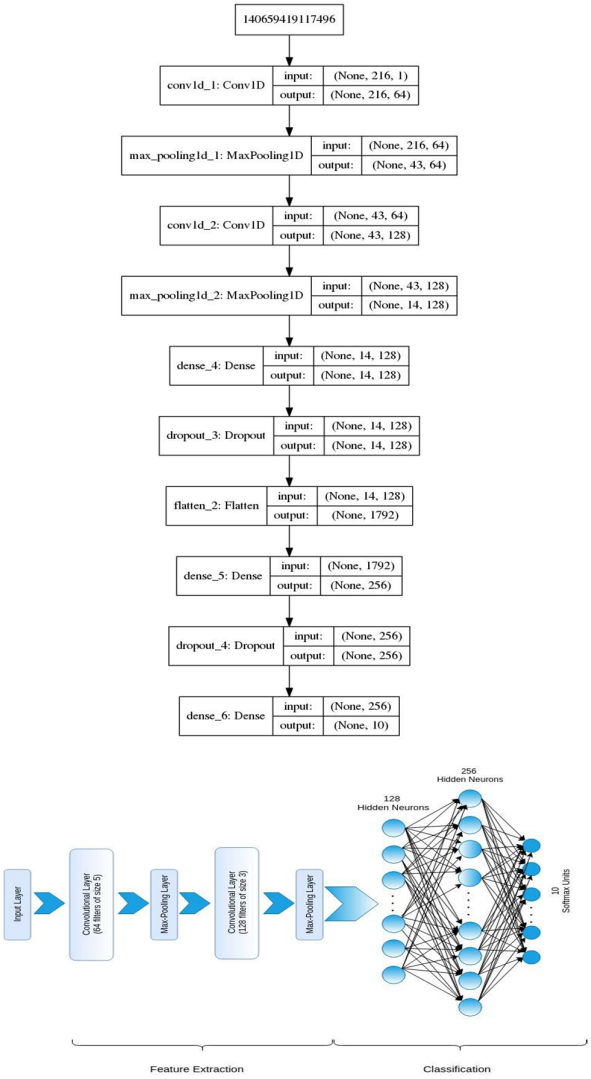

Figure 2: *Baseline architecture of CNN.*

4 Multi-Modal Analysis

Recently, we agree that the deep learning algorithms have successfully addressed problems in various fields, such as image classification, machine translation, speech recognition, text-to-speech generation and other machine learning related areas [30] [31] [32]. Similarly, substantial improvements in performance have been obtained when deep learning algorithms have been applied to statistical speech processing [28]. Even though

various types of deep learning methods have been applied, this problem is still considered to be challenging for several reasons; first, the scarcity of emotion tagged data for training deep neural models and second, the characteristics of emotions must be learned from low-level speech signals. However, feature-based models display limited skills when applied to this problem.

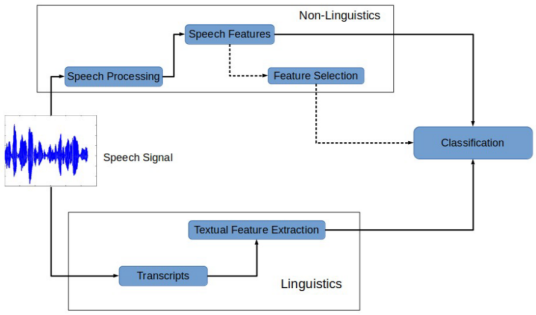

Figure 3: Multi-modal architecture of Audio (non-Linguistic) and Text (linguistic) models for speech emotion classification

In order to overcome these limitations, we have developed a model (as shown in Figure 3) that uses high-level text transcription, as well as low-level audio signals, to utilize the information contained within low-resource datasets to a greater degree. The emotional content of speech is clearly indicated by the emotion words contained in a sentence [29], such as "lovely" and "fantastic," which carry strong emotions compared to generic (non-emotion) words, such as "person" and "day." Thus, we hypothesize that the speech emotion recognition model will be benefit from the incorporation of high-level textual input with the low-level audio features. Moreover, this multimodal approach encodes both audio and textual information simultaneously via a dual recurrent encoder.

