

Unveiling Personality Traits through Bangla Speech using Morlet Wavelet Transformation and Soft-Voting Classifier

by

Md. Sajeebul Islam Sk.
22366027

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

Department of Computer Science and Engineering
BRAC University
December 2023

© 2023. BRAC University
All rights reserved.

Declaration

It is hereby declared that

1. The thesis submitted is my 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:



Md. Sajeebul Islam Sk.

22366027

Approval

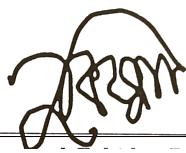
The thesis titled “Unveiling Personality Traits through Bangla Speech using Morlet Wavelet Transformation and Soft-Voting Classifier” submitted by

1. Md. Sajeebul Islam Sk. (22366027)

Of fall, 2023 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of Master of Science in Computer Science and Engineering on December 01, 2023.

Examining Committee:

External Examiner:
(Member)



Dr. Mohammad Zahidur Rahman

Professor
Department of Computer Science and Engineering
Jahangirnagar University

Internal Examiner:
(Member)



Dr. Muhammad Iqbal Hossain

Associate Professor
Department of Computer Science and Engineering
BRAC University

Internal Examiner:
(Member)



Moin Mostakim

Senior Lecturer
Department of Computer Science and Engineering
BRAC University

Supervisor:
(Member)



Dr. Md. Golam Rabiul Alam

Professor

Department of Computer Science and Engineering
BRAC University

Program Coordinator:
(Member)



Dr. Md Sadek Ferdous

Associate Professor

Department of Computer Science and Engineering
BRAC University

Head of Department:
(Chairperson)



Sadia Hamid Kazi, Ph.D.

Chairperson and Associate Professor

Department of Computer Science and Engineering
BRAC University

Ethics Statement

The conducted thesis adhered meticulously to the established research ethics, norms, and codes of practice outlined by BRAC University. It is affirmed that all sources referenced in the thesis have been duly cited. As the sole author of this work, I acknowledge and assume full responsibility for any potential violations of ethics codes that may arise. The research process was conducted with utmost regard for ethical considerations, underscoring the commitment to upholding the standards set by BRAC University throughout the study.

Abstract

Speech serves as a potent medium for expressing a wide array of psychologically significant attributes. While earlier research on deducing personality traits from user-generated speech predominantly centered on other languages, there is a noticeable absence of prior studies and datasets for automatically assessing user personalities from Bangla speech. In this paper, the speaker's objective is to bridge the research gap by generating speech samples, each imbued with distinct personality profiles. These personality impressions are subsequently linked to OCEAN (Openness, Conscientiousness, Extroversion, Agreeableness, and Neuroticism) NEO-FFI personality traits. To gauge accuracy, human evaluators, unaware of the speaker's identity, assess these five personality factors. The dataset is predominantly composed of around 90% content sourced from online Bangla newspapers, with the remaining 10% originating from renowned Bangla novels. We perform feature level fusion by combining MFCCs with LPC features to set MELP and MEWLP features. We introduce MoMF feature extraction method by transforming Morlet wavelet and fusing MFCCs feature. We develop two soft voting ensemble models, DistilRo (based on DistilBERT and RoBERTa) and BiG (based on Bi-LSTM and GRU), for personality classification in speech-to-text and speech modalities respectively. The DistilRo model has gained F-1 score 89% in speech-to-text and the BiG model has gained F-1 score 90% in speech.

Keywords: bangla speech, OCEAN, NEO-FFI, personality classification, MoMF, MEWLP, DistilRo, BiG

Acknowledgement

My name is Md. Sajeebul Islam Sk., a student from the Computer Science and Engineering (CSE) department at BRAC University. This paper represents the culmination of two years of academic exploration and is presented as my final thesis. The focus of this work is to delve into the intricate relationship between Bangla speech and personality traits, aiming to consolidate the knowledge I have acquired during my academic journey.

I extend my deepest gratitude to my supervisor, Dr. Md. Golam Robiul Alam, for his unwavering support, invaluable guidance, and scholarly insights throughout the course of this research. His dedication to excellence and commitment to fostering a rich academic environment have been instrumental in shaping the trajectory of this study. I am indebted to Dr. Alam for his patience, encouragement, and the countless hours he devoted to mentoring me. His profound expertise and keen insights have been a beacon of inspiration, propelling me forward in my pursuit of knowledge.

I would like to express my heartfelt appreciation to my dear parents for their unconditional love, unwavering belief in my abilities, and constant encouragement. Their sacrifices, encouragement, and unwavering support have been the bedrock upon which this academic journey has been built. Their selfless dedication to my well-being and academic pursuits have been a source of strength and motivation.

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

Table of Contents

Declaration	i
Approval	ii
Ethics Statement	iv
Abstract	v
Acknowledgment	vi
Table of Contents	vii
List of Figures	ix
List of Tables	xii
Nomenclature	xiii
1 Introduction	1
1.1 Research Background	1
1.2 Research Scopes	2
1.3 Research Objectives	2
1.4 Research Contributions	3
1.5 Outline of Research	4
2 Literature Review	5
2.1 Big Five personality traits	5
2.2 NEO-FFI Questionnaire	6
2.3 The landscape of Bengali speech and personality linked studies	7
2.4 Feature extraction and dataset of personality linked studies	8
3 Methodology	11
3.1 Data Collection	12
3.2 Data Annotation	13
3.3 Reliability Analysis	14
3.4 Data Preprocessing and Feature Extraction	15
3.4.1 MFCCs	18
3.4.2 Mel-Frequency Cepstral Coefficients with Linear Predictive Coding (MELP)	32

3.4.3	Mel-Frequency Cepstral Coefficients with Wiener Linear Predictive Coding (MEWLP)	38
3.4.4	Morlet-based Mel-frequency Cepstral Coefficients (MoMF)	39
3.5	DistilRo and BiG: Soft Voting Ensemble Models for Personality Classification in Speech, and Speech-to-Text Modalities	40
3.5.1	Distilled Bidirectional Encoder Representations from Transformers (DistilBERT) and Robustly optimized BERT approach (RoBERTa)	41
3.5.2	DistilRo	42
3.5.3	Bidirectional Long Short-Term Memory (Bi-LSTM)	43
3.5.4	Gated Recurrent Unit (GRU)	44
3.5.5	BiG	44
4	Results and Discussions	46
4.0.1	Parameter Selection for Feature Extraction	46
4.1	Personality Classification using DistilRo	46
4.2	Personality Classification using BiG	47
4.2.1	MFCC base findings	49
4.2.2	MoMF base findings	52
4.2.3	MELP base findings	52
4.2.4	MEWLP base findings	58
5	Conclusion	62
	Bibliography	68

List of Figures

3.1	Toplevel Overview of the Proposed System	11
3.2	The graph shows the probability distribution of Agreeableness. On the left, green bars represent ratings for speech acted to low personality trait scores. On the right, there are purple bars, which represent ratings for speech acted to high personality trait scores.	15
3.3	The graph shows the probability distribution of Conscientiousness. On the left, green bars represent ratings for speech acted to low personality trait scores. On the right, there are purple bars, which represent ratings for speech acted to high personality trait scores.	16
3.4	The graph shows the probability distribution of Extroversion. On the left, green bars represent ratings for speech acted to low personality trait scores. On the right, there are purple bars, which represent ratings for speech acted to high personality trait scores.	16
3.5	The graph shows the probability distribution of Neuroticism. On the left, green bars represent ratings for speech acted to low personality trait scores. On the right, there are purple bars, which represent ratings for speech acted to high personality trait scores.	17
3.6	The graph shows the probability distribution of Openness. On the left, green bars represent ratings for speech acted to low personality trait scores. On the right, there are purple bars, which represent ratings for speech acted to high personality trait scores.	17
3.7	Lowerneurotic Waveplot	18
3.8	Lowerneurotic Plot of Spectrogram	19
3.9	Lowerneurotic Plot of MFCCs	19
3.10	Highneurotic Waveplot	20
3.11	Highneurotic Plot of Spectrogram	20
3.12	Highneurotic Plot of MFCCs	21
3.13	Highagree Waveplot	21
3.14	Highagree Plot of Spectrogram	21
3.15	Highagree Plot of MFCCs	22
3.16	Lowagree Waveplot	22
3.17	Lowagree Plot of Spectrogram	23
3.18	Lowagree Plot of MFCCs	23
3.19	HighConscientious Waveplot	24
3.20	HighConscientious Plot of Spectrogram	24
3.21	HighConscientious Plot of MFCCs	25
3.22	LowConscientious Waveplot	25
3.23	LowConscientious Plot of Spectrogram	25

