Conversational AI for Companionship

by

Zakaria Khan
18101404
Salauddin Md Akash
18101494
Afia Mobassira Jamima
18101016
Isfar Hasan Ishan
18101042

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science

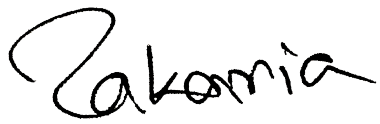
Department of Computer Science and Engineering
Brac University
May 2022

Declaration

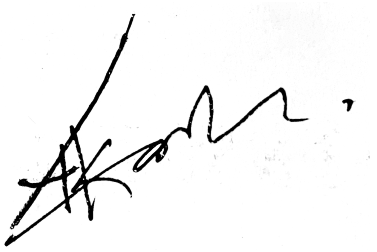
It is hereby declared that

1. The thesis submitted is our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

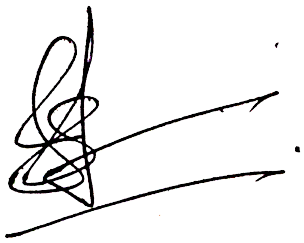


Zakaria Khan
18101404



Salauddin Md Akash
18101494

Afia Mobassira Jamima
18101016



Isfar Hasan Ishan
18101042

Approval

The thesis titled "Conversational AI for Companionship" submitted by

1. Zakaria Khan(18101404)
2. Salauddin Md Akash(18101494)
3. Afia Mobassira Jamima(18101016)
4. Isfar Hasan Ishan(18101042)

Of Spring, 2022 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on May 29, 2022.

Examining Committee:

Supervisor:
(Member)

Dr. Md. Golam Rabiul Alam
Associate Professor
Department of Computer Science and Engineering
Brac University

Co-Supervisor:
(Member)

Shaily Roy
Lecturer
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)

Sadia Hamid Kazi, PhD
Chairperson and Associate Professor
Department of Computer Science and Engineering
Brac University

Abstract

The conversational style of a human is estimated by humor, personality, voice tone, etc. These characteristics are necessary for virtual assistants that are artificially intelligent for conversation. This research recommends an intelligent system capable of holding an appropriate human-like dialogue, including the emotion and personality of a specific character. To draw the pattern of the attributes of specified emotion, a method can be used to transmit voice tone. In order to determine all the necessary characteristics mentioned above, the goal is to use different categories of machine learning models. Since the pattern of conversation, style varies from one individual to another and geographically, our goal is to create a virtual assistant. In addition, a conversational model will be applied to it. It will read the category of emotions(exclamation, assertion, negation, interrogation) of human beings and respond accordingly. Many methodologies are being utilized to predict sentiments through AI and react accordingly. IVA is one of them but with its limitations and boundaries. Therefore, this paper comes with several methodologies that can be used alongside IVA; such as HMM, GMM, SVM, NLU, BoAW, BERT, etc. These algorithms and methodologies will help to predict the sentiments used in a context and precisely predict the outcome of an inquiry. To sum it up, this thesis aims to create a conversational AI for companionship, which will create an emotional bridge between itself and the user.

Keywords: Self Disclosure; Machine Learning; Virtual Assistant; Sentiment; NLP; Mercantile; Interlocutor

Acknowledgement

Firstly, all praise to the Great Allah for whom our thesis have been completed without any major interruption.

Secondly, to our supervisor, Dr. Md. Golam Rabiul Alam sir for guiding us throughout the process of conducting this thesis. Without his direct supervision, it would not have been possible to carry on such research and submit a comprehensive thesis.

Thirdly, to our co-advisor Ms. Shaily Roy, for her kind support and advice in our work. She helped us whenever we needed help.

We want to express our gratitude to the unmet peoples on the internet for their selfless help and guidance throughout the research period. It would not have been easy to accomplish such a task without their help. We are also grateful to our family and friends who have contributed a lot in accomplishing this task.

Lastly, We would like to thank all our institute's teachers who have helped us gain valuable knowledge and insights throughout the four years of our under-graduation life.

Table of Contents

List of Figures

List of Tables

Nomenclature

The next list describes several symbols & abbreviation that will be later used within the body of the document

ASR Automated Speech Recognition

CNN Convolution Neural Network

DCT Discrete Cosine Transform

FFT Fast Fourier Transform

IVA Intelligent Virtual Assistant

$MFCC$ Mel-frequency Cepstral Coefficients

$MFEC$ Mel-frequency Energy Coefficients

ML Machine Learning

OTT Over The Top

RNN Recurrent Neural Network

$STFT$ Short-time Fourier Transform

TTS Text To Speech

Chapter 1

Introduction

1.1 Introduction

Human faces have inscrutable countenances while having conversations. Similarly, the voice of a human being also consists of resonance, relaxation, rhythm, and pacing. If we want to propagate an artificially intelligent virtual assistant, the authentication of human voice impersonation is the key aspect. The goal for creating an intelligent system is that they should be capable of having a human-like conversation. A human-like personal assistant must be capable of acknowledging the sentiments of a person, such as intuition, affection, sympathy, ecstasy, admiration, anger, embarrassment, etc. Several methodologies are being used for human assistance. Intelligent Virtual Assistant (IVA) is one of those systems that emulates a human and delivers personalized responses through cognitive computing and analytics[39]. A well-configured IVA is capable of serving as virtual assistance, customer service representative, and contributing to administrative tasks. Several OTT platforms have used IVA for commercial benefits in recent years, notably. In today's era, the world's biggest tech giants such as Apple, Amazon, Microsoft, Google have developed commercial IVAs like Siri, Alexa, Cortana, Google assistant. Even though the commercial benefits consumed by these tech giants are abundant, using IVA technology disregards the personality traits of a normal human being. It is not capable of understanding the context of a situation; as a result, the communication gap between humans and virtual assistants appears explicitly. People find IVA technology unrealistic because they don't own emotional attachments and personality traits. They also do not reciprocally self-disclose themselves while having a conversation similar to with typical chatbots. This study will address all the lacking of traditional IVA technology and generate an IVA capable of having a human-like conversation by understanding the context of a situation. They can bring off transmission tasks with some level of compassionate intellect and human emotions so that it doesn't seem artificial. The characteristics of this human-like virtual assistant must be natural, just like a human who must be compassionate while interacting with other people. Hu-

man beings are reckoned as the most passionate; besides, what makes them different and more relevant is how they interact, socialize, communicate, share, and convey emotional support to other people. Self-disclosure is a key characteristic to make people less stressed and worried practically. When someone elicits their oppression which encircles sentiments, emotional assistance from a partner through confabulation can ennoble that person's psychological sequel. This research aims to surpass the limitations and drawbacks that the current IVA has in terms of recognizing the sentiment of an AI-based conversation and responding accordingly.

1.2 Problem Statement

The idea behind Intelligent Virtual Assistants is to generate an artificial conversational, personal assistant who must have the characteristics of a human to make communication more intuitive and realistic. Unfortunately, the IVA technology is insufficient to impersonate actual human characteristics. Human voices have a wide range of factors to express the novelty of an event. When humans communicate in real life, non-verbal cues convey different emotions. It is easier to parse a sentence's meaning from facial expressions and the tone of voice. The current virtual assistants are not that advanced to have a human-like conversation that offers quality and diction. IVA falls short in many areas, particularly emotions and empathy. They can only understand and can respond to concise questions such as "weather updates", "schedule fixing", "traffic updates", "fixing meeting dates" etc. Although the current IVA's successfully dig up the offensive words, their responses to those questions are incomprehensive in manner. Their answer does not depend on emotion, it depends on the question as the voice-recognition model transmutes speech to text which follows the digitized model sounds. If asked, "What is the time in New York?", we will get the desired answer concisely. To have a human-like conversation, the limitation of the current IVA technology must be addressed.

Emotions are sensations, feelings, and physiological states that help determine how to distinguish human nature. It is complicated, affected by personality and actions. When someone slams a car door on your face, you probably feel angry. Similarly, you feel delighted and honored when someone holds the door open for you. Understanding the context of a situation plays a vital role in human emotions. During a conversation, understanding the emotional ambiance of the next person is considered to be the most effective way of generating an empathetic connection. In addition, variants of grammar, usage of words such as "umm", "uhh" also helps to create substantive communion. Even though IVA technology has progressed significantly in the modern period, it lacks in many areas, such as understanding the intricacy of human behavior. There are popular mercantile IVA's like Google Assistant of

Google, Cortana of Microsoft, Siri of Apple, Alexa of Amazon. These assistants are booming day by day, but they focus on empathetic manners, they can't even compete with a family pet. The AI assistant might not recognize the difference between an inquiry uttered by the user and their severe frustration or sarcasm, making them incapable of responding accordingly. In a crisis, if a person wants to communicate to reduce their stress and worry, the virtual assistant must provide emotional support to change the people's relational perception by demonstrating social cues.
Furthermore, the current IVA's cannot distinguish which conversations are private and which are shareable with others. As a result, the integrity of the discussion is always questioned. Maintaining the secrecy of a conversation with other users who have access to the same system is sometimes challenging. If a person shares their own experiences while communicating with another stressed individual, this type of self-disclosure is very significant to create a reliable connection between the speakers[40]. This sort of mutual vulnerability is constructive to calm the distress of the characteristic. While self-disclosing, if no emotional support was offered, then the outcome will be more repulsive. The assurance of when and what they should self-disclose must be maintained studiously.
Personality in IVA's is important because current IVA does not have personality. When human beings interact, it's imperative to have a similar nature to relate and connect with others To have a proper human-like conversation, the personality of IVA must reflect compassion, reliability, respectfulness, and honesty. Personality will greatly impact how people engage with IVAs, giving us the impression that we have a real-life partner.[39].
The limitations mentioned above are very intricate; researchers are trying to visualize the solutions to address those boundaries. Our research purpose and contribution is not only to generate a norm capable of identifying the sentiments of a person but also to hold a comprehensive human-like conversation with compassion.

1.3 Research Objectives

The progression of IVAs of this generation is quite salient if we glance at the chronic form of IVAs. But still, there are numerous sectors where IVAs can be exalted. The main goal of our research is to make the system more tangible and specific. To emerge with a new complexion of IVA where emotion awareness and novelty will be the main feature, and the main goal is to objectify a person's personality. To impersonate a person's character, we need to nourish enough audio of that person. Our crucial research objectives will be as follows:

- Take the unprocessed audio input as well as process the audio.

- Identify the emotional characteristics from that tone of the audio
- Convert that audio into text (written) format for further analysis
- Process the generated text and identify what the user wants
- Based on identified emotions and the user's query, generate appropriate responses.
- Generate the voice response with a proper sentiment from the generated text response.

Chapter 2

Literature Review

The idea of recognizing emotion from the speech is not new, but several limitations need to be sorted out. To overcome the limitations mentioned in the problem statement that IVA has, several methodologies and representations can be used to recognize sentiments in any sort of speech. The long-established approach recognizing emotions from audio uses Modeling, Textual Features, Audio Features Before using any kind of algorithm to detect emotion, the first task is to process the audio and gather data from it[24][15]. Then, the audio data can be processed and converted to a waveform graph. In a paper published in 2016[12], the Hidden Markov Model(HMM) makes the waveform from the audio input. The generated waveform can be further processed and retrieved features from that graph. For emotional recognition, the major attributes from the audio that need to be considered are pitch, volume, tone, voice condition, and speed of the speech. In past works, most of the systems have used temporal features and statics features. The HMM model can also classify the modeled frame-based feature, described in the paper[6]. It is also clarified that, for statistical characteristics, the Support Vector Machine (SVM) can be used to build a classifier. As stated in that paper, the Gaussian Mixture Model (GMM) can also be used, as stated in that paper[6]. The paper of Pathak and Kulkarni described some steps of processing the audio, which are- Signal Processing, Emotional Feature Extraction, Classification, Training, and Testing[7]. For the feature extraction, the paper used the LPC Analysis algorithm. The processes that are expressed are described further below.

NLP is the most relevant and fundamental methodology for artificial virtual assistant systems. When a system receives audio input, the task is to process the audio to make it understandable. The input is predominantly speech voice, which is natural language. To process it, Natural Language Processing(NLP) algorithms can help proceed further. Our research purpose reflects the research paper[24], which is based on a competition arranged in 2017 by the Amazon Alexa team, where 15 university students from around the world participated. The key purpose of the

competition was to find out the improved technology behind social bots. In that competition, most university students used the Natural Language Understanding (NLU) methodology to process the audio input. According to Ram and Prasad[24], Most of the participant teams of that competition used some advanced algorithms, such as StanfordCoreNLP, Google NLP, SpaCy, Alexa's ASK NLU, etc. Models from Automatic Speech Recognition (ASR) can be used to analyze the audio.Models from ASR, like- Acoustic model, Language model, might gratify to get the desired result. The acoustic model turns the sounds in the form of phonetics representation. On the contrary, the language model maps possible phonetic pictures into words and sentence structure[8]. Furthermore, it is described by Fadhil and Villafiorita in their paper[15] that, to make a virtual assistant system user-friendly, the method used Natural Language Processing algorithms for the natural conversion of the speech. This study was to develop a mobile phone application predominantly for children. All the data used in that paper were apprehended from the school children aged 8 to 14 years who used the system[15].