4.1 Audio-Only Encoder (AoE)

We have built an *Audio-only Encoder* (AoE) to predict the emotional class of a given audio signal based on only audio features. Once Mel-frequency cepstral coefficients (MFCCs) features have been extracted from an audio signal, a subset of the sequential features is fed into the recurrent neural networks (RNN), which is composed of gated recurrent units (GRUs), which in turn leads to the formation of the network's internal hidden state $\boldsymbol{h}_t$ to model the time series pattern. The updates of the hidden state is performed with the input data $\mathbf{x}_t$ and the hidden state output of the previous time step $\boldsymbol{h}_{t-1}$, which is basically the main working principle of a recurrent neural network. The present time hidden state $\boldsymbol{h}_t$ can be mathematically modeled as following:

$$\boldsymbol{h}_t = \boldsymbol{f}_w\,(\boldsymbol{h}_{t-1}, \boldsymbol{x}_t), \qquad (1)$$

where $\boldsymbol{f}_w$ is a function which imitates the function of an RNN with weight parameter $\boldsymbol{w}$, $\boldsymbol{h}_t$ represents the hidden state at $\boldsymbol{t}^{th}$ time step, and $\mathbf{x}_t$ represents the $\boldsymbol{t}^{th}$ MFCC features in $\mathbf{x} = \{\mathbf{x}_{1:\,ta}\}$. After encoding the audio signal $\mathbf{x}$ with the RNN, the last hidden state of the RNN, $\boldsymbol{h}_{ta}$, is considered to be the representative vector that contains all of the sequential audio data. In this model, we have also incorporated the prosody features of an audio signal. The prosody of an audio signal is characterized by the inherent pattern of stress and intonation in a language. We have incorporated this characteristic in order to better classify the emotional content in an audio signal.

However, in order to implement, we have developed a prosodic feature vector, $\mathbf{p}$, which models the prosody of audio files. Then, we have concatenated the last hidden state vector, $\boldsymbol{h}_{ta}$, with the prosodic feature vector, $\mathbf{p}$, in order to generate more informative vector representation of the audio signal. We denote this more informative vector as $\mathbf{e}$, where $\mathbf{e} = concat\{\boldsymbol{h}_{ta}, \mathbf{p}\}$. The MFCC and the prosodic features are extracted from the audio signal using the openSMILE toolkit [26] and $\mathbf{x}_t$ has 39 and $\mathbf{p}$ has 35 MFCC features.

Finally, the emotion class is predicted by applying the *softmax* function to the vector $\mathbf{e}$. For a given audio sample $\mathbf{i}$, we assume that $\boldsymbol{y}_i$ is the true label vector, which contains all zeros but contains a one at the correct class, and $\boldsymbol{y}'_i$ is the predicted probability distribution from the *softmax* layer. The training objective then takes the following form:

$$\boldsymbol{y}'_i = \boldsymbol{e}^{\boldsymbol{T}}\boldsymbol{M} + \boldsymbol{b} \qquad (2)$$

$$\partial = -log \prod_{i=1}^{N} . \sum_{c=1}^{C} y_{i,c}\, log\ (\boldsymbol{y}'_{\boldsymbol{i,c}})$$

where, **e** is the calculated representative vector of the audio signal with dimensionality $\mathbf{e} \in \mathbf{R}^d$. The $\mathbf{M} \in \mathbf{R}^{d \times C}$ and the bias **b** are learned model parameters, **C** is the total number of classes, and **N** is the total number of samples used in training.

4.2 Text-Only Encoder (ToE)

Apart from the audio, we tried to use the textual information as another modality in predicting the emotion class of a given signal. To use textual information, the speech transcripts are tokenized and indexed into a sequence of tokens using the Natural Language Toolkit (NLTK) [27]. Each token is then passed through a word embedding layer that converts a word index to a corresponding 300-dimensional vector that contains additional contextual meaning between words. The sequence of embedded tokens is fed into a *Text-only Encoder* (ToE) in such a way that the audio MFCC features are encoded using the AoE represented by equation 1. In this case, $\mathbf{x}_t$ is the t^{th} embedded token from the text input. Finally, the emotion class is predicted from the last hidden state of the text-RNN using the *softmax* function. We use here the same training objective as we adopted for the AoE model, and the predicted probability distribution for the target class is as follows:

$$\boldsymbol{y}_i' = softmax\left(\boldsymbol{h}_{last}{}^{T}\boldsymbol{M} + \boldsymbol{b}\right) \quad (3)$$

where $\boldsymbol{h}_{last}$ is the last hidden state of the text-RNN, $\boldsymbol{h}_{last} \in \mathbf{R}^d$, and $\mathbf{M} \in \mathbf{R}^{d \times C}$ and the bias **b** are learned model parameters. The lower part of Figure 3 and Figure 4 indicates the architecture of the ToE model.