3.24	LowConscientious Plot of MFCCs	26
3.25	HighExtrover Waveplot	26
3.26	HighExtrover Plot of Spectrogram	27
3.27	HighExtrover Plot of MFCCs	27
3.28	LowExtrover Waveplot	28
3.29	LowExtrover Plot of Spectrogram	28
3.30	LowExtrover Plot of MFCCs	29
3.31	HighOpen Waveplot	29
3.32	HighOpen Plot of Spectrogram	29
3.33	HighOpen Plot of MFCCs	30
3.34	LowOpen Waveplot	30
3.35	LowOpen Plot of Spectrogram	31
3.36	LowOpen Plot of MFCCs	31
3.37	Lowneurotic Plot of LPC	32
3.38	Highneurotic Plot of LPC	33
3.39	Lowagree Plot of LPC	34
3.40	Highagree Plot of LPC	34
3.41	Lowconscientious Plot of LPC	35
3.42	Highconscientious Plot of LPC	36
3.43	Lowextrover Plot of LPC	36
3.44	Highextrover Plot of LPC	37
3.45	Lowopen Plot of LPC	37
3.46	Highopen Plot of LPC	38
3.47	An overview of DistilBERT work flow	42
3.48	Structure of DistilRo Model	43
3.49	Structure of BiG Model	45
4.1	Confusion matrix of RoBERTa. RoBERTa effectively captures the Agreeable and Open personality traits. However, a discernible inconsistency of 17% in data alignment with Agreeable is observed within the Conscientious trait. Additionally, a 14% data mismatch is identified between Neurotic and Open traits.	48
4.2	Confusion matrix of DistilBERT. DistilBERT effectively captures the Extrover personality traits. However, 13% data mismatch is identified between Conscientious and Extrover traits. Additionally, a 14% data mismatch is identified between Neurotic and Open traits.	49
4.3	Confusion matrix of DistilRo. DistilRo highly captures the Agreeable, Conscientious, Neurotic, and Open personality traits. It encounters challenges in accurately representing the Extrover trait, particularly in establishing distinctions with Neurotic and Open personality traits.	50
4.4	Each Model Accuracy and F-1 score of Speech-to-text Modality	50
4.5	Confusion matrix of Bi-LSTM based on MFCCs. Bi-LSTM effectively capture all the personality traits.	52
4.6	Confusion matrix of GRU based on MFCCs. GRU effectively capture all the personality traits except LA, HO and LO. However, most data of LA are mismatch with LC, LN, and HA traits. Additionally, 52 and 40 data of HO and LO traits are considering as HN and HA traits respectively.	53

4.7	Confusion matrix of BiG based on MFCCs. BiG effectively capture all the personality traits except HO and LO.	53
4.8	Confusion matrix of Bi-LSTM based on MoMF. Bi-LSTM effectively capture all the personality traits but a discernible inconsistency of LA is observed alignment with LC.	54
4.9	Confusion matrix of GRU based on MoMF. GRU effectively capture all the personality traits except LA and HO. However, most data of LA are identifying as LC. Additionally, for HO trait data, model don't separate HO and HN traits adequately.	54
4.10	Confusion matrix of BiG based on MoMF. BiG effectively capture all the personality traits but model struggle with HO trait data.	56
4.11	Confusion matrix of Bi-LSTM based on MELP. Bi-LSTM effectively capture all the personality traits.	57
4.12	Confusion matrix of GRU based on MELP. GRU effectively capture all the personality traits except LO trait. Additionally, most of the data of LO trait are considering as HA trait.	57
4.13	Confusion matrix of BiG based on MELP. BiG correctly capture all the personality traits but struggle with LA trait. However, to identify LA trait it sometime consider LC trait.	59
4.14	Confusion matrix of Bi-LSTM based on MEWLP. Bi-LSTM effectively capture all the personality traits.	59
4.15	Confusion matrix of GRU based on MEWLP. GRU can't capture all the personality traits. However HA, HO, LA, LC, and LO are mismatch with LN, HN, LC, LE, and LN traits respectively.	60
4.16	Confusion matrix of BiG based on MEWLP. BiG effectively capture all the personality traits except LO.	60

List of Tables

3.1	Present an overview of the recordings	13
3.2	speech-to-text modality	13
3.3	speech modality	14
3.4	speech-to-text modality	15
3.5	Baseline Models parameters of DistilRo	42
3.6	Baseline Models parameters of BiG using MFCCs & MoMF	44
3.7	Baseline Models parameters of BiG using MELP & MEWLP	45
4.1	RoBERTa classification results	47
4.2	DistilBERT classification results	47
4.3	DistilRo classification results	47
4.4	Comparing F-1 score of three models	49
4.5	Bi-LSTM classification results based on MFCCs	51
4.6	GRU classification results based on MFCCs	51
4.7	BiG classification results based on MFCCs	51
4.8	Bi-LSTM classification results based on MoMF	55
4.9	GRU classification results based on MoMF	55
4.10	BiG classification results based on MoMF	55
4.11	Bi-LSTM classification results based on MELP	56
4.12	GRU classification results based on MELP	58
4.13	BiG classification results based on MELP	58
4.14	Bi-LSTM classification results based on MEWLP	61
4.15	GRU classification results based on MEWLP	61
4.16	BiG classification results based on MEWLP	61

Nomenclature

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

- BERT Bidirectional Encoder Representations from Transformers
- Bi-LSTM Bidirectional Long Short-Term Memory
- BiG Bidirectional Long Short-Term Memory & Gated Recurrent Unit
- DistilBERT Distillable Bidirectional Encoder Representations from Transformers
- DistilRo Distillable Bidirectional Encoder Representations from Transformers & Robustly optimized BERT approach
- GRU Gated Recurrent Unit
- LPC Linear Predictive Coding
- MELP Mel-Frequency Cepstral Coefficients with Linear Predictive Coding
- MEWLP Mel-Frequency Cepstral Coefficients with Wiener Linear Predictive Coding
- MFCCs Mel-Frequency Cepstral Coefficients
- MoMF Morlet-based Mel-frequency Cepstral Coefficients
- NEO-FFI NEO Five-Factor Inventory
- RoBERTa Robustly optimized BERT approach

Chapter 1

Introduction

Personality is like the fingerprint of our inner selves. An individual is a unique combination of traits, behaviors, and characteristics [6]. Some of us light up in social gatherings, while others find solace in quieter moments. Our interests, the way we react to challenges, our sense of humor – they’re all threads that weave together into the beautiful tapestry of who we are [7]. Nature gives us a head start with certain traits, but life experiences add their own splash of color to the canvas [22]. Personality remains a dynamic and evolving essence, a key to understanding ourselves and connecting with others on a profound level.

1.1 Research Background

Speech serves as a potent medium for the expression of a multitude of psychologically significant phenomena. For instance, within a matter of a few hundred milliseconds of encountering speech, humans have the remarkable ability to consistently deduce an extensive array of details about the speaker [43]. Beyond mere directed dialogues and basic command-and-control interfaces, the realm of voice-based human-machine interaction is broadening. Machines must now possess the capability to comprehend inputs and generate responses within a distinct context and it is influenced by numerous factors, prominently the voice quality, necessitating a more intricate level of interpretation and output generation [14]. The modern landscape is reminiscent of the vivid characters that inhabit the pages of famous novels, each showcasing a unique facet of human nature. The increasing number of online platforms including news portals, social media, and blogs [49] make it easier for people to raise their voices on a multitude of subjects. Various systems have been suggested to characterize an individual’s personality. A substantial number of these systems revolve around the framework of the Big Five personality traits [1]. This model endeavors to depict a person’s personality by utilizing five distinct factors, somewhat akin to a vector.

Previously, many methodologies have utilized a variety of lexicons, linguistic elements, psycholinguistic factors, and emotional attributes within a supervised learning framework to ascertain a user’s personality from their textual and spoken interactions. These approaches have employed a spectrum of learning models, ranging from conventional SVM [14], KNN, BPT, TF-IGM[57], naive Bayes, and so on, to the contemporary deep learning strategies like CNN, MLP [49], Bi-LSTM [61], GRU [28] and so on. In this study, we harness the power of state-of-the-art language mod-

els to address the intricacies of personality. Specifically, we leverage the capabilities of BoBERTa (Bidirectional Encoder Representations from Transformers) [41] and DistilBERT [50], two prominent transformer-based models that have demonstrated exceptional prowess in natural language understanding tasks. BoBERTa excels in capturing bidirectional contextual information, enabling nuanced comprehension of linguistic nuances [46]. On the other hand, DistilBERT strikes a balance between computational efficiency and performance, making it an attractive choice for tasks with resource constraints [52]. In [65], the authors harnessed the power of RoBERTa for capturing semantics and contextual information from YouTube comments in both English and Russian languages. Furthermore, in [64], the authors employed DistilBERT to undertake the classification of personality traits within the realm of social media. Ensemble methods represent learning algorithms that build a collection of classifiers and subsequently categorize novel data points by aggregating their predictions through a weighted voting mechanism [2]. The synergy achieved through ensemble methods not only mitigates the limitations of individual models but often results in superior generalization and resilience to noise. The majority of these techniques have primarily been developed for languages such as English, German [14], and others. To the best of our understanding, there exists a solitary prior effort and dataset in Bangla that dealt with the intricacies of detecting personality from social media Bangla text [49]. However, no previous endeavor or dataset is accessible for identifying personality traits from Bangla speech. As a result, it becomes imperative to bridge this gap in research and establish resources that can pave the way for future investigations in this domain.