The speech's emotional tones are different for males and females. In the paper[6], two different models have been used for males and females. Moreover, emotions can be categorized in other forms. Happiness, sadness, anger, and fear- these four types of emotions are categorized by Dellaert, Polzin, and Waibel in their paper[1]. On the contrary, Anger, Happiness, Fear, Sadness, Neutral, and Surprise- these six types of categories are described in the paper[6]. The paper by Pathak and Kulkarni refers to another six types of emotions- Sad, Anger, Happy, Fear, Disgust, and Boredom[7]. The paper by Pathak and Kulkarni refers to another six types of emotions- Sad, Anger, Happy, Fear, Disgust, and Boredom[7]. Therefore, the dataset must be divided into different emotions into other classes. Before that, the proper emotional audio dataset is needed. According to Devillers, the dataset has to be appropriately labeled, and it has to be suited to the model[8]. The importance of labeling the dataset described by Bhat and Yadav in their paper is that it has to be labeled into the appropriate emotional category. Another model has been proposed, where the semi-supervised model is used, and dataset labeling is not required to determine the emotion from the speech[6]. The labeled dataset has to be split into train and test categories to use the dataset both for training the algorithm and testing the efficacy of the algorithm[6][42][7].

Recognizing emotions from utterances is challenging as there are different emotions, starting with happiness, anger, sadness, fear, and many more. In the paper[1], the voting algorithm (20.5%) has the closest error rate compared to humans (18%). The author mentioned in the paper[7], different forms of emotion are characterized using the Prediction Coefficient(LPC). Neural Network(NN) and the network trained using EBPTA to perform this task. However, the outcome was not as high as ex-

pected. The efficiency is close to 46%[7]. In their paper, "Emotion Dynamics Modeling via BERT", Yang and Shen stated that Emotion Dynamics Modeling is one of the noteworthy functions in Sentiment Recognition in a conversational context[42]. The goal is to anticipate conversational sentiments while creating an empathetic dialogue system. They made a series of BERT-based models to change the inter-interlocutor and intra-interlocutor interdependence of informal emotion dynamics in a more distinct manner. They used Flat-structured BERT (F-BERT) to affiliate the expressions in a context instantaneously alongside Hierarchically-structured BERT (H-BERT) so that they can apprehend the interlocutors while affiliating the expressions[42]. Moving forward, they recommended identifying the emotional influence amongst interlocutors using a Spatial-Temporal-structured BERT (ST-BERT). Furthermore, the datasets they exploited in the entire experiment are two prominent Emotion Recognition in Conversation (ERC) datasets. Last but not least, they demonstrated that the models they proposed based on the ERC datasets are capable of procuring around 5%-10% enhancement over baselines that are currently considered to be state-of-the-art. Emotion Recognition in Speech (ERS) is one of the most enjoyable working fields due to its practical life utilization[13]. To predict the proper emotion specified in any context, several representations can be used. Schmitt, Ringeval, and Schuller used Bag-of-Audio-Words (BoAW), which is one of them. In audio recognition tasks, Bag-of-Audio-Words (BoAW) has been implied successfully[13]. A method was proposed using BoAW, which is created only by using Mel-frequency Cepstral Coefficients (MFCCs) in "At the Border of Acoustics and Linguistics: Bag-of-Audio-Words for the Recognition of Emotions in Speech." BoAW representation of MFCCs performs exceptionally better than functionals and excels in late-proclaimed deep learning methods, including convolutional and recurrent networks. They evaluated their methodology over the Remote Collaborative and Affective Interactions (RECOLA) corpus since it has fully extemporaneous and genuine practical actions. Their database of the dataset consists of 46 multimodal recordings[13]. Also, in Chan, Hao, and Lee's work, emotion recognition has three phases: Feature Extraction, Feature Selection, Classification. These are the phases where the sentiment in a context is selected, identified, and compared[1].

AI chatbots have gradually been exploited to promote compassion and subsidize people who are confronting trying times[40]. People return to computers as community delegates, according to the CASA framework. Still, it does not assert that individuals would deal with computers exactly as real humans will. When the interaction is between a machine and people, people automatically assume stereotypes like it is rationalistic, objective, restrained, and cold[28], which basically compose the outcome of interactions. People tend to believe that artificial chatbots are not still capable enough to feel the sentiments of an individual and their reactions towards

the individual are also customized. The sentimental support based on algorithms may seem artificial, so an individual is less inclined to take into account artificial chatbots as genuine origins of sentimental support. On the contrary, compassionate associates can come up with sentimental associations that acquaint genuine inclination, comprehension, and caring[19]. Although the traditional chatbots have limitations about understanding the context and sentiments, modification with the help of machine learning methodologies is also a viable possibility now.

Additionally, Augmentation can be another strategy employed in our work in order to achieve an improved accuracy ratio. The study[41] uses the Multi-Window Data Augmentation (MWA-SER) technique for speech emotion identification. MWA-SER is a unimodal strategy that relies on two fundamental ideas: constructing a speech augmentation technique and establishing a deep learning model for discerning a sound signal's inherent sentiment. By applying several window widths in the audio feature extraction method, our referred multi-window augmentation technique provides a greater set of data from the spoken signal. When paired with a deep learning model, Multi-Window Data Augmentation increases vocal emotion identification ability.[25]

Chapter 3

Data Procurement

To acquire expected accuracy, multiple datasets must be run on the algorithms of the intended research. Therefore, a well-organized and neat dataset is regarded as one of the most important aspects of any research. The algorithms' performance in research is primarily determined by how convenient the dataset is. This task necessitates using data consisting of various audios expressing various emotions. RAVDESS, TESS, CREMA-D, SAVEE, Indic TTS, Libri TTS, LJ Speech, are among the types of datasets utilized to conduct the emotion recognition algorithm in our study. Each one contains a large amount of audio information.

3.1 Dataset Description

To acquire well-rounded data for our models, we used a range of datasets. Some of the data was created by professionals, while others were gathered from the public.We needed numerous datasets to modify a few crucial variables like audio quality and sampling rate to improve our accuracy. We sought to balance each data item with a comparable alternative to eliminate bias in our model. We used the CREMA-D dataset, which is fairly varied and includes audio in a variety of dialects and quality levels, to ensure that we receive an accurate portrayal of the reality. In addition to CREMA-D, we employed other datasets that will contribute to our overall accuracy with their respective quality and efficiency.

3.1.1 RAVDESS

Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) is the acronym[22]. However, we only used voice audio files in our investigation. SR Livingstone and FA Russo (2018) Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of North American English face and vocal expressions. The human audio dataset contains 1440 samples (60 Trial x 24 Actors) of actors, twelve male, twelve female uttering two lexically comparable

statements together in North American idiom. The speech dataset includes expressions that are fearful, neutral, surprised, disgusted, calm, happy, sad, and angry, with two stages of emotional intensity: strong and normal.

3.1.2 TESS

TESS[34] is a shorthand for the Toronto Emotional Speech Set created by the Psychology Department at the University of Toronto in 2010. This collection includes the voices of two actresses, ages 26 and 64. This dataset contains a total of 200 voiced specific words. An audio recording represents each of the seven emotions: sadness, happiness, pleasant, disgust, anger, fear, surprise, and neutral. The two women utter 200 target words with accompanying emotions in 2800 wav audio files. This dataset must be paired with male data to remove bias in the developed model.

3.1.3 CREMA -D

The abbreviation for "Crowd-sourced Emotional Multimodal Actors Dataset" is CREMA-D[44]. There are 7442 audio recordings in this collection, with 48 male and 43 female participants ranging in age from 20 to 74. African Americans, Asians, Caucasians, Hispanics, and Unspecified are among the actors of various races and ethnicities. This is the most diversified dataset among those included in this research. The performers are given a choice of 12 sentences to speak utilizing one of six emotion categories that are following Happy, Disgust, Neutral, Fear, Anger, and Sad. Low, Medium, High, and Unspecified were the four levels of intensity.

3.1.4 SAVEE

Audio-Visual Expression in Surrey SAVEE is actually abbreviated as Emotion[43]. Only mail actor's audios are represented in this sample. The University of Surrey features four adult native English speakers in this dataset. The models that are constructed will be influenced by the use of this straight male dataset. As a result, it's best to combine this dataset with another that has more female speakers (in our instance, TESS). This dataset contains seven behavioral categories: happy, disgust, neutral, fear, anger, surprised and sad. . These thirteen males are between the ages of 27 and 31 years old. With each emotion, the text content consists of 15 TIMIT phrases. There are two sentiment-specific, three common and last but not the least, ten general phrases for each emotion that are phonetically balanced. To create 30 neutral utterances, the three basic and 12 sentiment-specific phrases were registered as neural. There are overall 120 utterances per actor.

3.1.5 Indic TTS

Indic TTS[45] is an application that creates text-to-speech algorithms for thirteen Indian languages. Including Bengali, using a consortium of high-quality corpora[11]. The collection comprises both audio and textual descriptions of talks. There are male and female actors for every basic language who read the script from fiction, science, newspapers and other sources. Audio talks are also recorded in a calm, echo-free location. The voice signals were captured at a sample rate of 48KHz. There are 15.23 hours and 15.75 hours of audio recordings of male and female performers delivering English utterances with Bengali and Hindi accents, respectively. This speech corpus[11] is being utilized to develop new speech-recognition algorithms in both the Indian and English languages, with a focus on the Indian accent.

3.1.6 Libri TTS

This dataset is completely free to use. Linda Johnson, the speaker, spewed thirteen thousand one hundred small audio excerpts from seven nonfiction books. The audio snippets comprise over 24 hours in duration. Each audio sample is given an English transcription. Additionally, the period of the audio samples varies from one second to ten seconds. The developers of the dataset physically coupled textual transcribed to sound, as well as a quality assurance check was performed to ensure that the transcript phrases actually matches the audio dialogues.

3.1.7 LJ Speech

This dataset is completely free to use. Linda Johnson, the speaker, spewed thirteen thousand one hundred small audio excerpts from seven nonfiction books. The audio snippets comprise over 24 hours in duration. Each audio sample is given an English transcription. Additionally, the period of the audio samples varies from one second to ten seconds. The dataset's authors physically coupled text transcription to audio, and a quality assurance check was done to guarantee that the transcript phrases matched the audio discussions precisely.

3.2 Data Splitting

Data splitting, which divides the dataset into three parts: validation data, testing data, and training data, is another facet of the dataset preparation.The entire dataset is separated into 80 percent training data, 10 percent testing data, and 10 percent validation data using an 8:1:1 ratio. In addition, 10% of the training data is being utilized to validate the results.

Data Type	**Percentage**	**Parameters**
Training Data	80%	·train_indices ·x_train ·y_train
Testing Data	10%	·test_indices ·x_test ·y_test
Validation Data	10%	·valid_indices ·x_valid ·y_valid

Table 3.1: Data Splitting with Parameters

Chapter 4

Feature and Augmentation

4.1 Features

Before this study can use the audio recordings from the dataset in the algorithms, they must first be processed. Furthermore, precise data is required for classification and prediction models, as well as a proper data representation, known as features.These audio recordings include a variety of components that have been shown to boost the outcomes of various models.

4.1.1 Short-Time Fourier Transform (STFT)

The Short-Time Fourier Transform (STFT) is used to divide the acoustic waves into overlapping and short sections at first. Then power spectrograms are generated using the Fourier transform of each segment. Making power spectrograms has the objective of identifying resonant frequencies in waveforms. In addition, an STFT has the advantage of detecting audio signal changes in time series data. Fast Fourier transformation creates a Mel-scale representation of frequencies known as Mel-Spectrogram.

4.1.2 Mel-Spectrogram

A spectrogram is a visual representation of a signal's frequency spectrum as it changes over time. The Mel-spectrogram is just a scale of Mel depiction of waves obtained through a quick Fourier transformation. It's a type of spectrogram that can be used to extract useful data from audio. A short-time Fourier transform is used to translate audio wave impulses from the time domain to the frequency domain utilizing brief and overlapping chunks across the audio waves to construct the spectrogram. Frequencies surpassing a particular threshold are represented in a logarithmically portrayed mel-spectrogram (the corner frequency). Because humans have a restricted range of comprehension of frequencies and amplitudes, the

frequency axis of the spectrogram is then transformed to a log-scale. The color dimension is also translated to decibels. Decibels are used to represent the color dimension. Finally, the frequency axis on the non-linear Mel Scale is plotted to create a Mel-spectrogram. The Mel-spectrogram is utilized to provide sound information to our models that is similar to what a human would hear. Mel-spectrograms represent analog approximations of power spectrograms on the Mel-frequency scale that have been streamlined. To construct the Mel-spectrogram, raw audio waveforms are routed through filter banks. Once one has a basic understanding of spectrograms, learning other types of them becomes a breeze. All that's needed is a new framework upon which the spectrograms are built. Another feature that can be employed in different classification approaches is this one.