4.3 Merged Recurrent Encoder (MRE)

In order to obtain the benefits from both the audio and text modes, we present an architecture called the merged recurrent encoder (MRE) to overcome the limitations of existing approaches. In this study, we consider multiple modalities, such as MFCC features, prosodic features and transcripts, which contain sequential audio information, statistical audio information and textual information, respectively. These types of data are the same as those used in the AoE and ToE cases.

However, the MRE model employs two RNNs to encode data from the audio signal and textual inputs, independently. The audio-RNN encodes MFCC features from the audio signal using equation 1. The last hidden state of the audio-RNN is concatenated with the prosodic features to form the final vector representation **e**, and this vector is then passed through a fully connected neural network layer to form the audio encoding vector **A**. On the other hand, the text-RNN encodes the word sequence of the transcript using equation 1. The final hidden states of the text-RNN are also passed through another fully connected neural network layer to form a textual encoding vector **T**. Finally, the emotion class is predicted by applying the *softmax* function to the concatenation of the vectors **A** and **T**. We use the same training objective as the AoE model, and the predicted probability distribution for the target class is as follows:

$$\mathbf{A} = \boldsymbol{g}_{\boldsymbol{\theta}}(\mathbf{e}),\ \mathbf{T} = \boldsymbol{g}'_{\boldsymbol{\theta}}(\mathbf{h}_{last})$$

$$\boldsymbol{y}_i' = softmax\left(concat\left(\boldsymbol{A}, \boldsymbol{T}\right)^{T}\boldsymbol{M} + \boldsymbol{b}\right) \quad (4)$$

where $\boldsymbol{g}_{\boldsymbol{\theta}}$, $\boldsymbol{g}'_{\boldsymbol{\theta}}$ is the feed-forward neural network with weight parameter $\boldsymbol{\theta}$, and **A**, **T** are final encoding vectors from the audio-RNN and text-RNN, respectively. $\mathbf{M} \in \mathbf{R}^{d \times C}$ and the bias **b** are learned model parameters.

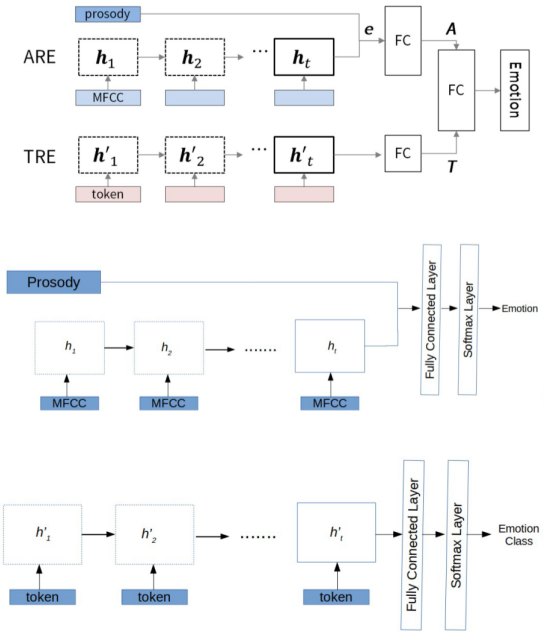

Figure 4: *Merged Recurrent Encoder. (The upper part shows AoE, which encodes audio signals and the lower part shows ToE, which encodes textual information).*

Dataset: We evaluated our multi-modal model on the Interactive Emotional Dyadic Motion Capture (IEMOCAP) [34] dataset. This dataset was collected theatrical theory in order to simulate natural dyadic interactions between actors. We use categorical evaluations with majority agreement. We use only four emotional categories viz. happy, sad, angry, and neutral to compare the performance of our model with other research using the same categories. The IEMOCAP dataset includes five sessions, and each session contains utterances from two speakers (one male and one female). This data collection process resulted in 10 unique speakers. For consistent comparison with previous work, we merge the excitement dataset with the happiness dataset. The final dataset contains a total of 5531 utterances (1636 happy, 1084 sad, 1103 angry and 1708 neutral).