1.2 Research Scopes

As far we have studied, no prior work on Bangla speech base personality traits classification. In [14] utilize traditional Machine Learning model and statistical features. Therefore, their proposed model is unable to extract depth level discriminative features to successfully classify speech base personality traits. There exist one Bangla paper that work on text base personality classification. In [49] refers text based personality classification based on Facebook and YouTube comments in Bangla Language and utilize TF-IDF base feature extraction and deep learning models for text base personality traits classification. However, this research is unable to grab the semantic relationship among the words. Therefore, the results are not impressive.

1.3 Research Objectives

Accessing personality from speech is crucial in understanding and interpreting human behavior in a more nuanced manner. Speech serves as a rich medium for conveying not only verbal content but also various non-verbal cues and nuances that reflect an individual's unique personality traits. By tapping into the acoustic features, intonations, and patterns within speech, we gain insights into aspects of personality that may not be easily discernible through other means. This capability has significant implications across diverse fields, from improving human-machine interactions and designing personalized user experiences to enhancing mental health

assessments and refining communication strategies in various professional settings. Ultimately, accessing personality from speech broadens our understanding of individuals, fostering more tailored and effective approaches in both technological and interpersonal contexts.

Our research focuses on evaluating personality traits using a concise dataset, specifically aiming to classify the personality of individuals based on short Bangla speech. Given the absence of pre-existing datasets for assessing personality from Bangla speech, we undertake the creation of our dataset. To achieve this, we engage a speaker to generate the necessary data, following two distinct trajectories: speech-to-text modality and speech modality.

In the speech modality, we apply feature extraction techniques and obtain initial predicted outcomes. Subsequently, we employ a feature level fusion technique, incorporating Morlet wavelet instead of the conventional Fast Fourier Transform in the Mel-Frequency Cepstral Coefficients (MFCCs) architecture. This involves the utilization of a Morlet low pass filter, capturing both frequency and time-domain information simultaneously, which is an enhancement over the in-built MFCCs that primarily capture frequency information.

In the training phase, we utilize Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU), and Bidirectional LSTM (Bi-LSTM). Ensemble methods are then applied to GRU and Bi-LSTM, aiming to enhance result accuracy. In the speech-to-text modality, we convert speech to text using Speech Recognition (SR) libraries. Subsequently, we utilize the in-built tokenizer of individual transformers and fine-tune a transformer model. Similar to the speech modality, we apply ensemble methods to improve the accuracy of the results in the model. The outcomes of our study demonstrate state-of-the-art performance compared to earlier research endeavors.

1.4 Research Contributions

The research presents a set of noteworthy contributions:

- The absence of a pre-existing dataset tailored for personality detection from Bangla speech necessitated the creation of our own dataset. This dataset stands out as a distinctive resource, encompassing speech samples derived from non-professional speakers. These scripts are meticulously curated from online Bangla newspapers and excerpts from famous Bangla novels. Importantly, each of these speech segments is diligently annotated with the quintessential Big Five NEO-FFI personality traits.
- Introduce Morlet-based Mel-frequency Cepstral Coefficients (MoMF) feature extraction method by transforming Morlet wavelet and fusing MFCCs feature.
- In this study, a key focus lies on an exploration of personality traits as assessed by the NEO-FFI questionnaire, considering both acted and non-acted speech. To ensure the reliability of our findings, we first computed Cronbach's alpha to gauge the internal consistency of individual traits within the NEO-FFI.

Additionally, we delved into the probability density distribution of NEO-FFI ratings for acted speech.

- We have proposed two distinct classification methodologies. The first one is speech-to-text multi-class classification, where we apply an ensemble model called DistilRo. DistilRo is developed based on DistilBERT and Roberta. The second approach involves speech multi-class classification, where we developed a feature extraction technique called MELF based on MFCC and LPC. We also apply an ensemble model called BiG. BiG is developed based on Bi-LSTM and GRU. This duality in classification methodologies enriches the depth of our research, catering to the intricate nuances of personality assessing in Bangla speech.

1.5 Outline of Research

The subsequent sections of this research study include the following.

In Section II, a comprehensive exposition unfolds, elucidating the intricacies of the personality test employed in this study. The examination and contextualization of this test are crucial for grasping the framework within which personality traits are assessed, providing readers with a solid foundation for the subsequent analyses.

Section III takes center stage, unveiling our distinctive dataset and introducing the model deployed in this research. A detailed exploration of the dataset's uniqueness and the model's architecture offers transparency into the methodologies applied, setting the stage for a nuanced understanding of the subsequent analytical processes.

Moving on to Section IV, a meticulous analysis of the results derived from our model is presented. This section serves as the empirical heart of the paper, where findings are scrutinized, patterns are identified, and correlations are unveiled, contributing substantively to the overarching goal of unraveling personality traits through Bangla speech.

Finally, Section V draws the threads together, offering a comprehensive conclusion to the paper. It encapsulates the key findings, highlights the significance of the research, and responsibly reports on its limitations. This structured approach ensures a cohesive narrative that guides the reader through the journey from methodology to conclusion, enhancing the overall coherence and impact of the study.

Chapter 2

Literature Review

In line with the foundational Big Five personality traits concept, as elucidated in [1], we have employed the Bangla version of the NEO-FFI personality inventory [40]. Our application involves the evaluation of vocal impressions derived from speech, with each of the 5 fundamental traits.

2.1 Big Five personality traits

Based on [14], we provide an overview of the general characteristics associated with these personality traits, as assessed by the NEO-FFI inventory:

Agreeableness (A): When individuals achieve high scores in agreeableness, it often signifies their natural inclination towards empathy and trustworthiness. They tend to place trust in others readily and are often eager to offer assistance. Conversely, those who score lower in agreeableness may tend to display traits such as egocentrism, competitiveness, and a predisposition towards skepticism and distrust.

Conscientiousness (C): Individuals with high conscientiousness scores are generally recognized for their precision, attentiveness, reliability, and effective planning skills. Conversely, individuals with low conscientiousness scores may display carelessness, thoughtlessness, and a tendency to act imprudently.

Extroversion (E): Individuals with high extroversion scores are often characterized by their outgoing, sociable nature, marked by a propensity for sociability and enthusiasm. They tend to thrive in independent roles and exhibit a vibrant, energetic demeanor. Conversely, individuals leaning towards introverted tendencies are frequently observed as reserved, deep in thought, and inclined towards more conservative outlooks.

Neuroticism (N): Individuals scoring high in neuroticism tend to exhibit emotional instability and are often susceptible to feelings of shock or embarrassment. They are easily overwhelmed by emotions and may lack self-confidence. Conversely, individuals with low neuroticism scores are generally characterized as composed and emotionally stable. They excel under pressure and remain unflustered.

Openness (O): Scores on the openness factor reflect an individual's receptivity

to new ideas and their willingness to embrace novel experiences in daily life. High scorers are often described as visionary and curious, with an openness to adventurous experimentation. In contrast, individuals with low scores tend to lean toward conservatism, favoring conventional wisdom over avant-garde thinking.

2.2 NEO-FFI Questionnaire

The NEO Five-Factor Inventory (NEO-FFI) is a personality inventory that assesses general personality using the five-factor model [12].

In our research study, we integrated a subset of 50 statements from the NEO-FFI questionnaire, a well-established and widely used personality assessment tool. The NEO-FFI is derived from the Revised NEO Personality Inventory, designed to evaluate the fundamental dimensions of personality, commonly known as the "Big Five" traits. These traits encompass Neuroticism, Extraversion, Openness, Agreeableness, and Conscientiousness, providing a comprehensive framework for understanding and categorizing individual differences in personality.

Our system is a fusion of the NEO-FFI questionnaire and innovative technological components tailored specifically for the analysis of personality traits through Bangla speech. The inclusion of a unique dataset, combined with a model designed for Bangla speech analysis, underscores the interdisciplinary nature of our study. This synthesis of established psychological measurement tools with state-of-the-art linguistic analysis techniques positions our research at the forefront of exploring the intricate relationship between language and psychology.

The system utilizes advanced linguistic analysis methods, including natural language processing and machine learning algorithms, to decipher the linguistic patterns embedded in Bangla speech. By leveraging this technology, we aim to unveil subtle connections between specific linguistic features and the underlying personality traits of the speakers. This approach not only enhances the applicability of traditional personality assessments but also contributes novel insights to the broader field of personality research.

Our raters construct an individual's personality profile by responding to 50 statements from the NEO-FFI questionnaire, utilizing a scale that ranges from 'strongly disagree' to 'strongly agree,' which corresponds to numeric values between 1 to 5. Each of the five personality factors can yield a score within the range of 0 to 50. The combination of these scores results in a comprehensive personality profile for the individual. The reliability of the questionnaire is consistently high, with intra-scale consistency coefficients, as measured by Cronbach's Alpha, consistently exceeding 0.8.