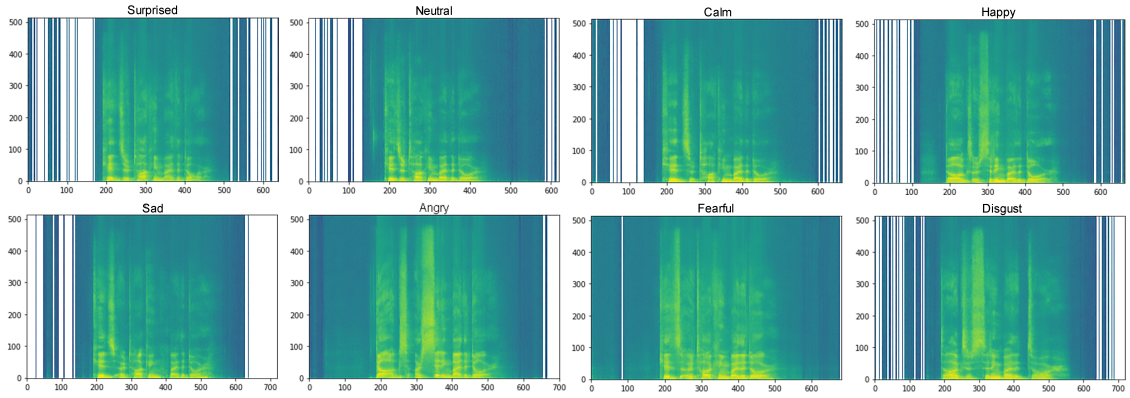


Figure 4.1: Mel-spectrogram of all emotional categories

4.1.3 Mel-frequency Energy Coefficient(MFEC)

MFEC stands for Mel-Frequency Energy Coefficients. This approach is well-known for its adaptability and is widely utilized in Automated Speech Recognition. The goal of this project was to build a CNN architecture that uses MFEC rather than MFCC to achieve high precision in emotion recognition. The proposed architecture was tested using the open-source RAVDESS dataset. MFECs are the cepstral coefficients calculated from the power spectrum of a signal with logarithmically divided frequencies as per the Mel-scale. This is a nonlinear frequency scale that measures noise absorption in the auditory system. This procedure follows some steps sequentially. Audio power spectrums are generated using STFT. After that, frequency bins are created by calculating the total of each window's energy, and the power spectrograms are treated using triangular, overlapping window functions. After that, the Mel Scale is employed to map the frequencies of the audio stream's power spectrograms. The number and arrangement of window functionalities and the frequency bins' width are all determined by this mapping. The Mel Scale is used because indi-

viduals perceive audio pitches determined by frequency ratios; also, it's a pitch scale that doesn't follow a linear pattern portraying audio pitches in "mels" of frequency. The contribution of this research is to introduce a new model that enables Mel-Frequency Energy Coefficients (MFEC) as the input to CNN rather than a raw feature to achieve a higher-level recognition of emotional characteristics.

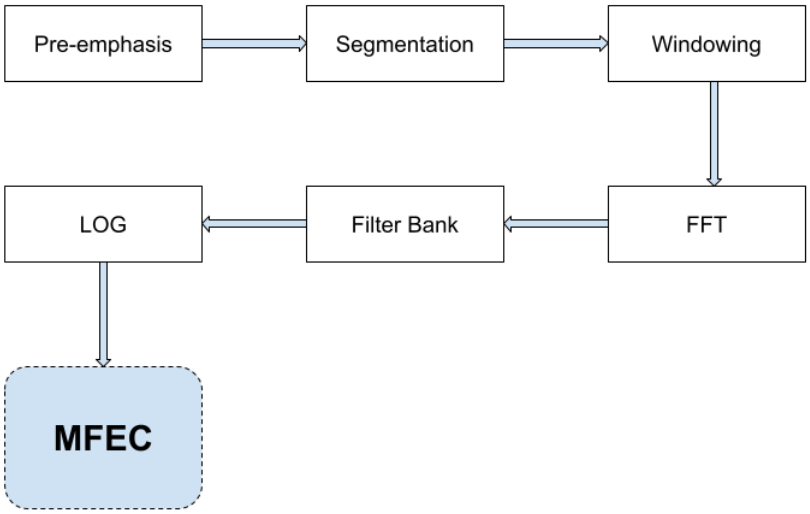


Figure 4.2: MFEC Feature Analysis

The voice signal is pre-emphasis to increase the frequency components that contain less energy than the lower frequency, and then to minimize signal discontinuity, which can lead to erroneous extracting features[38]. The framing approach is utilized since it enables a particular file to be partitioned into numerous extremely short and fixed-length periods with quasi-stationary signals[3]. To have all voice signals in the frame more precisely consistent, the generated frames are multiplied by a weighted window.

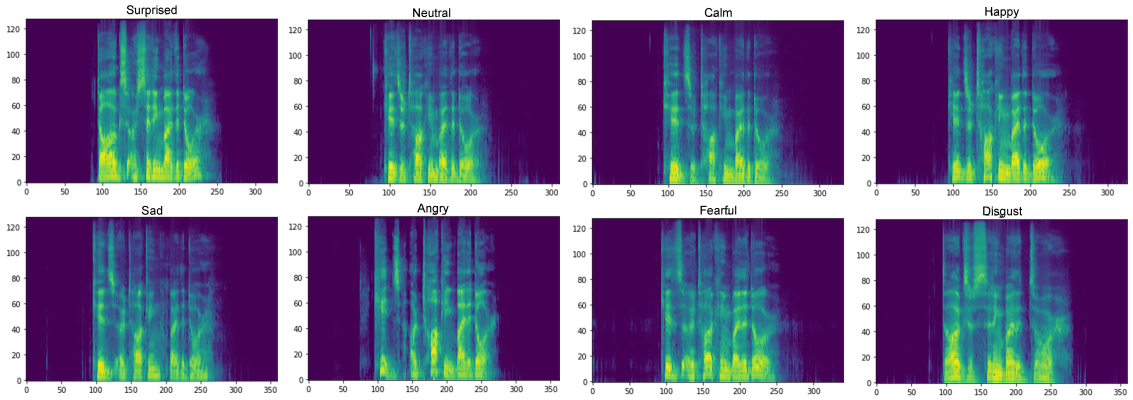


Figure 4.3: MFEC Features for Different Emotional Catagories

4.1.4 Chromagram

The chromagram is a powerful set of components for understanding music whose pitches may be labeled, and it is calculated from the wavelength or power spectrogram (chroma stft). Because the respiratory sounds range significantly in pitch, Chroma is a good feature in our instance.

The twelve unique pitch classes in audio are referred to as chromagrams or chroma features. Chroma-based qualities may help audio with semantically classified pitches and intonation that resembles the equal-tempered octave. Pitch class profiles is also a term for it[25]. Significant features of chroma characteristics is their proficiency to catch melodic and harmonic components of music as well as being reluctant to participate in instrumentation and pitch. Chroma character traits[31] are used to depict the harmonic content of a small time frame of sound.

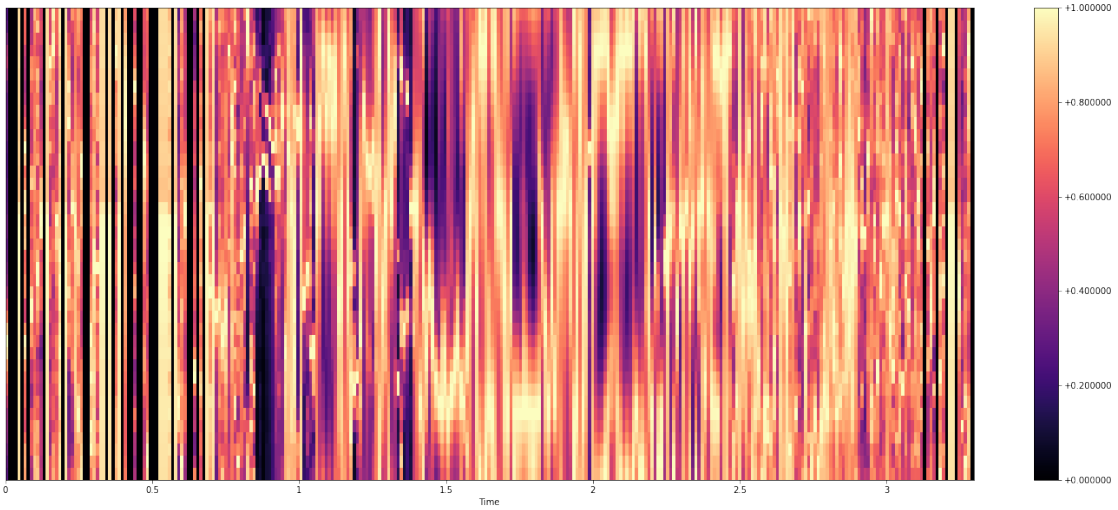


Figure 4.4: Chromagram Feature

4.1.5 Spectral Contrast

Spectral contrast is the audible disparity among the ebbs and flows in the spectrum. The non-rhythmic and rhythmic component allocation in the spectrum might be approximated by the Spectral Contrast criterion. Spectral allocation in each sub-band was averaged using MFCC and other prior characteristics, causing a loss of relevant spectral analysis. The overall spectral allocation is unlikely to establish audio spectral characteristics, even if two spectra with distinct spectral allocations have equal average spectral features. On the other hand, spectral contrast preserves greater data and may enable better categorization distinction. As a consequence, we've decided to incorporate this functionality into our own design.

A spectrogram is divided into sub-bands upon every frame. For every sub-band, energy disparity will be assessed by examining the mean energy of the largest per-

centile (maximum energy) to the mean energy of the lowest percentile (base energy) (valley energy). Broad-band noise is related with low disparity values, while clear, narrow-band signals are correlated with elevated disparity values[4].

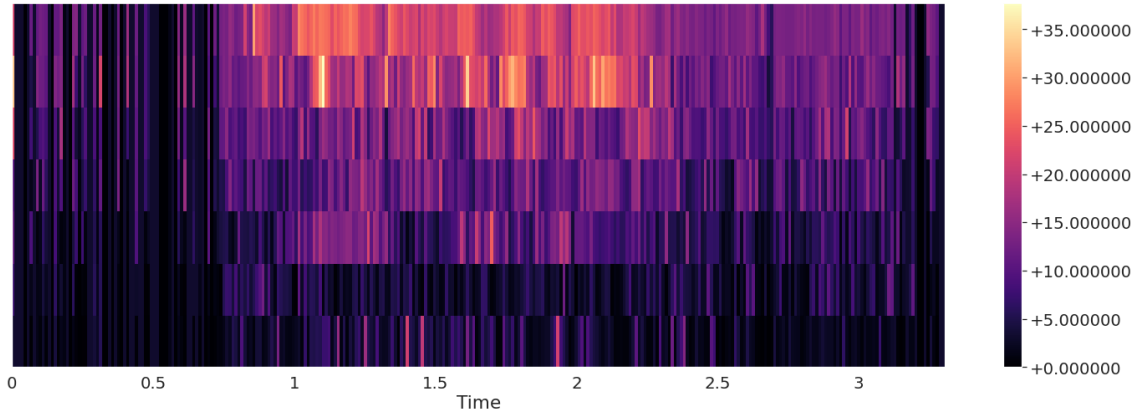


Figure 4.5: Spectral Contrast Feature

4.1.6 Tonal Centroid (Tonnetz)

The Tonnetz is a pitch space characterized by a set of relationships between melodic notes, according to a simple interpretation. An audible file can contain up to twelve pitch categories in the chroma characteristic, but it is processed even though it only had six pitch categories by combining sections of these. As an outcome, the chroma characteristic graph generated by this technique simply categorizes pitches into six-dimensional bases, which we call Tonnetz, and each represents the minor third, ideal fifth, and main third as two-dimensional coordinates. When utilizing machine learning algorithms to analyze pulmonary sound, the Tonnetz characteristic may be useful[5].