Feature Extraction: In order to extract speech information from audio signals, we use MFCC values, which are widely used in analyzing audio signals. The MFCC feature set contains a total of 39 features, which include 12 MFCC parameters (1-12) from the 26 Melfrequency bands and log-energy parameters, 13 delta and 13 acceleration coefficients. The frame size is set to 25 ms at a rate of 10 ms with the Hamming function. According to the length of each wave file, the sequential step of the MFCC features is varied. To extract additional information from the data, we also use prosodic features, which show effectiveness in affective computing. The prosodic features are composed of 35 features, which include the F0 frequency, the voicing probability, and the loudness contours. All of these MFCC and prosodic features are extracted from the data using the OpenSMILE toolkit [26].

Setup Details: Among the variants of the RNN function, we use GRUs as they yield comparable performance to that of the LSTM and include a smaller number of weight parameters [28]. We use a max encoder step of 750 for the audio input, based on the implementation choices presented in [33] and 128 for the text input because it covers the maximum length of the transcripts. The vocabulary size of the dataset is 3,747, including the "_UNK_" token, which represents unknown words, and the "_PAD_" token, which is used to indicate padding information added while preparing mini-batch data. The number of hidden units and the number of layers in the RNN for each model (AoE, ToE, MRE) are selected based on extensive hyper-parameter tuning.

5 Experiments & Results

This section discusses the experiments performed in this study to classify emotion class and gender of the input audio data using the FFNN and CNN based deep learning models as described in the above sections. In order to have a comparative discussion, we restricted ourselves to 100 epochs for both the models.

The datasets used to train both of these networks already have been discussed in the Section 3.1. We encourage the readers to consult that section to have a detailed idea about the datasets. We have merged the audio files from the two datasets, SAVEE and RAVDESS, to produce the raw data. There are approximately 1900 audio files after merging. However, we were not able to use all the audio files to train our networks as the emotion classes of the two datasets were not identical. The emotion classes reported in SAVEE database are *anger, disgust, fear, happiness, sadness, surprise* and *neutral* whereas the emotion classes in RAVDESS database are *neutral, calm, happy, sad, angry, fearful, disgust* and *surprised*.

As we know *neutral* emotion does not specifically portray any emotion specific feature, we discarded all the sentences belong to *neutral* class from our raw dataset. Furthermore, we considered only 5 main classes of emotions namely, *calm, happiness, sadness, fear* and *anger*. As a result, we have approximately 1200 sentences in our raw database. In case of training and testing our models, we need to split our raw dataset, which is described in the previous section, to form the training and test data. We have taken approximately 80% of the raw dataset as training and the remaining 20% as test data to evaluate our models. The performances of the *neural* network models on this training and test set have been demonstrated in the following sections.

5.1 Results of FNNN Model

The performance of the deep Feed Forward Neural Network is measured in terms of training vs. test accuracy graph, training vs. test loss graph and two confusion matrices, one for emotion classification and another for gender classification.

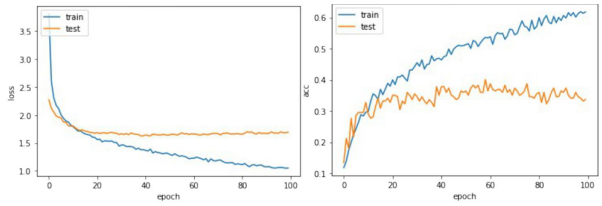

Figure 5: *Training vs. Test Loss and Training vs. Test Accuracy graph for FFNN model.*

It is very much clear from the above graphs in Figure 5 that the model did not perform very well; in fact it is very clear that over-fitting happened in this case. In order to investigate the reasons, we reported the confusion matrices. Figure 6 represents the overall confusion matrix for FFNN model.