Feature level fusion, employed in this study, integrates distinctive information from multiple sources or modalities at the feature level, enhancing the overall performance by merging diverse and complementary characteristics. This approach is a comprehensive understanding of the underlying data, promoting synergy and improved representation for robust analysis and decision-making [4]. For instance, in

the study outlined in [25], the author employs the Mel-frequency cepstral coefficients (MFCCs) and linear predictive coding (LPC) feature extraction techniques to evaluate personality traits from speech data. Another notable approach, highlighted in [56], involves the utilization of Morlet wavelet for the classification of electroencephalogram (EEG) signals. Furthermore, in [27], the author tackles the challenge of noise reduction by employing the Wiener filter and Spectral Subtraction techniques. Subsequently, a combination of LPCs, MFCCs, and Linear Prediction Cepstral Coefficients (LPCC) is utilized for feature extraction in this multi-faceted exploration of signal processing methods.

In [63], the authors employed ensemble methods for the classification of context and emotion in political speech, achieving accuracies of 73% and 53%, respectively. Utilizing an ensemble approach, authors integrated insights from personality recognition in text [21], [53], and speech [17], [58], demonstrating the effectiveness of this combined strategy in achieving high performance across their tasks. The progress in this area for Bangla language was not possible for the lack of dataset. The NEO-FFI questionnaire [8] is meant for people to assess themselves or for others who know them well to assess them. But in our experiments, we're not using it to judge a person's entire personality. Instead, raters assess the immediate vocal impression of an unfamiliar speaker listening to just a few seconds of speech. We expect the ratings to be consistent even with this short listening time.

2.3 The landscape of Bengali speech and personality linked studies

For the development of a voice search module for the Pipilika search engine [33], various Deep Neural Network (DNN-HMM) and Gaussian Mixture Model-Hidden Markov Model (GMM-HMM) models are explored for Bengali speech recognition. The study assembles a corpus of 9 hours of voice recordings from 49 speakers, contributing to the advancement of voice-enabled search functionality.

In the realm of sentiment recognition and emotion extraction [34], the authors present an extensive collection of methods for Bangla texts. Deep learning-based models are developed for categorizing phrases into three-class and five-class sentiment labels and identifying fundamental emotions. The effectiveness of these models is assessed using a fresh dataset of comments from various YouTube videos in Bengali, English, and Romanized Bengali.

Additionally, efforts have been made to create open-source Bengali corpora for sentiment analysis [44] and hate speech detection [35]. The sentiment analysis corpus comprises over 10,000 texts annotated for sentiment polarity, with an additional word corpus annotated for sentiment polarity. The hate speech detection study involves data collection from social networks, followed by classification using machine learning methods such as SVM and Naive Bayes.

In the domain of speech recognition [45], a convolutional neural network (CNN) is employed to create a Bengali number recognition system from spoken streams. Mel

Frequency Cepstrum Coefficient (MFCC) analysis is utilized for feature extraction from speech signals, with the trained CNN utilizing these characteristics for effective recognition of isolated Bengali digits.

Identifying Bengali broadcast speech is addressed in [51], where support vector machines (SVM) are employed with a linear kernel on the MATLAB platform. The study achieves promising results in distinguishing between different forms of noisy broadcast voice samples.

Another research endeavor [59] focuses on gathering data from Bengali comments on social media platforms, aiming to develop a classifier capable of swiftly distinguishing between social and anti-social remarks. Using supervised machine learning classifiers, including Logistic Regression, Random Forest, Multinomial Naive Bayes, and Support Vector Machine, alongside neural network models such as Gated Recurrent Unit (GRU), the study examines 2000 comments from Facebook and YouTube. Language models, incorporating unigrams, bigrams, and trigrams, enhance the classification process.

In [60] studies contribute significantly to the evolving landscape of Bengali Natural Language Processing (BNLP), covering diverse subfields such as sentiment analysis, speech recognition, optical character recognition, and text summarization. A critical analysis of contemporary BNLP tools and techniques is notably absent in available resources, prompting the authors to conduct an in-depth examination of 75 BNLP research papers, categorizing them into 11 distinct areas.

One notable study [55] introduces a deep learning-based approach for speech emotion identification, leveraging a combination of a time-distributed flatten (TDF) layer, a deep convolutional neural network (DCNN), and a bidirectional long-short-term memory (BLSTM) network. This model, evaluated on the SUBESCO Bengali emotional speech corpus, outperforms cutting-edge convolutional neural network (CNN)-based speech emotion recognition models, demonstrating superior temporal and sequential emotion representation.

The paper [49] addresses the growing interest in utilizing online platforms for expressing opinions and thoughts, recognizing the significance of user-generated content in studying and modeling personality traits. To design effective systems like recommendation systems, Q/A systems, employee assessments, and product promotions, detecting and analyzing user personality.

2.4 Feature extraction and dataset of personality linked studies

In [62] studies the dataset comprises recorded clinical diagnostic interviews (CDI) from 79 patients diagnosed with major depressive disorder, each classified into analytic and introjective personality styles. Feature extraction involves analyzing linguistic features associated with each style, utilizing standardized questionnaire responses, basic text features (TF-IDF scores), advanced text features using LIWC

(linguistic inquiry and word count), and context-aware features employing BERT (bidirectional encoder representations from transformers). Notably, automated classification based on LIWC outperforms questionnaire-based models, with the best performance achieved by combining LIWC with questionnaire features, suggesting the potential of linguistically based automated techniques for characterizing personality in psychopathological contexts.

In [49] works predominantly focus on personality detection from user-generated text in English, this paper introduces a pioneering effort in the realm of Bangla language. The research contributes a benchmark Bangla personality traits detection dataset, consisting of 3000 informal Bangla texts sourced from diverse online platforms, and presents baseline systems leveraging advanced supervised classification methods for a comprehensive performance analysis.

The paper [48] introduces a novel TB-APR (Text-Based Automatic Personality Recognition) approach by utilizing a projective test to construct a corpus for personality computing research. Unlike conventional personality inventories, which may exhibit limitations in controlling intentional or non-conscious omissions of undesired personality traits, the proposed model employs the Z-test projective instrument for labeling a textual corpus.

Additionally [14] utilizes German and English language datasets for the application of a personality assessment paradigm to speech input. The professional speaker is cued to produce speech with different personality profiles, and the resulting vocal personality impressions are encoded using the "Big Five" NEO-FFI personality traits. Human raters, unfamiliar with the speaker's identity, assess these traits based on the recordings. Signal-based acoustic and prosodic methods are then employed for analysis, revealing high consistency among the acted personalities, human raters' assessments, and initial automatic classification results. This marks a significant step toward incorporating personality traits into speech for potential use in voice-based communication between humans and machines.

The paper [67] aims to develop a technique for deducing a user's personality traits based on social media profiles, acknowledging the value of users' contributions and emotions in status updates for studying human behavior. The proposed method utilizes LSTM-CNN, with pre-processed and vectorized text documents fed into the model. Feature extraction employs SpectralNet Features (SNF), and feature selection is carried out using Correlation-based Feature Selection (CFS). However, the focus is on leveraging the LSTM-CNN model, which combines the strengths of CNN for extracting time-independent features and LSTM for capturing long-term dependencies.

Addresses the challenge of automatically inferring users' [30] personality from social network activities, emphasizing the critical role of data representation in the performance of such approaches. Deep learning methods are employed to autonomously learn effective data representations for personality recognition. The experiments conducted in the study utilize Facebook status updates data as the dataset. Several neural network architectures, including fully-connected (FC) networks, convo-

lutional networks (CNN), and recurrent networks (RNN), are investigated and compared with shallow learning algorithms in the context of the myPersonality shared task.

Explores emotion and personality detection from text [66], emphasizing its novelty as a sub-field of artificial intelligence closely tied to Sentiment Analysis (SA). In contrast to SA’s focus on positive, neutral, or negative sentiments, emotion analysis discerns specific emotions like disgust, fear, anger, happiness, surprise, and sadness expressed in text. Simultaneously, the article delves into the critical psychological concept of personality, aiming to efficiently and reliably identify and validate an individual’s unique characteristics. The review encompasses approaches in developing text-based emotion and personality detection systems, shedding light on the studies’ contributions, methodologies, datasets, and conclusions, along with their respective strengths and limitations.

In [20], conducted a comprehensive analysis of 700 million words, phrases, and topic instances extracted from Facebook messages of 75,000 volunteers. These volunteers had also taken standard personality tests, allowing for the exploration of language variations correlated with personality, gender, and age. The open-vocabulary technique employed in the analysis allows the data itself to drive an extensive exploration of language, uncovering connections not captured by traditional closed-vocabulary word-category analyses. The dataset used in this paper comprises Facebook messages, representing the largest study, by an order of magnitude, of language and personality to date.

Addressing [39] challenges associated with working with spatial data in the context of personality traits and their regional clustering. Two main challenges, the Modifiable Aerial Unit Problem and spatial dependencies, are tackled using data-analytic techniques specifically designed for spatial data. The research provides practical guidelines for working with spatial data in psychological research and explores the robustness of regional personality differences and their correlates.

Chapter 3

Methodology

Our research had five phases to classify personality from Bangla speech. We began with data collection (Phase 1). In Phase 2, we annotated this data to understand personality traits. Phase 3 involved reliability checks to ensure the accuracy of our annotations. Phase 4 involved data preprocessing and feature extraction. Lastly, in Phase 5, we developed two soft voting ensemble models called DistilRo and BiG. In Figure 3.1, we show an overview of our work.