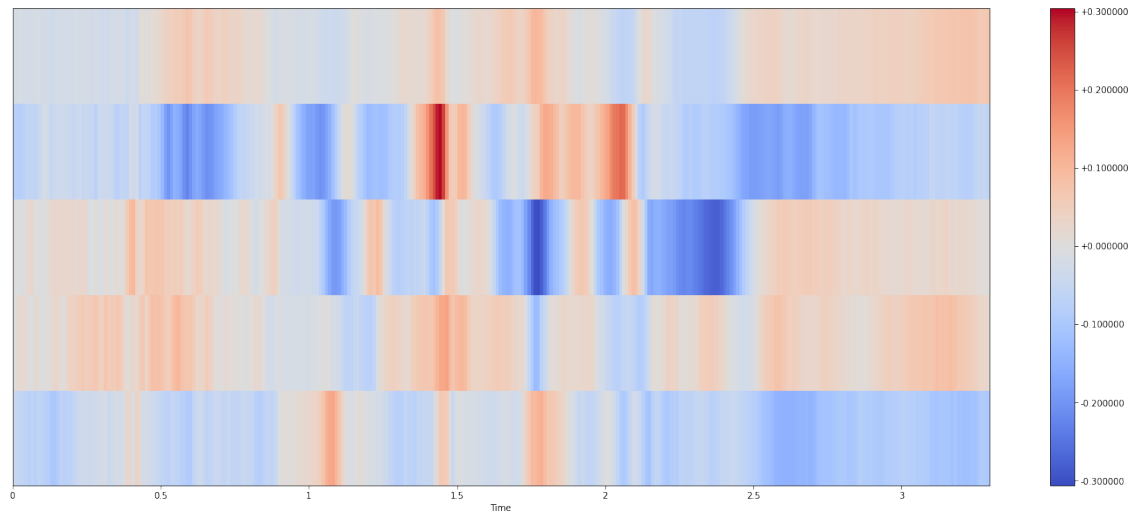


Figure 4.6: Tonal Centroid Feature

4.1.7 Mel-Frequency Cepstral Coefficients (MFCC)

Mel-frequency Cepstral Coefficients identify audio signal pitch fluctuations. It is a mathematical representation that converts the power spectrograms of a sound signal obtained using STFT into a small number of coefficients that indicate the audio signal's power in the frequency domain. For this transformation, certain mathematical techniques are performed sequentially. The initial application of STFT was to create audio power spectrums. The wavelength divisions are then created by summing the sum of the energy from each window and employing triangular, intertwining window functions to the power spectrograms. The Mel Scale is then used to depict the frequencies of the audio signal's power spectrograms. The amount and placement of window functions, alongside the breadth of the frequency bins, can be determined using this mapping. The reason behind utilizing Mel Scale is because people perceive audio pitches using frequency proportions, and the non-linear pitch range represents audio pitches in "mels" of frequency. Mel filterbanks are a combination of window functions and frequency bins. The logarithm of the summation of input audio's power spectrograms, usually referred as cepstrum, is then determined for each filterbank. Finally, because there are correlations between filterbank energies, the logarithm of the summation of the power spectrograms for each filterbank is used to decorrelate them using the Discrete Cosine Transform (DCT). The benefit of using a DCT is that it develops coefficients in such a way that just the topmost few coefficients correctly characterize the audio stream. The peaks of the DCT of the logarithm of the summation of the filterbank powers with respect to time are therefore mel-frequency cepstral coefficients. The MFCC lays the path for us to deconvolution audio waveforms in order to locate resonant frequencies.

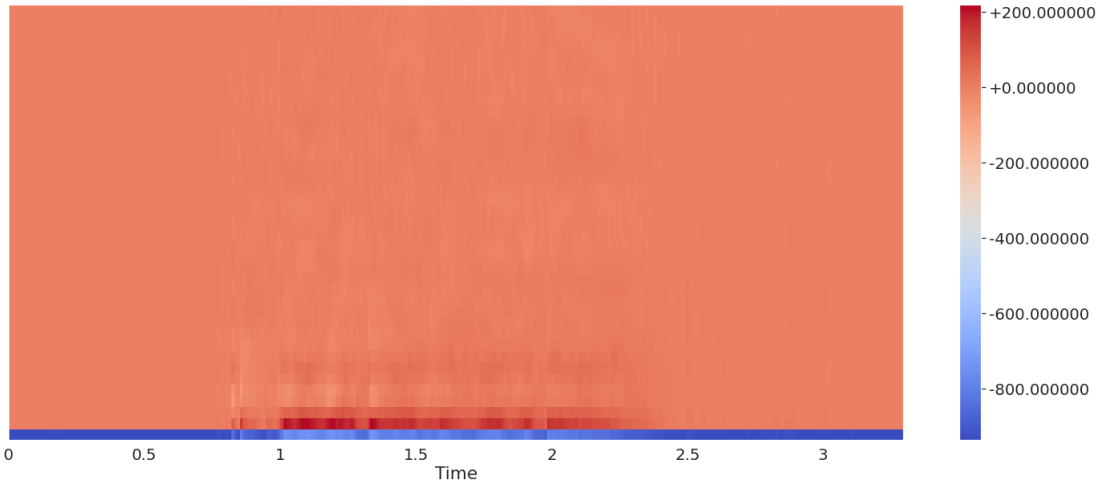


Figure 4.7: MFCC Feature

4.2 Augmentation

Experientially, one of the most typical issues in genuine data science problems is a shortage of data. Data augmentation facilitates the development of artificial data utilizing existing data sources, increasing the model's generalizability. Noise injection, shifting time, modifying pitch, and speed can all be used to create syntactic data for audio.

4.2.1 Add noise

Noise can help machine learning models generalize functions more effectively. Introducing noise to the sample set can help audio sentiment identification algorithms perform better. This is how a standard audio speech clip looks before and after injecting Additive White Gaussian Noise.

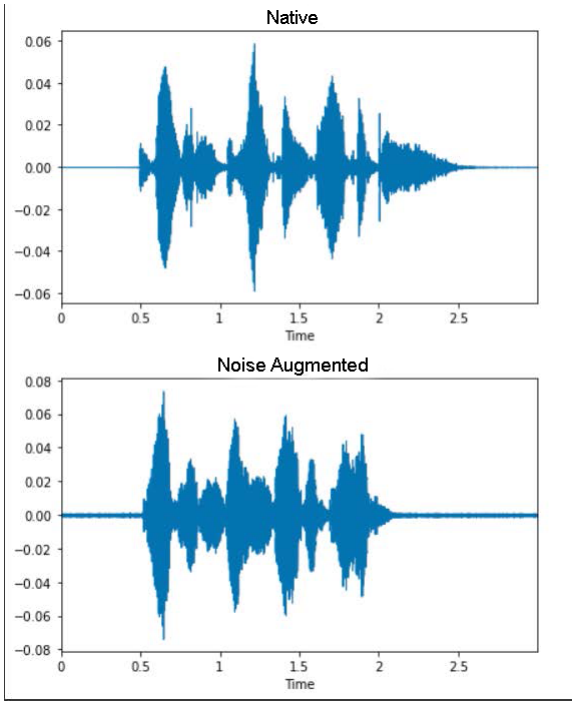


Figure 4.8: Noiseless vs With Noise Audio Waveforms

4.2.2 Increase Speed

Another procedure that is being used in this study to process an audio sample is the implementation of the speed increment on the audio samples. However, the outcomes of the audio sample processing due to speed increment approach were not satisfactory. Because increasing the speed of an audio sample causes the audio to become distorted. The processing of an audio sample is harmed by audio distortion because the audio becomes rapid, loud, or even broken. As a result, we abandoned

the audio processing approach through speed incrementation that we used at first since the results were unsatisfactory.

4.2.3 Time shifting

Shifting time is a straightforward concept. The audio is simply shifted to left side or right side at arbitrary intervals. If audio is fast forwarded by n seconds, its first n seconds will be indicated as 0 (silenced). When the audio is shifted to the right side for n seconds, the latest n seconds are indicated as 0. (will be silenced).

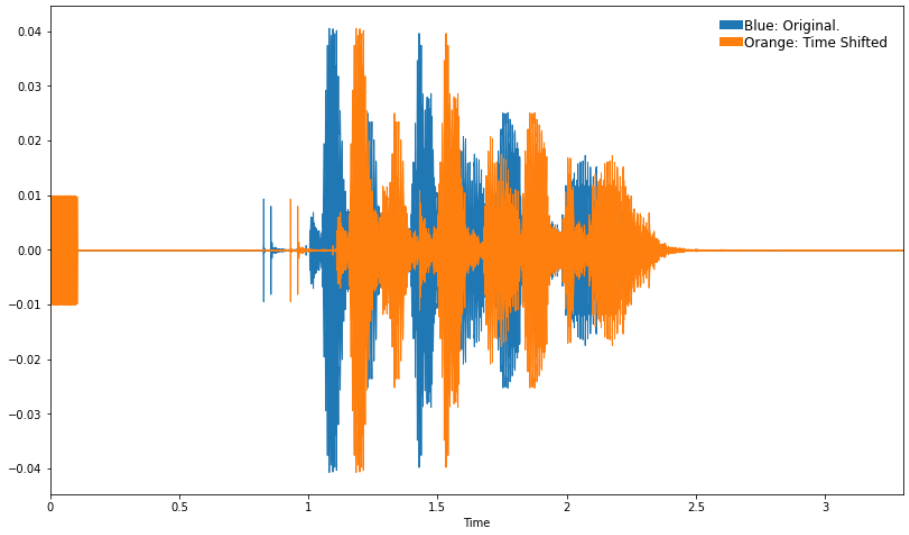


Figure 4.9: Original vs Time Shifted Audio Waveforms

4.2.4 Signal Loss

Recording audio in the real world might result in signal loss due to a variety of devices and delay difficulties. To increase data clarity, the majority of audio datasets are prepared in a noise-free setting. In natural environments, machine learning models must also perform better. As a consequence, the training dataset should reflect the real world. As a result, audio signals are enhanced by signal loss.

4.2.5 Change volume

In general, carefully managed audio speech data maintain a constant degree of loudness. Audio dataset developers are given enough of a break to minimize fatigue from long hours of recording and to maintain a constant tone throughout the operations. People do not engage with technologies in the same manner that actors are taught to do. They may speak rather loudly at times. It is possible that the machine learning system will perform badly if it is not robust to the loudness of the voice

or the environment. As a result, we seldom change the amount of any input during augmentation with the explicit objective of generalizing the machine learning model.

4.2.6 Spec Augmentation

SpecAugment[26], an unique augmentation approach for automated speech recognition, has been released by Google. Over the audio streams, conventional augmentation processes are performed. SpecAugmentation by Google augments the auditory spectrogram, which is a graphic illustration of the signal. Unlike previous augmentation approaches, this method incurs no additional computational costs or data. SpecAugment modifies spectrograms along the vertical and horizontal axes, which correspond to Mel-frequency bands and times, respectively. This method enables cnn architectures to generalize beyond information loss in speech sources.

Chapter 5

Methodology

5.1 Proposed Model

Some features utilize audio signals as input to categorize them into distinct emotions. The system evaluates the input and extracts features called MFEC (Mel-frequency energy coefficients), Mel-spectrogram, and Delta of coefficients to create a decent result. TThe system then leverages Mel-spectrograms as features to define the emotional state of the input (anger, happiness, sadness, disgust etc.) Using Parallel CNN and Transformer-Encoders[37]. The RNN-based Deepspeech model then uses the spectrograms of the audio inputs to generate English text transcripts[9]. The system then aggregates these two qualities and provides a contextual answer for users by employing a multi-domain conversational agent[2]. Following the response from the multi-domain conversational agent, the system uses Flowtron to generate a Mel-spectrogram, which synthesizes speech as naturally as feasible[36]. Finally, Waveglow is a model that generates spoken audio signals using Mel-spectrograms[29].