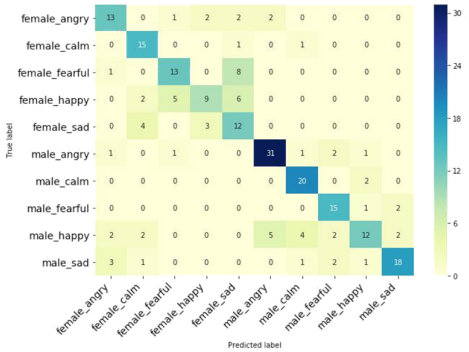

Figure 6: *Confusion matrix for Emotion and Gender classification using FFNN*

If we analyze the confusion matrix, we can conclude that the overall performance of the FFNN model is not good. We can see in the confusion matrix that *male_fearful, male_happy, male_sad* have been misclassified as *male_angry*. In addition, a considerable amount of *female_sad* and *male_sad* labels have been misclassified as *female_happy* and *male_fearful*, respectively. Overall, the accuracy achieved by this model is 40.82%. Figure 7 represents the confusion matrix only for gender labels that is *male* and *female*. The classification performed by this model for gender labels is far better than overall classification.

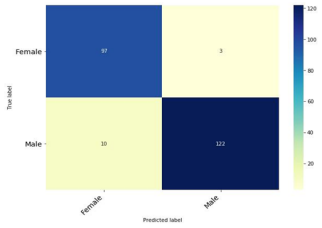

Figure 7: *Confusion matrix for Gender classification (Only) using FFNN model.*

5.2 Results of CNN Model

Unlike the FFNN model, CNN model did not suffer from over-fitting as can be seen in Figure 8. On the other hand, Figure 9 and Figure 10 represent the confusion matrices for overall classification and gender classification, respectively. If we analyze the confusion matrix for the overall classification, it surely outperforms our FFNN model as the misclassification rate is much lower in the case of CNN. Misclassification of *female_fearful* as *female_sad* is the only noticeable misclassification that happened in the whole confusion matrix. The accuracy for the overall classification is approximately 68.38%. This accuracy was achieved by running the training algorithm for 2000 epochs

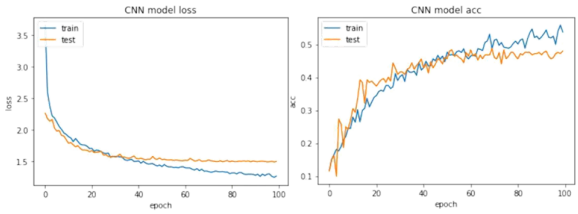

Figure 8: *Training vs. Test Loss and Training vs. Test Accuracy graph for CNN model*

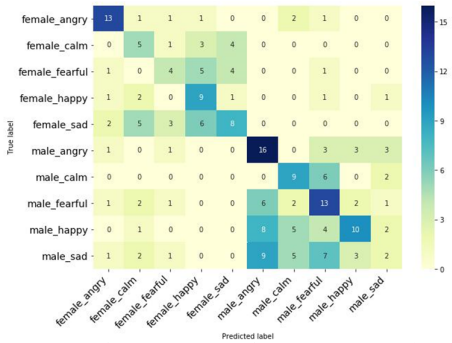

Figure 9: *Confusion matrix for Emotion and Gender classification using CNN.*

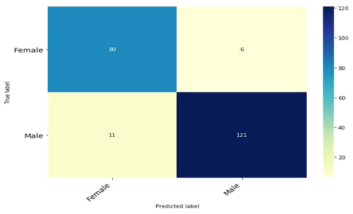

Figure 10: *Confusion matrix for Gender classification using CNN model.*

5.3 Results of Multi-Modal Model

The MRE model combines the benefits of AoE and ToE models and it receives approval if we see the curves of loss as well as accuracies over training and validation set, respectively. Moreover, the performances of individual models justify the multi-modal effects when used in combination. Table 1 shows the accuracies of various models.

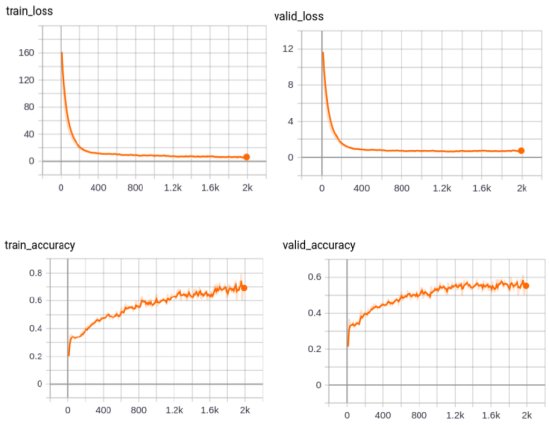

Figure 11: *Loss and accuracy curves for AoE model with (8:0.5:1.5) splitting into training, development and test data*