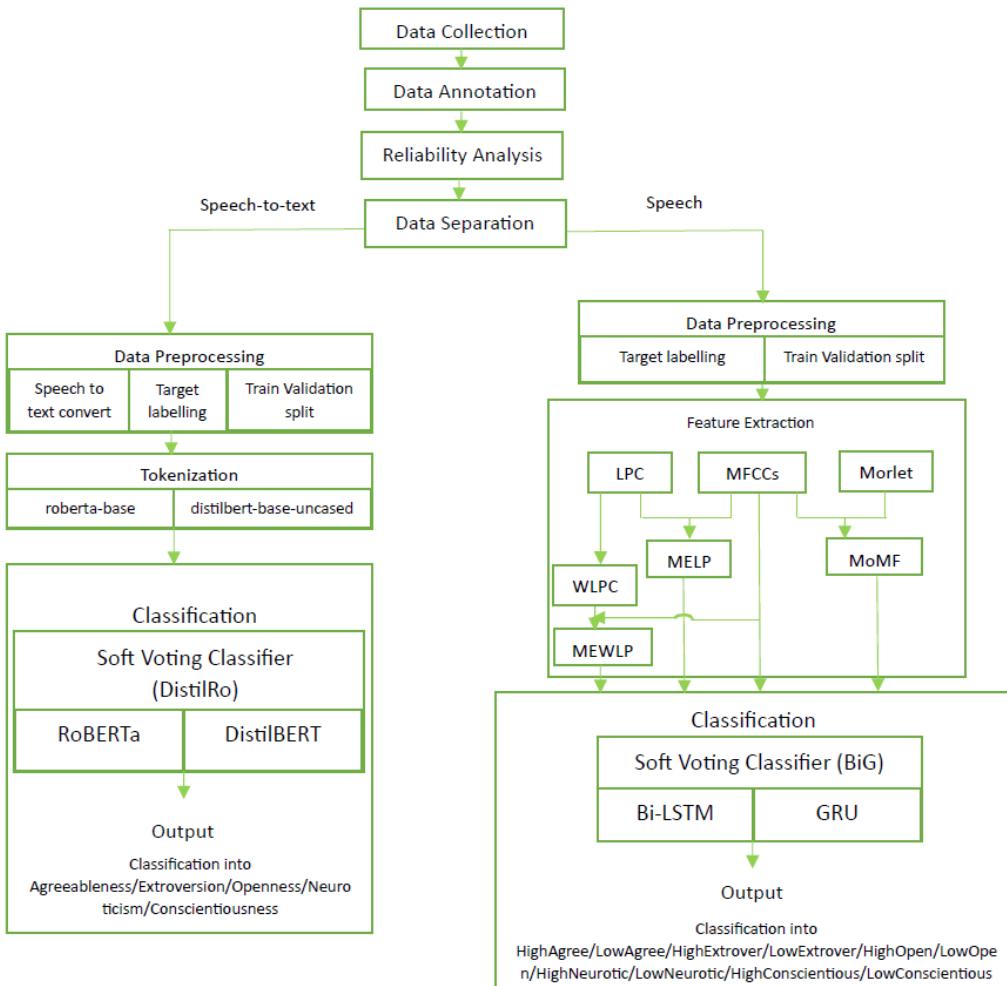


Figure 3.1: Toplevel Overview of the Proposed System

3.1 Data Collection

The majority of prior investigations conducted on various native languages, including English [19][11], German [5], French [23], Spanish [29], Mandarin [13], and Hindi [42], have predominantly relied upon data collection from online newspapers and renowned novels.

Based on these observations, we conducted a data collection process that involved gathering single-sentence text from various popular online Bangla newspapers and renowned Bangla novels. To obtain relevant content, we employed an empirical approach, using specific keywords aligned with the Big Five personality model. Following this initial data gathering phase, we took additional precautions to ensure the quality and relevance of the text. We enlisted graduate students from Bangla Department, who possessed a deep understanding of the Big Five personality traits, to review and verify the collected text. They were familiar with the Big Five personality traits and made sure the text matched our criteria. This process helped maintain the correct semantic information in our data. We obtained 90% of the text from various online Bangla newspapers, while the remaining 10% was sourced from well-known Bangla novels.

To prepare a realistic speech corpus for our experiments in the speech-to-text modality, we enlisted a non-professional speaker. The speaker was given the task of immersing himself in the NEO-FFI descriptions, which represent 5 personality profiles. The speaker recorded a 'natural' version of a predefined text, speaking in his ordinary manner without acting. We instructed speaker to perform this imitation at least 7 times for each text. All the audio recordings were conducted in a room that prevents external noise. Then we carefully examined all the audio recordings and selected the best 5 audio samples for each text. The complete audio database comprises a total of 1750 recorded files.

In the speech modality, we utilized 71% of the text from our predefined material. We provided specific instructions to our speaker to enact 10 distinct personality variations, resulting in the desired speech recordings. These instructions were crafted following the guidelines outlined in the NEO-FFI manual and Section II. The actor's performance was aimed at portraying extreme ends of the personality factors. To illustrate, for the trait of agreeableness, we directed the speaker to imitate both a highly agreeable person and a significantly less agreeable person. Since the speaker was not a professional, there was a higher likelihood of obtaining authentic and natural speech samples. We ensured that the speaker carried out this imitation process a minimum of 6 times for each text. We followed same procedure for recordings audio like before. Afterward, we carefully reviewed the audio and selected the best 4 recordings for each text. As a result of this effort, our audio database contains a total of 1000 recorded files. Similar approach of [25], we provides a summary of all the recorded data and conditions in TABLE 3.1.

Table 3.1: Present an overview of the recordings

	speech-to-text modality	speech modality
Predefine text	Yes	Yes
Domain	Newspapers and novels	Newspapers and novels
Speaker-dependency	Yes	Yes
Acted / non-acted	Non-acted	Acted
Linguistic diversity	”Short Text”	”Short Text”
Dataset size	$\approx 5\text{h}$	$\approx 3\text{h}$
Number of speaker	1	1
Audio capturing quality	44.1 KHz, mono	44.1 KHz, mono

3.2 Data Annotation

In the speech-to-text modality, we enlisted the assistance of 5 graduate students who were knowledgeable about the Big Five personality traits and were not known about the speaker. Each of them was allocated 350 randomly selected recorded audio files along with NEO-FFI questionnaires. These students carefully listened to each audio file multiple times and completed the questionnaires. Each question in the questionnaire offered five response options, ranging from ”strongly disagree” to ”strongly agree,” corresponding to numeric values between 1 to 5.

Subsequently, we calculated numerical values for each audio file based on the NEO-FFI questionnaire responses. The audio file that generated the highest value for a particular trait was selected as representative of that trait.

In the speech modality, we engaged 5 graduate students who were also well-versed in the Big Five personality traits. Similar to the speech-to-text phase, each of them was provided with 1000 recorded audio files along with NEO-FFI questionnaires and they were unknown to one another. These students followed the same procedure, listening to the audio files and completing the questionnaires to determine the personality traits represented in each audio recording. Following the annotation phase, our findings are presented in both TABLE 3.2 and TABLE 3.3

Table 3.2: speech-to-text modality

Label	Number of data
Agreeableness	350
Extroversion	350
Openness	350
Neuroticism	350
Conscientiousness	350

Table 3.3: speech modality

Label	Number of data
HighAgree	100
LowAgree	100
HighExtrover	100
LowExtrover	100
HighOpen	100
LowOpen	100
HighNeurotic	100
LowNeurotic	100
HighConscientious	100
LowConscientious	100

3.3 Reliability Analysis

We assess the reliability of the personality trait measures used in our research. The reliability analysis is crucial in ensuring that the measurement items consistently capture the underlying personality constructs. Cronbach's alpha [15], a well-established measure of internal consistency, was employed to evaluate the reliability of each personality trait scale utilized in our research. A high Cronbach's alpha value, typically ranging from 0.8 and higher, indicates strong internal consistency. This suggests that the items within the scale are closely related and collectively contribute to a reliable measurement of the intended variable [24].

During the speech-to-text modality, we took the numeric values that had been previously calculated for each audio file in the data annotation phase. These numeric values represented various aspects of personality traits. We used these numbers to compute Cronbach's alpha for each personality trait. In TABLE 3.4, we provide an overview of the Cronbach's alpha values obtained from our data.

In the speech modality where we used the NEO-FFI questionnaire to assess personality through speech, we assigned ratings using a five-point Likert scale [25]. Since we had 6 assessors for doing this assessment, and the total score for each personality trait could range from 0 to 50. To provide a clear sense of how ratings on this scale correspond to the various personality traits, we created histograms for each of the Big 5 personality traits and a Gaussian distribution curve in Figure 3.2,3.3,3.4,3.5,4.1. This figure simplifies a direct comparison between the ratings for high traits (depicted in purple) and low traits (depicted in green). If the difference between high and low traits was not noticeable, the Gaussian curves would mostly overlap. However, we observe that, for certain traits like extroversion, there is only a small area of overlap, indicating a clear distinction between high and low trait

Table 3.4: speech-to-text modality

Traits	Cronbach's alpha
Agreeableness	0.83
Extroversion	0.86
Openness	0.83
Neuroticism	0.85
Conscientiousness	0.81

ratings. In contrast, for other traits such as openness, there is more overlap between the two, suggesting that distinguishing between high and low trait ratings is less straightforward in this dataset. So, some pairs of high and low personality traits are more easily distinguishable based on the data, while others exhibit more similarity.