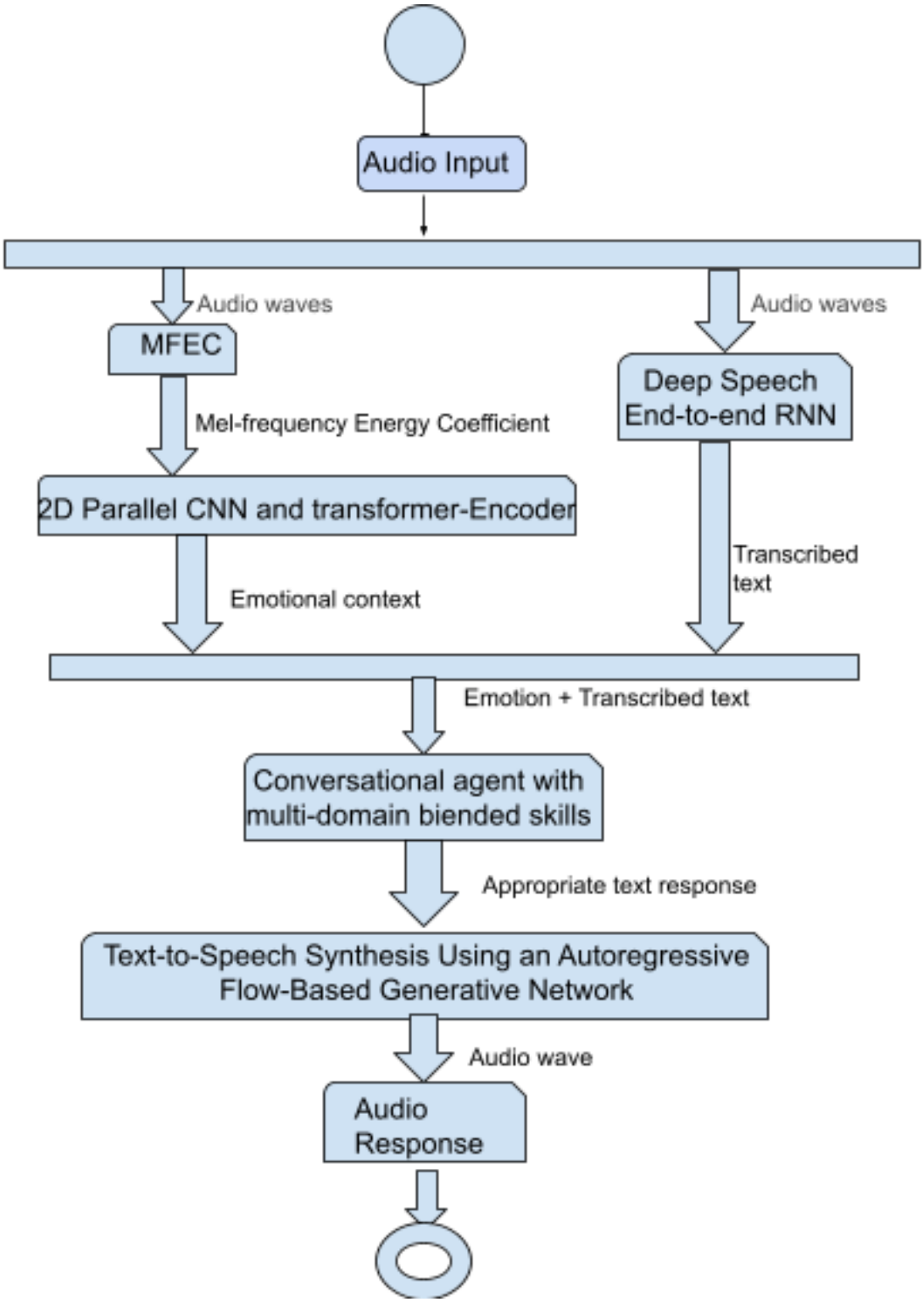


Figure 5.1: Block Diagram of the Proposed Model

5.2 Parallel 2D CNN and Transformer-Encoders

Artificial Neural Network's (ANN) another type is CNN. CNN model lowers processing costs while producing good performance. In this paradigm, larger features are filtered by smaller size filters known as Kernels. CNN is a neural network containing layers such as convolution, maxpool, dropout, and flattening. The entire input is filtered part by part in the convolution layer. When one frame of input has been filtered, the kernel moves to the next frame and filters that piece until the input is finished processing. The maxpooling procedure is performed after the conversion layer. This section collects the maximum value from each kernel size. Using that method, a fully - connected model may be made shorter and just the important parts employed. The kernel slides depending on the value of "stride". The output size of CNN model can be calculated using the value of input size, kernel size, stride size, and padding. Considering, P = Padding, S = Stride, K = Filter size, W = Input size. Then, output size= (((W - K + 2P)/S) + 1).

In this research, the parallel CNN is used in two parallel layers. With 2D parallel CNN, Transformer-Encoders are used to categorize audios into various emotional states. The Mel-frequency Energy Coefficient of audios is used as input of the model. This approach employs eight emotional states. Those are- fearful, neutral, surprised, disgusted, calm, happy, sad, and angry. The MFEC can be used like an image. The suggested model was constructed using the knowledge of earlier CNN research' findings[37]. This model implements the smaller kernel (3x3), and stride size=1. Previously, a larger kernel with 11x11, five strides was utilized in the AlexNet. VGGNet originated the concept of utilizing a smaller kernel. In general, a picture has three layers: R, G, and B. However, the MFEC characteristics may be utilized as black-and-white images using a single channel. There are 49C2 parameters in the 7x7 kernel. 3x3 kernels with 4 layers, on the other hand, have less parameters. As a result, the 3x3 kernel lowers processing costs. MFEC creates features with a size of (128,282). CNN blocks 1 and 2 are utilized in this model. There are four levels in each block. The input size and output sizes of every layers are computed below:

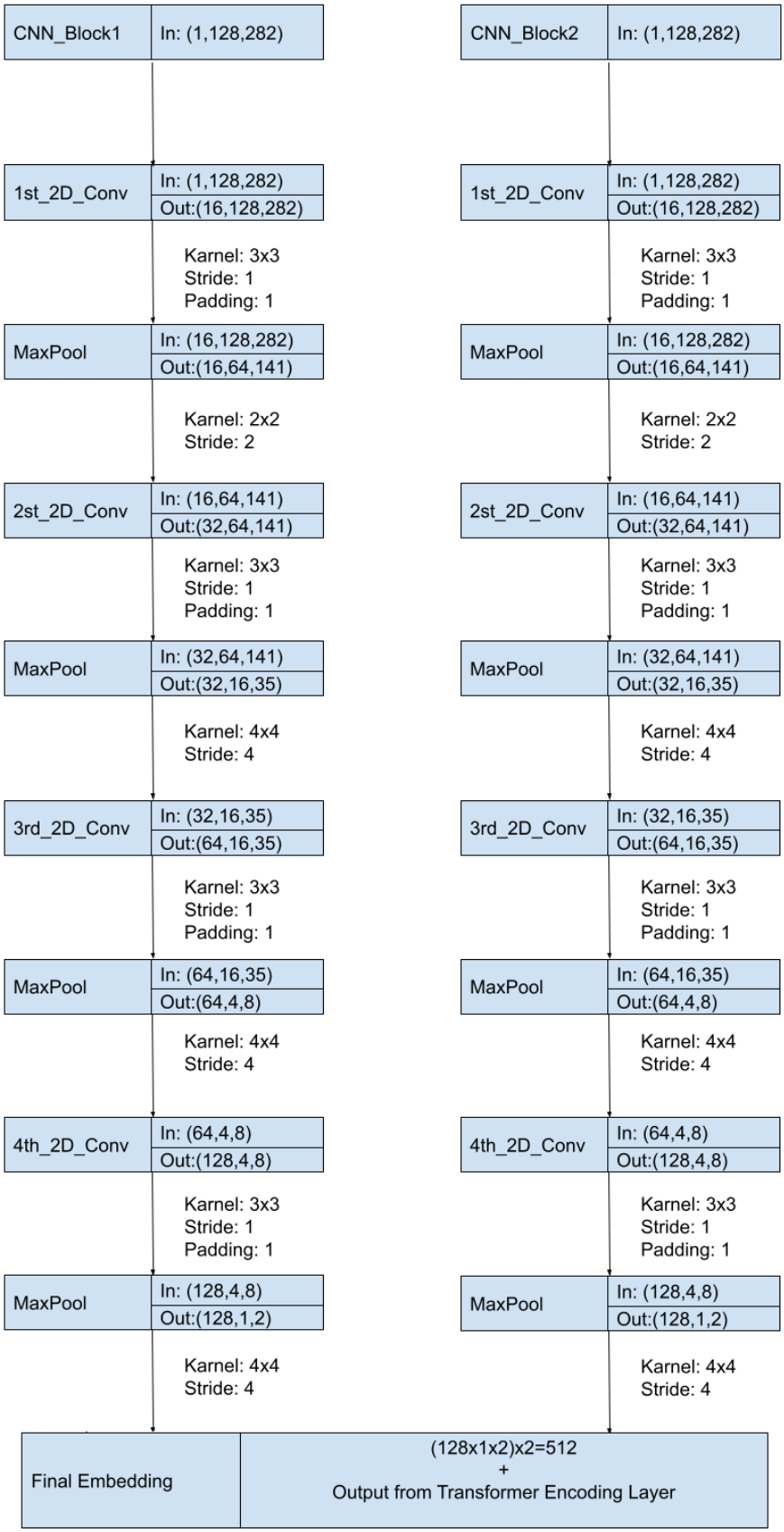


Figure 5.2: CNN Model Layers and Dimensions

The transformer encoder's purpose is to learn the temporal aspects of distinct emotions in the hopes of discovering the frequency distribution of every emotion depending on the global structure of every emotion's Mel-spectrogram. The transformer's encoder examines the specified frequency sequence and is suitable for developing sequential data.

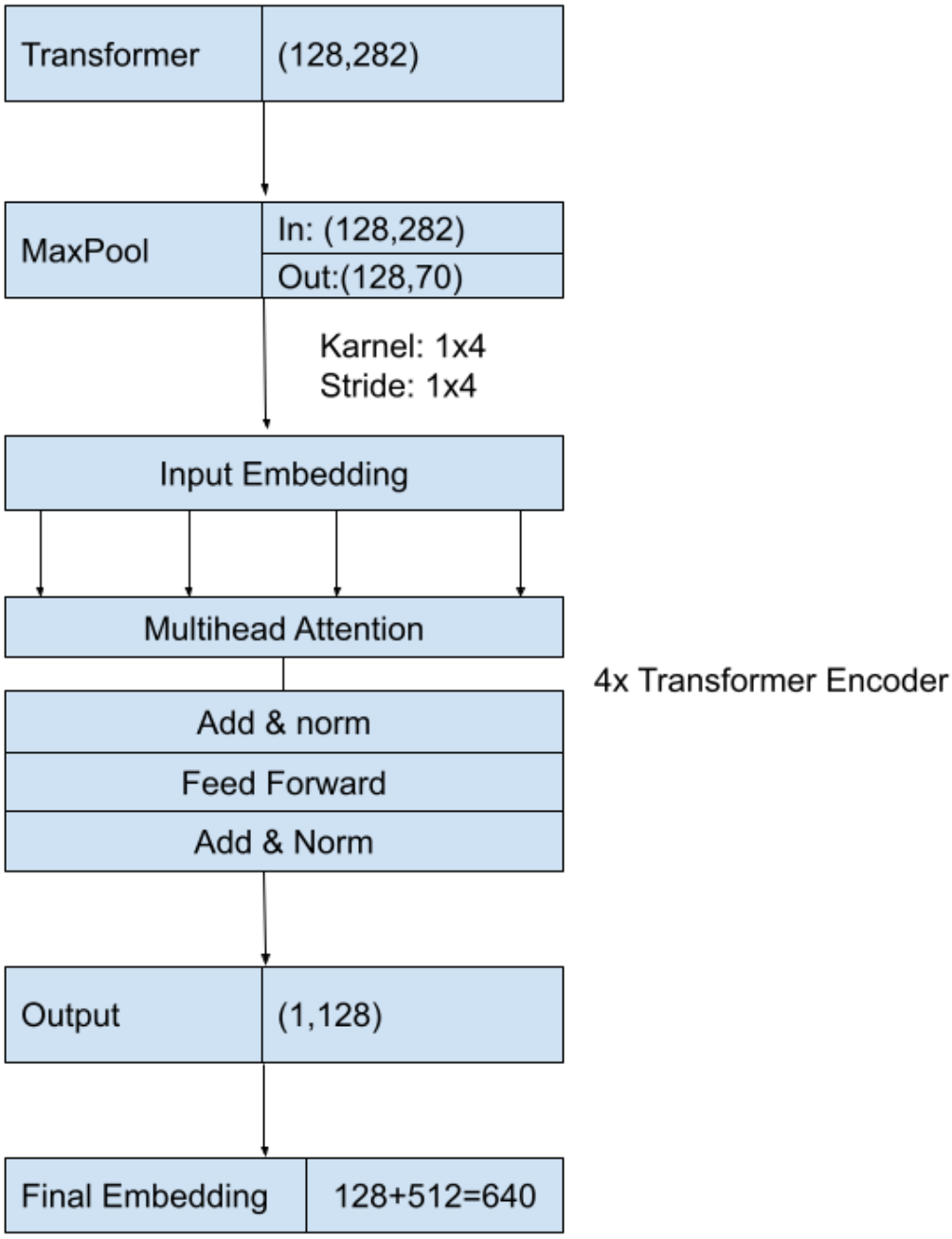


Figure 5.3: Transformer Encoder Layers

5.3 Deep Speech: End-To-End RNN

Deep Speech is a system with an end-to-end RNN speech trainer that has been extensively tuned. This system takes in speech spectrograms and generates an English text transcription of the speech.

$$X = \left\{ \left(x^{(1)}, y^{(1)}\right), \left(x^{(2)}, y^{(2)}\right), \ldots \right\} \tag{5.1}$$

Here, a single utterance is represented by x, and a label by y. For text transcription, the RNN is set up to convert an input order x into a character probability sequence [9]. The RNN model is made up of five hidden layers.

$$\begin{gathered} h_t^{(l)} = g\left(W^{(l)} h^{(l-1)} + b^{(l)}\right) \\ W^{(l)} = \text{ weight }, b^{(l)} = \text{ bias }, l = \text{ currentlayer} \end{gathered} \tag{5.2}$$

In this case, h(l) denotes a hidden layer, and l represents each layer.

$$\begin{aligned} h_t^{(f)} &= g\left(W^{(4)} h_t^{(3)} + W_r^{(f)} h_{t-1}^{(f)} + b^{(4)}\right) \\ h_t^{(b)} &= g\left(W^{(4)} h_t^{(3)} + W_r^{(b)} h_{t-1}^{(b)} + b^{(4)}\right) \end{aligned} \tag{5.3}$$

Forward recurrence computes h(f) progressively starting t=1 to t=T(i) for each utterance i, whereas backward recurrence computes h(f) backwards from t=T(i) to t=1. These two layers' outputs are then merged and transferred to the fifth layer, resulting in the construction of this function:

$$g\left(W^{(5)} h_t^{(4)} + b^{(5)}\right); h_t^{(4)} = h_t^{(f)} + h_t^{(b)} \tag{5.4}$$

Finally, the output layer employs the usual softmax method to forecast character probabilities.