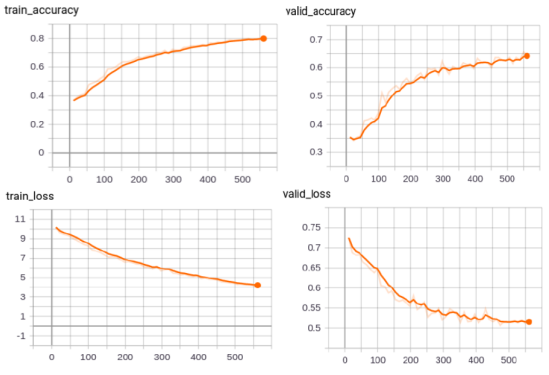

Figure 12: *Loss and accuracy curves for ToE model with (8:0.5:1.5) splitting into training, development and test data*

Model	Accuracy on Validation Set	Accuracy on Test Set
Model 1 (FFNN)	43.02%	40.82%
Model 2 (CNN)	64%-68.38%	52%-54.32%
Model 3.1 (AoE)	59.42%	59.5%
Model 3.2 (ToE)	63.58%	67.27%
Model 3.3 (MRE)	74.12%	74.64%

Table 1: Comparative analysis of accuracies of the various models on validation and test sets of IEMOCAP data.

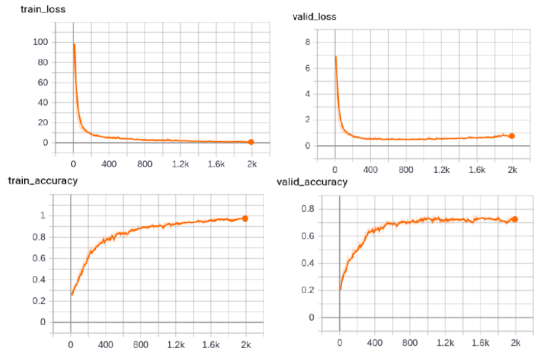

Figure 13: *Loss and accuracy curves for MRE model with (8:0.5:1.5) splitting into training, development and test data*

5.4 Error Analysis

We analyze the predictions of the AoE, ToE, and MRE models. Figure 14 shows the confusion matrix of each model. The ARE model (as shown in Figure 14(a)) incorrectly classifies most instances of *happy* as *neutral* (43.51%); thus, it shows reduced accuracy (35.15%) in predicting the *happy* class. Overall, most of the emotion classes are frequently confused with the *neutral* class. This observation is in line with the findings of [25], who noted that the *neutral* class is located in the center of the activation-valence space, complicating its discrimination from the other classes.

Interestingly, the ToE model (as shown in Figure 14(b)) shows gains in predicting the *happy* class when compared to the AoE model (35.15% to 75.73%). This result seems plausible because the model can benefit from the differences among the distributions of words in *happy* and *neutral* expressions, which gives more emotional information to the model than that of the audio signal data. On the other hand, it is unexpected that the ToE model incorrectly predicts instances of the *sad* class as the *happy* class 16.20% of the time, even though these emotional states are being present at oppose to one another.

The MRE model (as shown in Figure 14(c)) compensates for the weaknesses of the previous two models (AoE and ToE) and benefits from their strengths to a surprising degree. The values

arranged along the diagonal axis show that all of the accuracies of the correctly predicted class have increased. Furthermore, the occurrence of the incorrect "*sad-to-happy*" cases in the ToE model is reduced from 16.20% to 9.15%.

(a) AoE (b) ToE (c) MRE

Figure 14: *Confusion matrices of AoE, ToE and MRE models*

6 Speech Emotion Lexicon

Not only the classification of speeches into different emotion categories, but one of our partial objectives was also to explore the possibility of generating a speech emotional database consisting of emotional utterances also. Therefore, we have developed a speech emotion lexicon containing utterances of different emotional categories. However, it is difficult to simultaneously generate both *male* and *female* voices in a text-to-speech system and thus, we limited ourselves to synthesize only '*male*' voice in this present attempt. For this very reason, we selected our dataset with only *male* speakers

6.1 Pre-Processing

All audio i.e. our WAV files were resampled and filtered by an antialiasing FIR low pass filter to have frequency rate of 44.1 kHz prior to any processing. Silences and non-voiced parts at the start and the end have been removed from the files. The next step of developing the speech lexicon is to classify the emotion of each of the WAV files employing the best classifier.