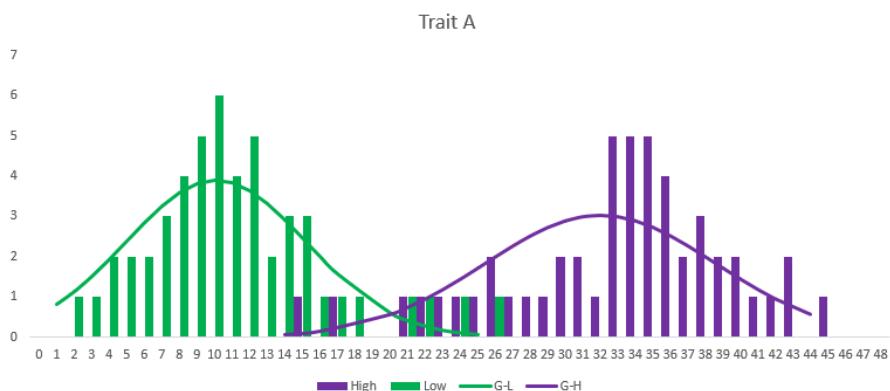


Figure 3.2: The graph shows the probability distribution of Agreeableness. On the left, green bars represent ratings for speech acted to low personality trait scores. On the right, there are purple bars, which represent ratings for speech acted to high personality trait scores.

3.4 Data Preprocessing and Feature Extraction

We performed data preprocessing in two distinct modalities to prepare our audio dataset for analysis.

In the first modality (Speech-to-Text), we focused on converting audio data into text format, which is essential for subsequent text-based analysis. The labels for the audio files were determined based on the directory structure of the dataset. To convert audio to text, we employed the SpeechRecognition library. This library allowed us to transcribe spoken words and convert them into a textual representation, making the data accessible for text-based processing. For tokenization, we used built-in DistilBERT tokenizer and the RoBERTa tokenizer.

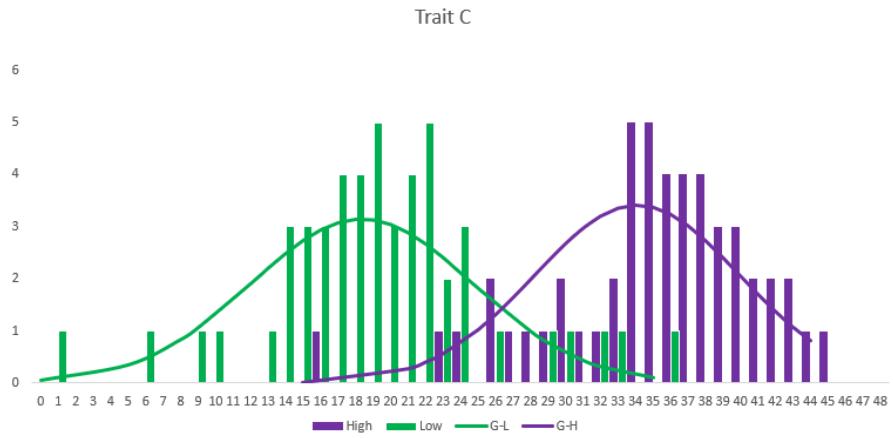


Figure 3.3: The graph shows the probability distribution of Conscientiousness. On the left, green bars represent ratings for speech acted to low personality trait scores. On the right, there are purple bars, which represent ratings for speech acted to high personality trait scores.

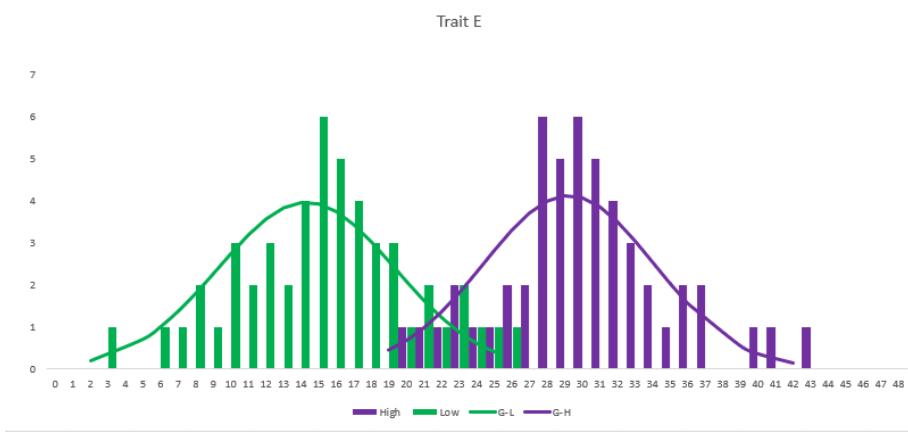


Figure 3.4: The graph shows the probability distribution of Extroversion. On the left, green bars represent ratings for speech acted to low personality trait scores. On the right, there are purple bars, which represent ratings for speech acted to high personality trait scores.



Figure 3.5: The graph shows the probability distribution of Neuroticism. On the left, green bars represent ratings for speech acted to low personality trait scores. On the right, there are purple bars, which represent ratings for speech acted to high personality trait scores.

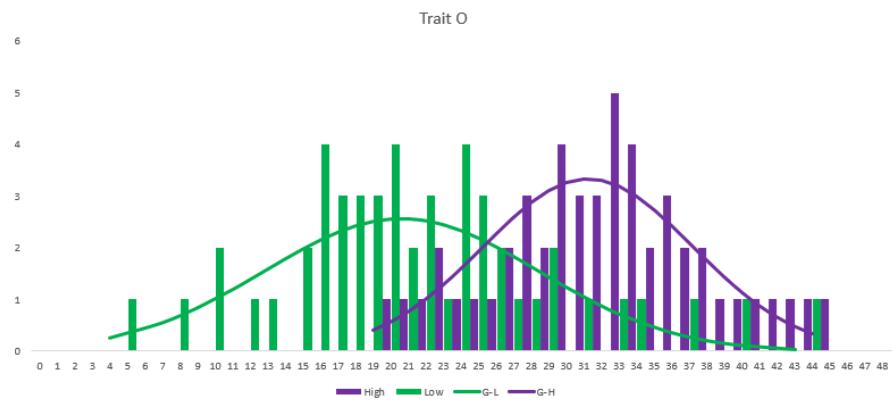


Figure 3.6: The graph shows the probability distribution of Openness. On the left, green bars represent ratings for speech acted to low personality trait scores. On the right, there are purple bars, which represent ratings for speech acted to high personality trait scores.

In the second modality (Speech Processing), We used audio files that were in WAV format. Similar to speech-to-text, we determined the labels for the audio files were determined based on the directory structure of the dataset. In our approach, we employ four feature extraction techniques: MFCCs, MELP, MEWLP, and MoMF.

3.4.1 MFCCs

For feature extraction, we turned to "Mel-frequency cepstral coefficients" (MFCCs) that captured a snapshot of the audio's acoustic characteristics and translates them into numerical data. MFCCs transform the speech signal into a frequency domain using Fourier Transform [16]. It divides the audio signal into small time frames and for each frame it calculates the energy. Mel scale is then applied to approximate the perception of pitch and the intensity of sound is converted to a logarithmic scale. Then the data is processed using DCT to convert the information into a set of coefficients and these coefficients capture the unique features of the sound [32]. From [32], we can define the MFCCs formula:

$$m(f) = 2595 * \log_{10}(1 + \frac{f}{700}) \quad (3.1)$$

$$\hat{a}_m = \sum_{m=1}^l (\log \hat{o}_l) \cos[m(l - \frac{1}{2}) \frac{\pi}{l}] \quad (3.2)$$

where f is frequency(Hz), l is number of mel ceptrum coefficients, \hat{o}_l is filterbank output, and \hat{a}_m is MFCCs coefficient.

For visual understanding,

we provide one lowneurotic audio sample's waveplot, spectrogram, and mfccplot in Figure 4.2,4.4,4.5 respectively.