5.4 Open-domain conversational agent

An agent must possess some key abilities to conduct a human-like dialogue. For example, an agent needs to be empathic, knowledgeable, intelligent, and capable of engaging with the user. A Facebook AI research team discovers that the agent should be pre-trained with vast corpora [32]. According to the study "Blended Skill Talk," conversational context and training data may be used to train open-domain

conversational bots (BST)" [35]. Decoding two models with the same accuracy but different algorithms might provide drastically different outcomes. The model also demonstrates the importance of conversational bots' replies to humans. Humans then see concise replies as uninteresting and express less engagement. On the other hand, humans will not listen to long responses. This model indicates that limiting the beam length to minimum yields better and more exciting results. The study utilized three distinct types of model architectures: Retriever architecture, Generative architecture, and Retrieve and Refine architecture. The Transformer model is used in these three architectures.

Firstly, the Retriever architecture produces a series of solutions and scores them. The response with the best score is then offered as the possible output. In this model, poly-encoder architecture is used to encode the global features of the context [33]. The benefit of utilizing poly-encoders over other retrieval models is that they deliver cutting-edge performance on a variety of conversation tasks.

$$y^i_{ctxt} = \sum_j w_j^{c_i} h_j \quad \text{where} \quad (w_1^{c_i}, .., w_N^{c_i}) = \text{softmax}\left(c_i \cdot h_1, .., c_i \cdot h_N\right) \tag{5.5}$$

$$y_{ctxt} = \sum_i w_i y^i_{ctxt} \quad \text{where} \quad (w_1, \ldots, w_m) = \text{softmax}\left(y_{cand_i} \cdot y^1_{ctxt}, .., y_{cand_i} \cdot y^m_{ctxt}\right) \tag{5.6}$$

Secondly, the Generator architecture creates replies by utilizing the Seq2Seq Transformer architecture. This approach generates output rather than retrieving from a set. Finally, tokenization is performed using the Byte-Level BPE tokenization technique [18]. It is similar to the Transformer model, except it can handle more extensive parameters.

Finally, the Retrieve and Refine architecture combines the Retrieve and Generator architectures. The results of the Retrieve layer are used as an entry to the Generative design. The outcome of the Retrieve architecture is separated using separator tokens before being provided as an input.

5.5 Autoregressive Flow-based Generative Network for TTS Synthesis

Text incorporating non-textual information, such as accent, pitch, and tone, is required for mel-spectrogram synthesis. Non-textual content must also correspond to the audio data's style. Suppose some audio data is fed to the model reflecting a

certain sentiment like fury, amazement„ and so on. In that case, the synthetic Mel-spectrogram should resemble the design, a phenomenon known as “Style Transfer.” NVIDIA researchers created the ”Flowtron” model, which can execute a similar function as previously stated. This is accomplished by Flowtron by increasing the likelihood of training data. Invertible data mapping to a latent space is learned by Flowtron., which may be used to control different voice synthesis aspects, including pitch, accent, speech tempo, tone, and so on [36]. Based on past Mel-spectrogram frames, it creates a new Mel-spectrogram frame. The total number of frames is p(X) = p(xtjx1:t1). Two distributions are sampled by the Flowtron's neural network, which acts as a generative model. z N, is a zero-mean spherical Gaussian (z; 0; I). The other type of mixture is a spherical Gaussian with static or updatable parameters.

$$z \sim \sum_{k} \hat{\phi}_k N\left(z; \hat{\mu}_k, \sum_{k}\right) \tag{5.7}$$

Using integrable and defined ”affine transformations,” the samples are transformed from p(z) to p(x). Flowtron already implements an integrable neural network. Coupling layers [21] [16], especially an affine coupling layer, are used to develop integrable neural networks [14]. A scale, s, and bias are generated for each input xt1. For example, the following input xt is transformed by this scale and bias affine:

$$\begin{gathered} (\log s_t, b_t) = NN\left(x_{1:t-1},\ \text{text}\ ,\ \text{speaker}\ \right) \\ x'_t = s_t \odot x_t + b_t \end{gathered} \tag{5.8}$$

In this case, NN() represents any type of autoregressive correlation transformation. Concatenation of a zero vector with the other NN() parameters can do this. Although the NN() does not have to be integrable, the affine coupling layer assures that the entire network is. Every t-th variable z0t in the autoregressive structure is dependent on its preceding timesteps from the star z1:t-1: [17].

$$z'_t = f_k\left(z_{1:t-1}\right) \tag{5.9}$$

Flowtron maximizes the log-likelihood of the data by employing the previously described parameterized affine transformation alongside autoregressive architecture. These are made feasible by changing variables:

$$\begin{gathered}\log p_\theta(x) = \log p_\theta(z) + \sum_{i=1}^{k} \log \left|\det\left(j\left(f_i^{-1}(x)\right)\right)\right| \\ z = f_k^{-1} \circ f_{k-1}^{-1} \circ \ldots f_0^{-1}(x)\end{gathered} \tag{5.10}$$

Mel-spectrograms are encoded to vectors and directed via numerous affine coupling layers given text and a predefined pseudo speaker embedding. The ”flow” refers to each affine coupling layer. Finally, the vectors that have been prepared are sent into the neural network for execution. To infer, the trained neural network is fed arbitrarily picked z values out of a Gaussian Mixture with constant or Flowtron anticipated parameters. A single WaveGlow [23] model previously generated on a single speaker is used to decode the inferred Mel-spectrograms into waveforms.

Chapter 6

Implementation and Results

6.1 Implementing Proposed Model

This study uses numerous approaches to determine which strategy best matches our situation to observe the accuracy rate. The parameters are identified that can influence the accuracy rate positively and negatively. Moreover, they are tweaked to verify the consistency of each model to establish the accuracy that is used in this work. Parallel 2D CNN with Transformer Encoder outperformed all other approaches in terms of computing efficiency. In this work, the RAVDESS dataset has performed best with the parallel 2D CNN model with Transformer Encoder in terms of accuracy.

6.1.1 Sequential 1D CNN

From the ground up, the endeavor is to build a sentiment classification system. A 1D convolutional neural network is utilized to predict the mood class using the mean value of the feature. This method is used to get a sense of the MFCC feature. The MFCC feature is an extraction technique that stands for Mel-Frequency cepstral coefficient. The MFCC coefficients contain information about rate changes in various spectrum bands. The approach tried here is inadequate and does not yield the best result. However, the accuracy rate is 51.01%. Therefore, the decision is not to waste time on it based on the accuracy rate of the approach. As a result, different methods are tried in the study rather than this one.

6.1.2 Sequential 2D CNN

There is another attempt to create an image using the MFCC value. Later, CNN is used to classify the image, which also assisted this study in classifying the emotion of that particular image. Initially promising results are obtained in this approach and then attempted to tinker with the parameters to ensure accuracy. It has to be

noted that the MFCC coefficients significantly impact the accuracy rate. Using this method, an accuracy of 68.02% is obtained. Number of 100 epochs are tried here. When the number of epochs is increased, the accuracy does not vary significantly.

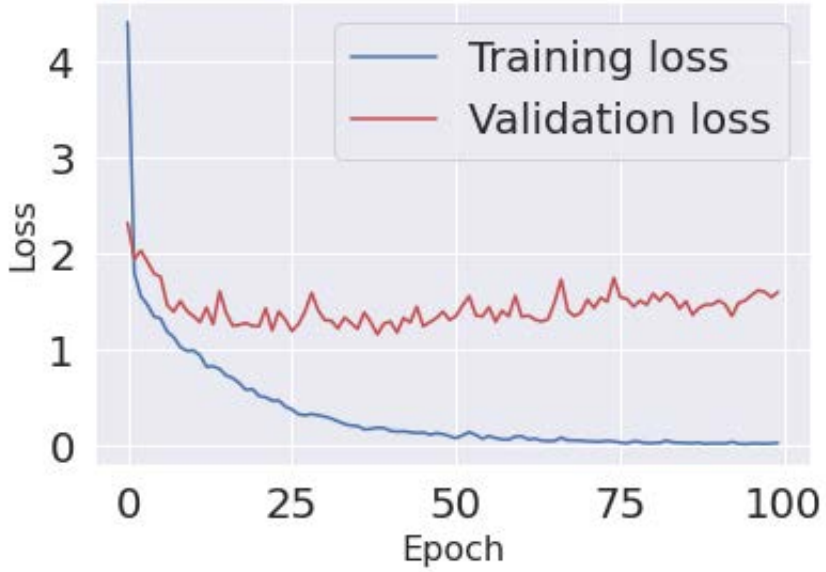


Figure 6.1: Loss Grpah of Sequential 2D CNN

6.1.3 XResnet Models (Transfer Learning)

The experiment is attempted to determine the accuracy with xResnet50 and xResnet18, prioritizing speed and accuracy[20]. In this approach, augmentation has been used, which includes the removal of silence, noises, signal loss, and volume changes, along with various spectrogram-augmentation techniques. It helps to temper the spectrogram so that it is resistant to spectrogram distortion in the time direction[27]. This method does not allow much because the accuracy rate is so low, on the other hand, using heavy augmentation has reduced the accuracy rate, even though both light and heavy augmentation are being tried.

6.1.4 VGG19 (Transfer-Learning)

VGG19 is a 19-layer VGG containing 16 convolution layers, three entirely linked layers, five max-pooling levels, and one SoftMax layer[10]. Comparatively being an old and simple model it has fundamental classification architecture that is also effective. That prompts the study to try out this model. Same augmentation as in the XResnet model approach is used in this approach. However, the accuracy is not up to par.

6.1.5 MFCC and Parallel 2D CNN with Transformer Encoder

2D parallel CNN has a filter complexity that increases in stages. It gradually reduces feature maps, resulting in high-quality features with increased computation performance. Furthermore, by focusing on temporal data, the transformer encoder learns the frequency distribution of various emotional categories. As a feature for this model, a Spectrogram produced from Mel-frequency cepstral coefficients (MFCC) of the audio signal was employed. As an augmentation method, we've applied time shifting to the data.

Feature	Epoch	Augmentation	Accuracy
MFCC	100	No	72.03%
MFCC	120	Yes	75.45%

Table 6.1: Accuracy of Parallel 2D CNN with Transformer Encoder using MFCC

We acquired 72.03% accuracy for 100 epochs without any augmentation. But when we added augmented data along with the real data and increased the epoch to 120, the accuracy jumped to 75.45%.

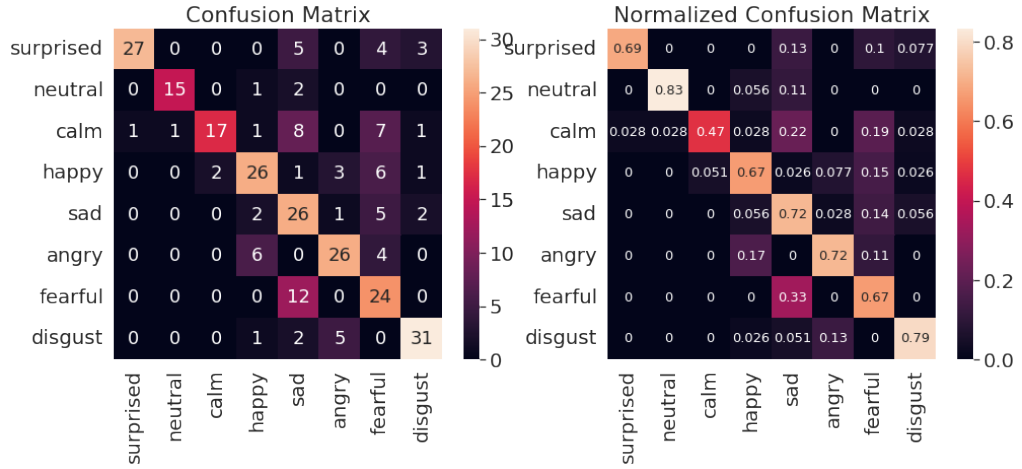


Figure 6.2: Confusion matrix of Parallel 2D CNN with Transformer Encoder using MFCC

6.1.6 Tonnetz and Parallel 2D CNN with Transformer Encoder

On Parallel 2D CNN with Transformer Encoder, we also evaluated the Tonal Centroid function. We encountered some runtime issues while testing this feature on 2D CNN using Transformer Encoder. In comparison to other processes, it is a sluggish

one and the output was below expectation. With augmentation, we obtained an accuracy of 66.67 percent on the original data. We also tried to improve the outcome by using tonal centroid on MFCC and Chromagram.