6.2 Transcript Generation

To make our model robust, we made sure that we can build the lexicon from the WAV files which do not have any transcript associated with it. We made use of IBM Speech to Text API [1] service to obtain the transcript of a given WAV file.

6.3 POS Tagging and Word Segmentation

At this stage, we tag the words based on their part of speech and also segment the words of the audio using the transcript generated by the Text-to-Speech service. At first, Parts of Speech (POS) tagging of all the segmented words has been performed and we discarded the *proper nouns* (names, places etc.) as it conveys very little emotional features than *adjectives* or *adverbs* etc.

After this, we segment the words based on their start and end time in the audio files. We get the start and times of all the words from the results obtained from Text-to-Speech service described in the previous sections. Finally, we use this information to extract the words using Pydub[2], a Python library for audio processing.

6.4 Emotion Word Lexicon

The first column of the lexicon represents the words that have been spoken and second column represents the gender and third column represents the emotion in which the corresponding word has been spoken. The last column represents the location of the WAV file containing the utterance of the corresponding word in the specified emotion. We have also grouped same words spoken in different emotions. Presently, the lexicon contains only 1K words in 5 different emotion categories. Three native speakers have evaluated the emotional utterances and an agreement score of pair wise kappa k=0.92 was found. The minute disagreement was happened due to the segmentation and such words have been discarded from the lexicon.

7 Conclusions

In the present task, two classification models based on deep neural network, one using normal Feed-forward Neural Network (FFNN) and another using Convolutional Neural Network (CNN) architecture has been implemented and also a comparative study between these two models has been reported. Among the two models it has been shown that CNN model outperformed the FFNN model. The models have been developed from a training set which

[1] https://cloud.ibm.com/apidocs/speech-to-text

[2] https://pydub.com/

consists of only English language. It will be an interesting study to apply other languages to train the model and compare the performances for the same.

In addition to that, we have implemented a multi-modal version of the deep neural model using Recurrent Neural Network, in order to improve the classifier system. We have used features of both audio and text and merged them into a single framework to investigate the effectiveness of the system. It is observed that the performance of the multi-modal system is far better that the FFNN or CNN based system in classifying emotions.

As we have mentioned earlier that we have developed a database which contains emotional utterances. However, the size of the database is not very large, we had a limited number of test samples and hence it affected the development of the proposed database.

Acknowledgement

The work is supported by the SERB sponsored IMPRINT-II Project, DST, Government of India.

References

[1] Dong Yu and Li Deng., Automatic Speech Recognition. Springer,2016.

[2] Lawrence R Rabiner and Biing-Hwang Juang. Fundamentals of speech recognition, volume 14. PTR Prentice Hall Englewood Cliffs, 1993.

[3] Picard, R.W. Affective Computing. M.I.T. Press, Cambridge, MA. 1997.

[4] Picard R. W., Healey J., Affective Wearables, Personal Technologies Vol 1, No. 4, pages 231-240. 1997.

[5] Picard R.W., Affective Computing for HCI. In Proc. of the 8th International Conference on Human- Computer Interaction: Ergonomics and User Interfaces-Volume I. Lawrence Erlbaum Associates, Inc. 1999.

[6] Picard R.W., Vyzas E., Healey J. Toward Machine Emotional Intelligence -Analysis of Affective Physiological State. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol 23 No. 10, 2001.

[7] Norman, D.A. Emotional Design: Why we love (or hate) everyday things. Basic Books. 2003.

[8] F. Dipl and T. Vogt, "Real-time Automatic Emotion Recognition from Speech", 2010.

[9] S. Lugovic, I. Dunder, and M. Horvat, Techniques and applications of emotion recognition in speech, 2016 39th Int. Conv. Inf. Commun. Technol.797979 Electron. Microelectron. MIPRO 2016 - Proc., November 2017, pages 1278–1283, 2016.

[10] B. Schuller, G. Rigoll, and M. Lang, "Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine - belief network architecture," Acoust. Speech, Signal Process., vol. 1, pages 577–580, 2004.

[11] R. W. Picard, "Affective Computing for HCI," In HCI (1), pages 829– 833, 1999.