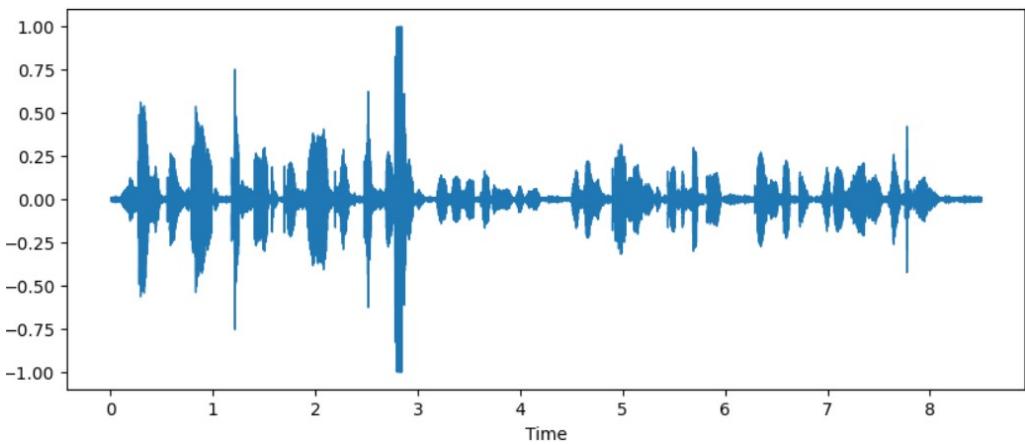


Figure 3.7: Lowerneurotic Waveplot

In 4.4, the x-axis represents time in seconds, and the y-axis represents frequency in Hertz. The colors represent the amplitude of the signal at each frequency and time, with blue being the lowest amplitude and red being the highest amplitude. The graph shows a series of vertical lines, indicating a repeating pattern in the signal.

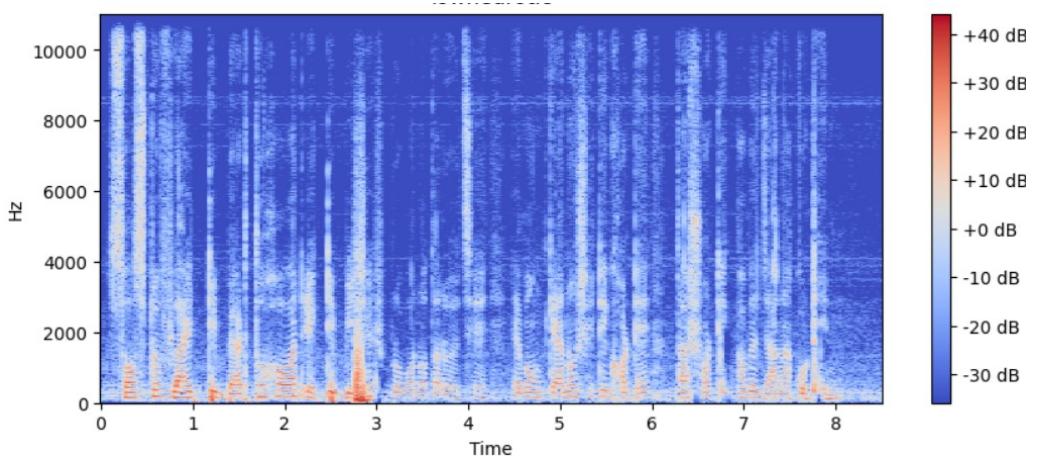


Figure 3.8: Lowerneurotic Plot of Spectrogram

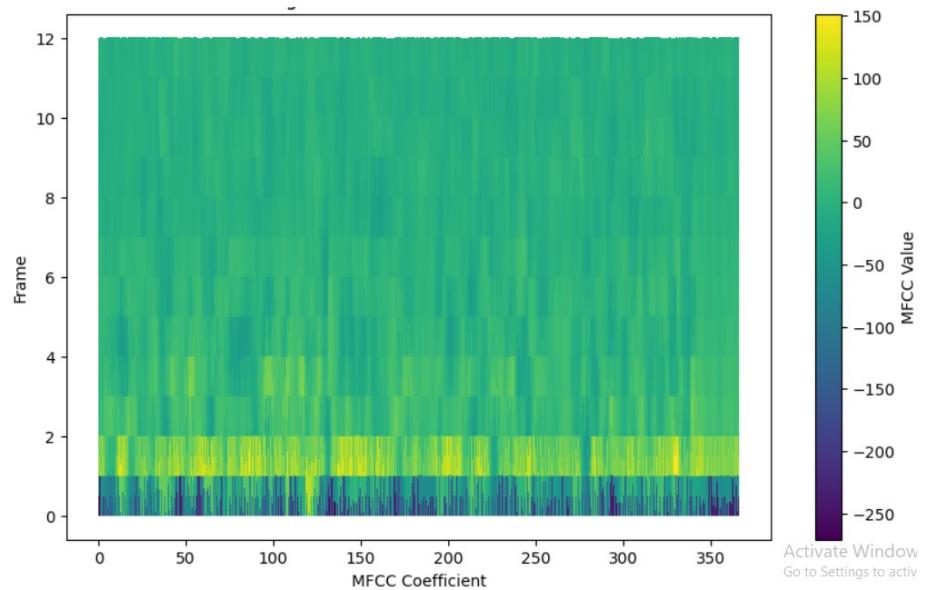


Figure 3.9: Lowerneurotic Plot of MFCCs

In 4.5, the x-axis is labeled “MFCC Coefficient” and ranges from 0 to 350. The y-axis is labeled “Frame” and ranges from 0 to 12. The color scale is on the right side of the graph and ranges from -250 to 150.

We provide one highneurotic audio sample’s waveplot, spectrogram, and mfccplot in Figure 4.6,3.11,3.12 respectively.

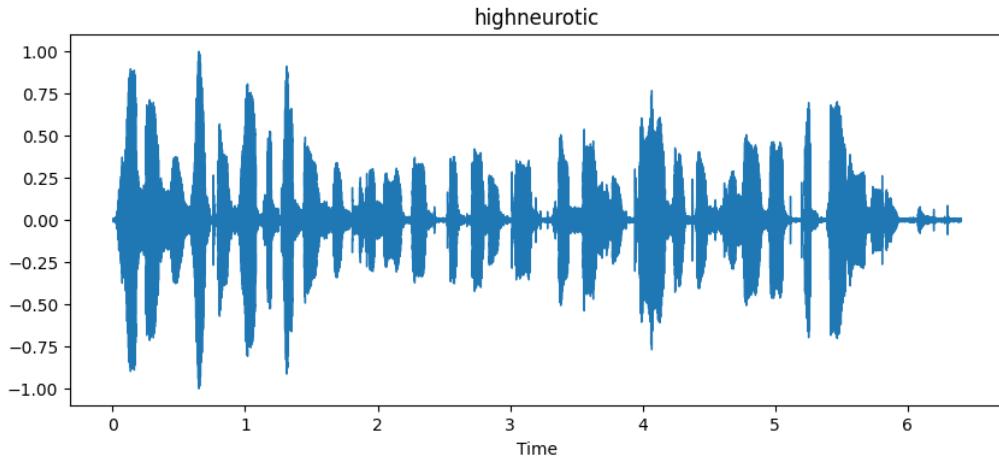


Figure 3.10: Highneurotic Waveplot

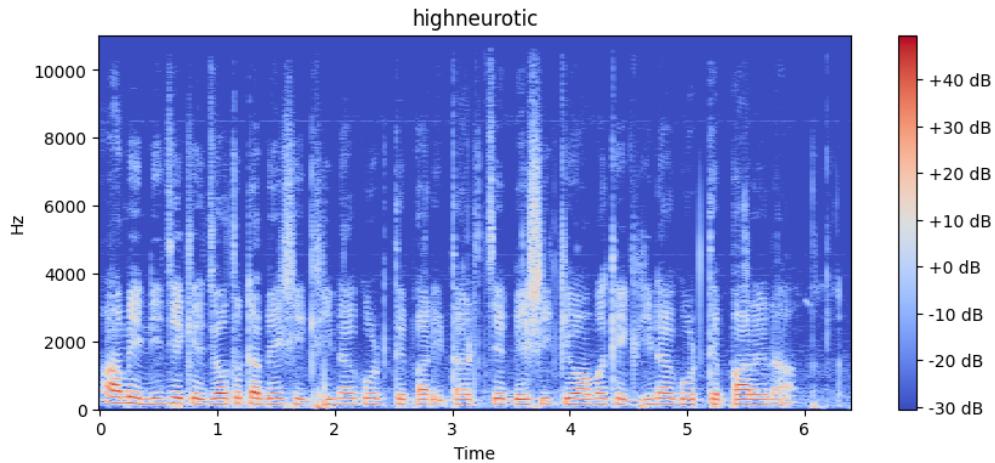


Figure 3.11: Highneurotic Plot of Spectrogram

In 3.11, the x-axis represents time in seconds, and the y-axis represents frequency in Hertz. The colors represent the amplitude of the signal at each frequency and time, with blue being the lowest amplitude and red being the highest amplitude. The graph shows a series of vertical lines, indicating a repeating pattern in the signal.

In 3.12, the x-axis is labeled “MFCC Coefficient” and ranges from 0 to 350. The y-axis is labeled “Frame” and ranges from 0 to 12. The color scale is on the right side of the graph and ranges from -250 to 150.

We provide one highagree audio sample’s waveplot, spectrogram, and mfccplot in Figure 3.13,3.14,4.11 respectively.