Feature	**Epoch**	**Augmentation**	**Accuracy**
Tonal Centroid	100	Yes	66.67%
Tonal Centroid + MFCC	120	Yes	68.89%
Tonal Centroid + Chromagram	100	Yes	51.28%
Tonal Centroid + Chromagram Delta	100	Yes	42.27%

Table 6.2: Results for Tonnetz features

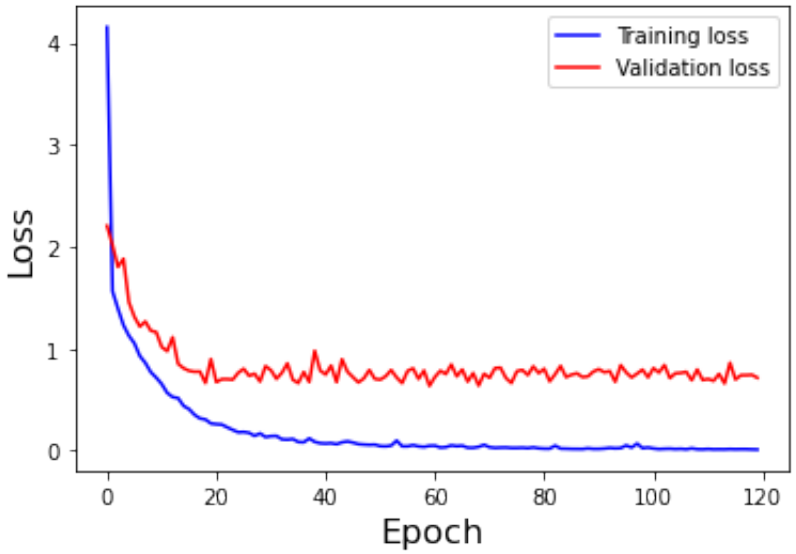


Figure 6.3: Loss graph using MFCC with Tonal Centroid

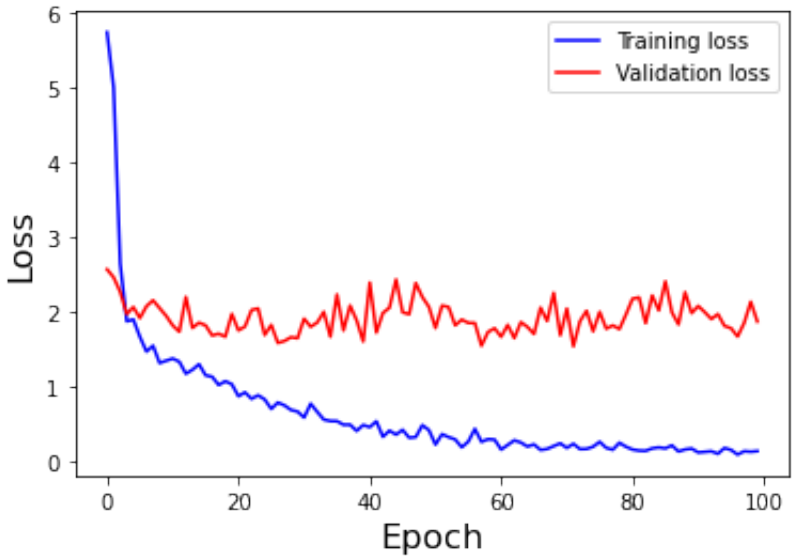


Figure 6.4: Loss graph using Chromagram with Tonal Centroid

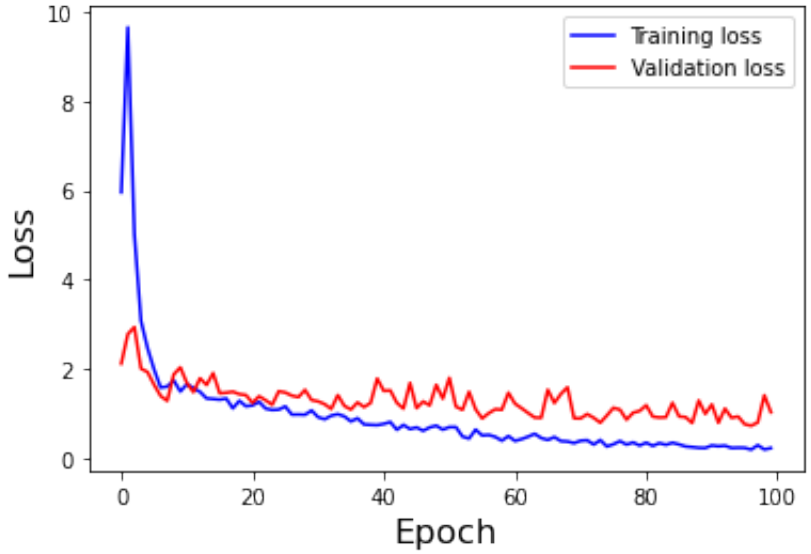


Figure 6.5: Loss graph for Chromagram with Tonal Centroid Delta

These experiments yielded no notable improvements in accuracy. We also discovered that extensive augmentation reduces model accuracy. We repeated the experiment numerous times but were unable to make the model stable enough for our purposes. As a result, we abandoned the technique.

6.1.7 Chromagram and Parallel 2D CNN with Transformer Encoder

In our case, Chromagram is an excellent feature to use with 2D CNN and Transformer Encoder. Audio with semantically categorized pitches and intonation that mimics the equal-tempered octave may benefit from chroma-based features. However, this feature also fell short of our expectations. This trait yielded the lowest accuracy of all, at 30.95%.

6.1.8 MFEC and Parallel 2D CNN with Transformer Encoder

The 2D parallel CNN model is well-known for its exceptional computational performance and high-quality features. Mel-frequency Energy Coefficients of audio signals are employed as a characteristic for this model. This model employs both augmented and non-augmented features. The model was trained using 2186 MFEC features of RAVDESS dataset. As a result, the model's accuracy rate for the dataset is pretty impressive. The maximum level of accuracy is about 87.69% with 200 epochs. RAVDESS and CREMA-D were the first two datasets we used.

Table 6.3: Accuracy using MFEC features with different datasets

Feature	Epoch	Augmentation	Dataset	Accuracy
MFEC	200	Yes	RAVDESS	87.69%
MFEC	120	Yes	CREMA-D	71.56%

In RAVDESS, male and female actors were evenly distributed (12 male, 12 female). The parallel 2D CNN model worked well in one region since the data set's idioms were all in North American. However, the CREMA-D dataset did not match the RAVDESS accuracy level since there were various regions in CREMA-D, and as a result, the diversity of ethnicity and race restricted the accuracy level that could be achieved using our proposed model. There were just two actors in the TESS dataset, both of whom were female. We decided not to proceed with this dataset since it would be gender biased if we worked on it. We also worked on SAVEE, which is a different dataset. The actors in this dataset were also gender biased because there were only male performers and their ages were restricted between 27 to 31 years old.As both datasets had seven features, the optimum solution was to merge TESS and SAVEE. However, the idioms for both datasets are different, and it did not work as expected. Only RAVDESS provided a stable and optimum solution among these datasets, and its accuracy was greater than the others. As we conclude, using RAVDESS on 2D parallel CNN with Transformer Encoder yields the best results.

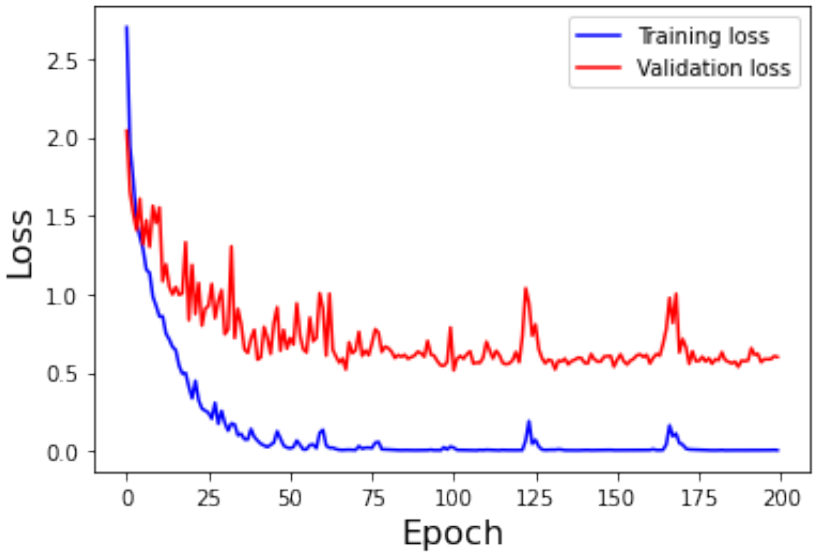


Figure 6.6: Loss graph for the model using RAVDESS.

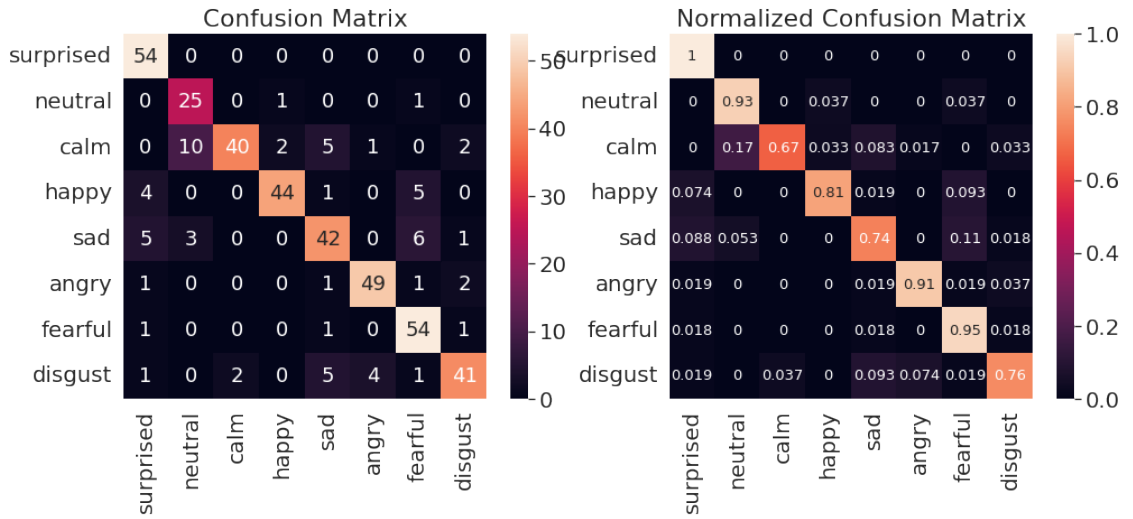


Figure 6.7: Confusion matrix for the model using RAVDESS

We concluded that, RAVDESS provides the best feasible result of 2D parallel CNN with Transformer encoder using MFEC as feature.

6.1.9 Deep Speech Tuning

This model is used to generate an English transcript from audio clips of speeches[9]. Nonetheless, the model does not produce a convincing outcome for the input data that is given. Given the difficulties, it is possible that the model's accuracy is affected by the accent. This model is attempted to be used with Indic TTS to overcome the accent discrepancy [30]. Although the native accent differs significantly from the Asian accent, using Indic TTS made a significant difference in inaccuracy. Significant accuracy can be attained if a vast dataset is repeatedly trained.

6.2 Result Analysis

After implementing described models, mixed results in terms of accuracy and efficiency are being achieved. Parameters are adjusted to ensure the uniformity of each model and determine the accuracy employed in our work. Sequential 1D CNN was not providing us with relevant results in our instance. In terms of Computational efficiency, using MFEC with the Parallel 2D CNN and Transformer Encoder surpassed all other techniques.