[12] F. Ren, "From cloud computing to language engineering, affective computing and advanced intelligence," International Journal of Advanced Intelligence, vol. 2(1), pages 1–14, 2010

[13] Stuart Jonathan Russell, Peter Norvig, John F Canny, Jitendra M Malik and Douglas D Edwards. Artificial intelligence: a modern approach, volume 2. Prentice hall Upper Saddle River, 2003.

[14] H. Cao, R. Verma, and A. Nenkova, "Speaker-sensitive emotion recognition via ranking: Studies on acted and spontaneous speech," Comput. Speech Lang., vol. 28, no. 1, pages 186–202, Jan. 2015.

[15] L. Chen, X. Mao, Y. Xue, and L. L. Cheng, "Speech emotion recognition: Features and classification models", Digit. Signal Process., vol. 22, no. 6, pages 1154–1160, Dec. 2012.

[16] J. P. Arias, C. Busso, and N. B. Yoma, "Shape-based modeling of the fundamental frequency contour for emotion detection in speech," Comput. Speech Lang., vol. 28, no. 1, pages 278–294, Jan. 2014.

[17] S. Wu, T. H. Falk, and W.-Y. Chan, "Automatic speech emotion recognition using modulation spectral features," Speech Commun., vol. 53, no. 5, pp. 768–785, May 2011.

[18] J. Rong, G. Li, and Y.-P. P. Chen, "Acoustic feature selection for automatic emotion recognition from speech,"Inf. Process. Manag., vol. 45, no. 3, pp. 315–328, May 2009.

[19] C.-H. Wu and W.-B. Liang, "Emotion Recognition of Affective Speech Based on Multiple Classifiers Using Acoustic-Prosodic Information and Semantic Labels,"

IEEE Trans. Affect. Comput., vol. 2, no. 1, pp. 10–21,Jan. 2011.

[20] Changqin Quan, Fuji Ren, "An Exploration of Features for Recognizing Word Emotion", Proceedings of the *23rd International Conference on Computational Linguistics (Coling 2010)*, pages 922–930, Beijing, August 2010

[21] Sanaul Haq, Philip JB Jackson, and J Edge. Speaker-dependent audio-visual emotion recognition. In AVSP, pages 53–58, 2009.

[22] Ekman, P., "Universals and cultural differences in facial expressions of emotion", Nebraska Symposium on Motivation, pages 207-283, 1972.

[23] Zeng, Z., Pantic, M., Roisman, G.I. and Huang, T.S., "Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions", IEEE Trans. PAMI, 31(1), pages 39-58, 2009.

[24] Livingstone SR, Russo FA (2018) The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PLoS ONE 13(5): e0196391.

[25] Michael Neumann and Ngoc Thang Vu, "Attentive convolutional neural network based speech emotion recognition: A study on the impact of input features, signal length, and acted speech," Proc. Interspeech 2017, pp. 1263– 1267, 2017.

[26] Florian Eyben, Felix Weninger, Florian Gross, and Bjorn Schuller, "Recent developments in opensmile, the ¨ munich open-source multimedia feature extractor," in Proceedings of the 21st ACM international conference on Multimedia. ACM, 2013, pp. 835–838.

[27] Steven Bird and Edward Loper, "Nltk: the natural language toolkit," in Proceedings of the ACL 2004 on Interactive poster and demonstration sessions. Association for Computational Linguistics, 2004, p. 31.

[28] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," arXiv preprint arXiv:1412.3555, 2014.

[29] Linhong Xu, Hongfei Lin, Yu Pan, Hui Ren, and Jianmei Chen, "Constructing the affective lexicon ontology," Journal of the China Society for Scientific and Technical Information, vol. 27, no. 2, pp. 180–185, 2008.

[30] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in Advances in neural information processing systems, 2012, pp. 1097–1105.

[31] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio, "Neural machine translation by jointly learning to align and translate," arXiv preprint arXiv:1409.0473, 2014.

[32] Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, et al., "Deep speech 2: End-to-end speech recognition in English and mandarin," in International Conference on Machine Learning, 2016, pp. 173–182.

[33] Michael Neumann and Ngoc Thang Vu, "Attentive convolutional neural network based speech emotion recognition: A study on the impact of input features, signal length, and acted speech," Proc. Interspeech 2017, pp. 1263– 1267, 2017.

[34] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," Language resources and evaluation, vol. 42, no. 4, pp. 335, 2008.