Models	**Epoch**	**Class**	**Accuracy**	**Data Augmentation**	**Feature**
Sequential 1D CNN	100	7	50.03%	No	
	100	14	42.92%	No	
Sequential 2D CNN	50	8	66.67%	Yes	
	100	8	68.02%	Yes	
	150	8	68.13%	Yes	
XResnet50	25		68.04%	AddNoise, RemoveSilence, ChangeVolume	MFCC+Delta
	25		70.13%	AddNoise	MFCC+Delta
VGG19	15		67.97%	AddNoise	MFCC+Delta
	15		70.43%	AddNoise, ChangeVolume, Trim, SignalLoss	MFCC+Delta
Parallel 2D CNN And Transformer-Encoder	350	8	67.34%	Yes	Mel Spectrogram
Parallel 2D CNN and Transformer-Encoder	100	8	72.03%	No	MFCC
Parallel 2D CNN and Transformer-Encoder	120	8	75.45%	Yes	MFCC
Parallel 2D CNN and Transformer-Encoder	100	8	51.28%	Yes	Chromagram + Tonal Centroid
Parallel 2D CNN and Transformer-Encoder	100	8	30.95%	Yes	Chromagram
Parallel 2D CNN and Transformer-Encoder	100	8	68.89%	Yes	Tonal Centroid + MFCC
Parallel 2D CNN and Transformer-Encoder	100	8	66.67%	Yes	Tonal Centroid
Parallel 2D CNN and Transformer-Encoder	100	8	42.27%	Yes	Tonal Centroid + Chromagram Delta
Parallel 2D CNN and Transformer-Encoder	**120**	**8**	**87.67%**	**Yes**	**MFEC**
Parallel 2D CNN and Transformer-Encoder	**200**	**8**	**87.69%**	**Yes**	**MFEC**

Table 6.4: Results Comparison

Chapter 7

Conclusion

This study elaborated on the existing IVA technology and its limitations. The three major limitations are addressed: context, sentiments, and personality. Our research aims to identify the above-mentioned categorized limitations of artificial virtual assistants, exterminate them, and advance them with interactive personality traits with diverse characteristics that will hold human-like conversations with end-users. Upon consideration, there are still sectors of furtherance in the queue. IVAs can recognize emotions but not in a precise manner. Speech Emotion Recognition isn't a brand-new field of study. Although the academic interest in software emotion recognition has remained relatively constant, the approach has changed significantly. The methodologies discussed in this research, such as BoAW representation, BERT methodology, etc., will make the emotion recognition outcomes more precise and accurate. Another methodology we mentioned for our research purpose is Natural Language Processing(NLP), making IVA user-friendly. Thus, IVA understands the user's tendency through communication, and it responds accordingly. Machine Learning(ML) is essential to train and test datasets, and thus the machine learns whether the audio refers to happiness, sadness, fear, or even anger. After the machine has trained all the datasets, it needs to test on another dataset, known as the test set. Without machine learning (ML) methodologies, IVAs cannot recognize the correct form of emotion. We can leap towards, after adjusting all the methods and overcoming the limitations and barriers of current existing IVAs. This will transform a traditional IVA closer to human emotion with the same energy, tendency, and integrity.

Bibliography

[1] F. Dellaert, T. Polzin, and A. Waibel, "Recognizing emotion in speech," in *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP '96*, vol. 3, 1996, 1970–1973 vol.3. DOI: 10.1109/ICSLP.1996.608022.

[2] F. Dellaert, T. Polzin, and A. Waibel, "Recognizing emotion in speech," *International Conference on Spoken Language Processing, ICSLP, Proceedings*, vol. 3, Dec. 1996.

[3] B. Juang and T. Chen, "The past, present, and future of speech processing," *IEEE Signal Processing Magazine*, vol. 15, no. 3, pp. 24–48, 1998. DOI: 10.1109/79.671130.

[4] D.-N. Jiang, L. Lu, J.-H. Tao, and L.-H. Cai, "Music type classification by spectral contrast feature," Feb. 2002, 113–116 vol.1, ISBN: 0-7803-7304-9. DOI: 10.1109/ICME.2002.1035731.

[5] C. Harte, M. Sandler, and M. Gasser, "Detecting harmonic change in musical audio," in *Proceedings of the 1st ACM Workshop on Audio and Music Computing Multimedia*, ser. AMCMM '06, Santa Barbara, California, USA: Association for Computing Machinery, 2006, pp. 21–26, ISBN: 1595935010. DOI: 10.1145/1178723.1178727. [Online]. Available: https://doi.org/10.1145/1178723.1178727.

[6] J. Liu, C. Chen, J. Bu, M. You, and J. Tao, "Speech emotion recognition using an enhanced co-training algorithm," in *2007 IEEE International Conference on Multimedia and Expo*, 2007, pp. 999–1002. DOI: 10.1109/ICME.2007.4284821.

[7] S. Pathak and A. Kulkarni, "Recognizing emotions from speech," in *2011 3rd International Conference on Electronics Computer Technology*, vol. 4, 2011, pp. 107–109. DOI: 10.1109/ICECTECH.2011.5941867.

[8] J. Deng, R. Xia, Z. Zhang, Y. Liu, and B. Schuller, "Introducing shared-hidden-layer autoencoders for transfer learning and their application in acoustic emotion recognition," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 4818–4822. DOI: 10.1109/ICASSP.2014.6854517.

[9] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates, and A. Y. Ng, *Deep speech: Scaling up end-to-end speech recognition*, 2014. arXiv: 1412.5567 [cs.CL].

[10] K. Simonyan and A. Zisserman, *Very deep convolutional networks for large-scale image recognition*, 2015. arXiv: 1409.1556 [cs.CV].

[11] A. Baby, A. Thomas, N. L, and T. Consortium, "Resources for indian languages," in. Sep. 2016, p. 8.

[12] M. Nishimura, K. Hashimoto, K. Oura, Y. Nankaku, and K. Tokuda, "Singing Voice Synthesis Based on Deep Neural Networks," in *Proc. Interspeech 2016*, 2016, pp. 2478–2482. DOI: 10.21437/Interspeech.2016-1027.

[13] M. Schmitt, F. Ringeval, and B. Schuller, "At the Border of Acoustics and Linguistics: Bag-of-Audio-Words for the Recognition of Emotions in Speech," in *Proc. Interspeech 2016*, 2016, pp. 495–499. DOI: 10.21437/Interspeech.2016-1124.

[14] L. Dinh, J. Sohl-Dickstein, and S. Bengio, *Density estimation using real nvp*, 2017. arXiv: 1605.08803 [cs.LG].

[15] A. Fadhil and A. Villafiorita, "An adaptive learning with gamification amp; conversational uis: The rise of cibopolibot," in *Adjunct Publication of the 25th Conference on User Modeling, Adaptation and Personalization*, ser. UMAP '17, Bratislava, Slovakia: Association for Computing Machinery, 2017, pp. 408–412, ISBN: 9781450350679. DOI: 10.1145/3099023.3099112. [Online]. Available: https://doi.org/10.1145/3099023.3099112.

[16] D. P. Kingma and J. Ba, *Adam: A method for stochastic optimization*, 2017. arXiv: 1412.6980 [cs.LG].

[17] D. P. Kingma, T. Salimans, R. Jozefowicz, X. Chen, I. Sutskever, and M. Welling, *Improving variational inference with inverse autoregressive flow*, 2017. arXiv: 1606.04934 [cs.LG].

[18] A. Miller, W. Feng, D. Batra, A. Bordes, A. Fisch, J. Lu, D. Parikh, and J. Weston, "Parlai: A dialog research software platform," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Copenhagen, Denmark: Association for Computational Linguistics, Sep. 2017, pp. 79–84. DOI: 10.18653/v1/D17-2014.

[19] J.-P. Stein and P. Ohler, "Venturing into the uncanny valley of mind—the influence of mind attribution on the acceptance of human-like characters in a virtual reality setting," *Cognition*, vol. 160, pp. 43–50, 2017, ISSN: 0010-0277. DOI: https://doi.org/10.1016/j.cognition.2016.12.010. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0010027716303055.

[20] T. He, Z. Zhang, H. Zhang, Z. Zhang, J. Xie, and M. Li, *Bag of tricks for image classification with convolutional neural networks*, 2018. arXiv: 1812.01187 [cs.CV].

[21] D. P. Kingma and P. Dhariwal, *Glow: Generative flow with invertible 1x1 convolutions*, 2018. arXiv: 1807.03039 [stat.ML].

[22] S. R. Livingstone and F. A. Russo, "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*, vol. 13, no. 5, pp. 1–35, May 2018. DOI: 10.1371/journal.pone.0196391. [Online]. Available: https://doi.org/10.1371/journal.pone.0196391.

[23] R. Prenger, R. Valle, and B. Catanzaro, *Waveglow: A flow-based generative network for speech synthesis*, 2018. arXiv: 1811.00002 [cs.SD].

[24] A. Ram, R. Prasad, C. Khatri, A. Venkatesh, R. Gabriel, Q. Liu, J. Nunn, B. Hedayatnia, M. Cheng, A. Nagar, E. King, K. Bland, A. Wartick, Y. Pan, H. Song, S. Jayadevan, G. Hwang, and A. Pettigrue, "Conversational ai: The science behind the alexa prize," Jan. 2018.

[25] J. Jogy. (2019). "How i understood: What features to consider while training audio files?" [Online]. Available: https://towardsdatascience.com/how-i-understood-what-features-to-consider-while-training-audio-files-eedfb6e9002b (visited on 03/28/2022).

[26] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition," in *Proc. Interspeech 2019*, 2019, pp. 2613–2617. DOI: 10.21437/Interspeech.2019-2680.

[27] ——, "Specaugment: A simple data augmentation method for automatic speech recognition," *Interspeech 2019*, Sep. 2019. DOI: 10.21437/interspeech.2019-2680. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2019-2680.

[28] S. S. Sundar and J. Kim, "Machine heuristic: When we trust computers more than humans with our personal information," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, ser. CHI '19, Glasgow, Scotland Uk: Association for Computing Machinery, 2019, pp. 1–9, ISBN: 9781450359702. DOI: 10.1145/3290605.3300768. [Online]. Available: https://doi.org/10.1145/3290605.3300768.

[29] N. Svenningsson and M. Faraon, "Artificial intelligence in conversational agents: A study of factors related to perceived humanness in chatbots," ser. AICCC 2019, Kobe, Japan: Association for Computing Machinery, 2019, ISBN: 9781450372633. DOI: 10.1145/3375959.3375973. [Online]. Available: https://doi.org/10.1145/3375959.3375973.

[30] R. Yamamoto, E. Song, and J.-M. Kim, "Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," Oct. 2019.

[31] K. Behera. (2020). "Feature extraction from audio," [Online]. Available: https://coeai19.wixsite.com/coe-ai-bbsr/post/feature-extraction-from-audio (visited on 04/05/2022).

[32] C. Fierro, "Recipes for building an open-domain chatbot," Jun. 2020.

[33] S. Humeau, K. Shuster, M.-A. Lachaux, and J. Weston, *Poly-encoders: Transformer architectures and pre-training strategies for fast and accurate multi-sentence scoring*, 2020. arXiv: 1905.01969 [cs.CL].

[34] M. K. Pichora-Fuller and K. Dupuis, *Toronto emotional speech set (TESS)*, version DRAFT VERSION, 2020. DOI: 10.5683/SP2/E8H2MF. [Online]. Available: https://doi.org/10.5683/SP2/E8H2MF.

[35] E. M. Smith, M. Williamson, K. Shuster, J. Weston, and Y.-L. Boureau, *Can you put it all together: Evaluating conversational agents' ability to blend skills*, 2020. arXiv: 2004.08449 [cs.CL].

[36] R. Valle, K. Shih, R. Prenger, and B. Catanzaro, *Flowtron: An autoregressive flow-based generative network for text-to-speech synthesis*, 2020. arXiv: 2005.05957 [cs.SD].

[37] I. Zenkov, *Transformer-cnn-emotion-recognition*, https://github.com/IliaZenkov/transformer-cnn-emotion-recognition, 2020.

[38] D. Abdiche and K. Harrar, "Text-independent speaker identification using mel-frequency energy coefficients and convolutional neural networks," Feb. 2021, pp. 204–209. DOI: 10.1109/IHSH51661.2021.9378726.

[39] A. Mamun, H. Md. Abdullah, and M. G. R. Alam, "Affective social anthropomorphic intelligent system," 2021.

[40] J. Meng and Y. N. Dai, "Emotional Support from AI Chatbots: Should a Supportive Partner Self-Disclose or Not?" *Journal of Computer-Mediated Communication*, vol. 26, no. 4, pp. 207–222, May 2021, ISSN: 1083-6101. DOI: 10.1093/jcmc/zmab005. eprint: https://academic.oup.com/jcmc/article-pdf/26/4/207/40342390/zmab005.pdf. [Online]. Available: https://doi.org/10.1093/jcmc/zmab005.

[41] N. J. Shoumy, L.-M. Ang, D. M. M. Rahaman, T. Zia, K. P. Seng, and S. Khatun, "Augmented audio data in improving speech emotion classification tasks," in *Advances and Trends in Artificial Intelligence. From Theory to Practice*, H. Fujita, A. Selamat, J. C.-W. Lin, and M. Ali, Eds., Cham: Springer International Publishing, 2021, pp. 360–365.

[42] H. Yang and J. Shen, "Emotion dynamics modeling via bert," 2021. [Online]. Available: https://arxiv.org/abs/2104.07252.

[43] S. Kanwal, S. Asghar, A. Hussain, and A. Rafique, "Identifying the evidence of speech emotional dialects using artificial intelligence: A cross-cultural study," *PLOS ONE*, vol. 17, e0265199, Mar. 2022. DOI: 10.1371/journal.pone.0265199.

[44] *"cheyneycomputerscience/crema-d: Crowd sourced emotional multimodal actors dataset (crema-d)"*. [Online]. Available: https://github.com/CheyneyComputerScience/CREMA-D.

[45] "Indic tts," [Online]. Available: https://www.iitm.ac.in/donlab/tts/index.php.