In 3.14, the x-axis represents time in seconds, and the y-axis represents frequency in Hertz. The colors represent the amplitude of the signal at each frequency and

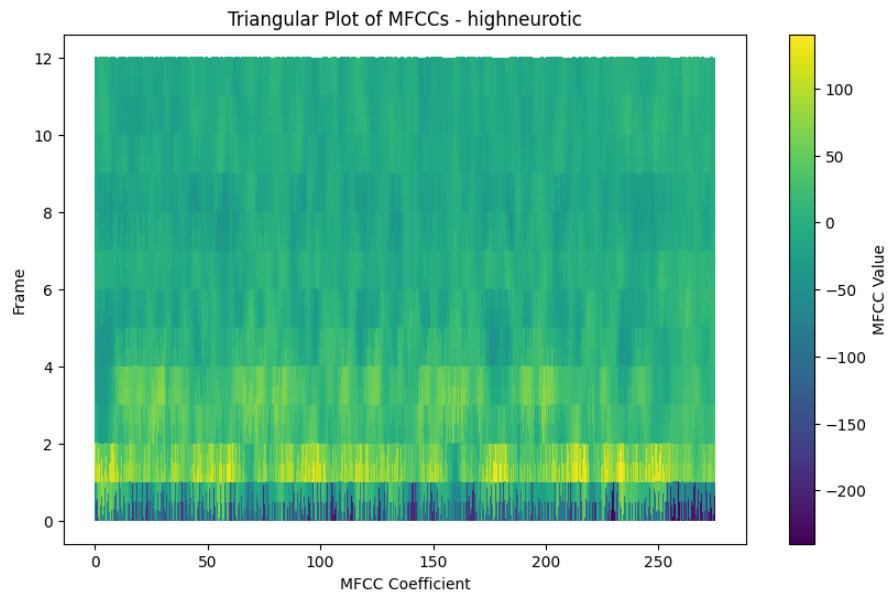


Figure 3.12: Highneurotic Plot of MFCCs

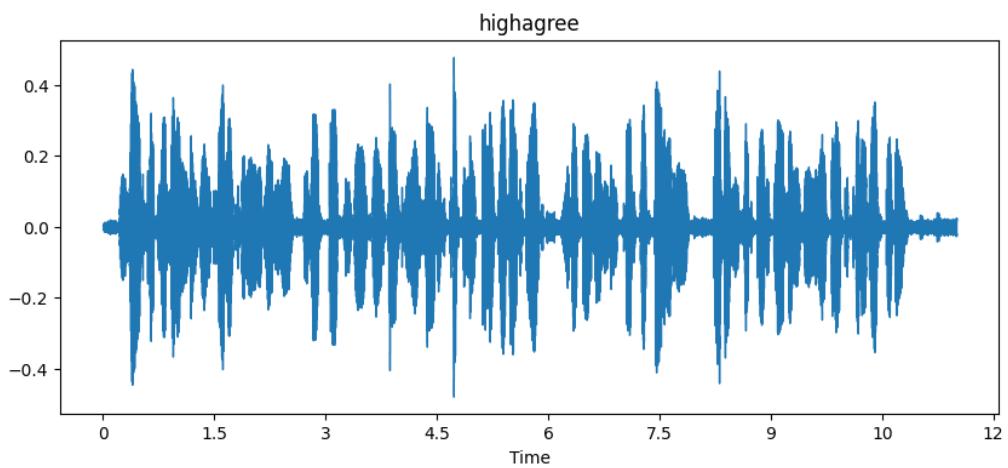


Figure 3.13: Highagree Waveplot

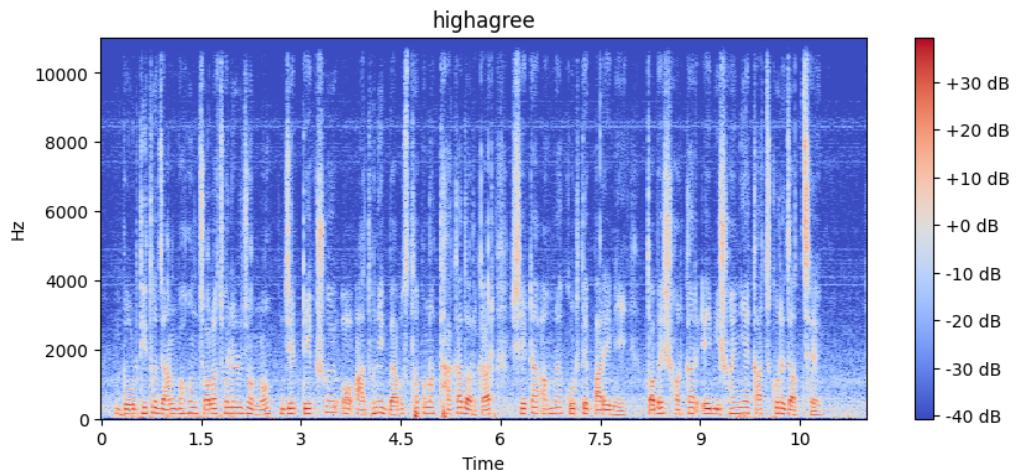


Figure 3.14: Highagree Plot of Spectrogram

time, with blue being the lowest amplitude and red being the highest amplitude. The graph shows a series of vertical lines, indicating a repeating pattern in the signal.

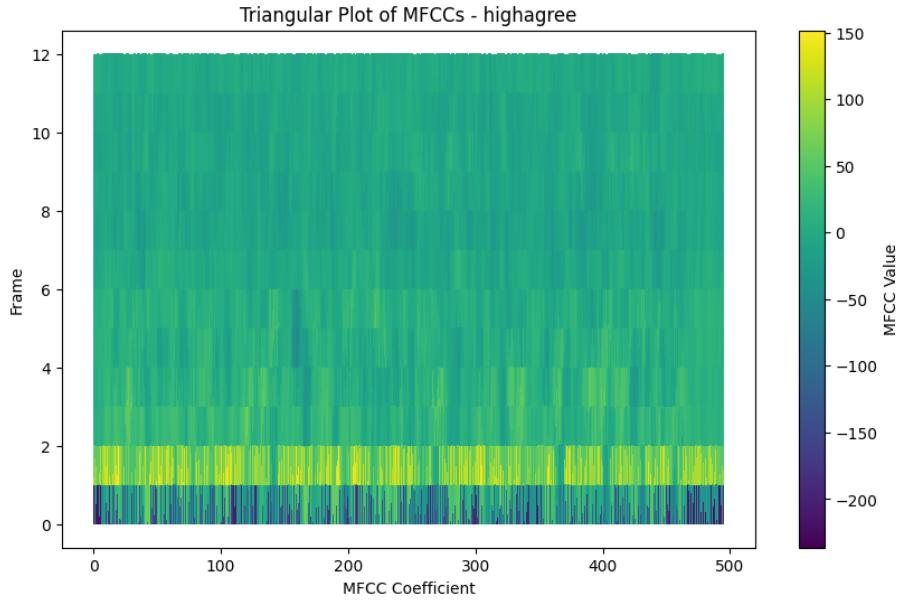


Figure 3.15: Highagree Plot of MFCCs

In 4.11, the x-axis is labeled “MFCC Coefficient” and ranges from 0 to 350. The y-axis is labeled “Frame” and ranges from 0 to 12. The color scale is on the right side of the graph and ranges from -250 to 150.

We provide one lowagree audio sample’s waveplot, spectrogram, and mfccplot in Figure 4.12, 4.13, 3.18 respectively.

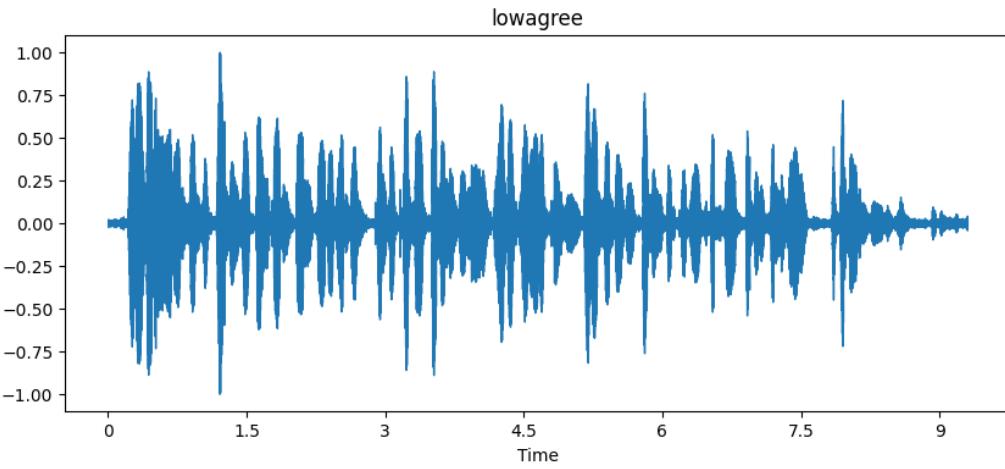


Figure 3.16: Lowagree Waveplot

In 4.13, the x-axis represents time in seconds, and the y-axis represents frequency in Hertz. The colors represent the amplitude of the signal at each frequency and time, with blue being the lowest amplitude and red being the highest amplitude. The graph shows a series of vertical lines, indicating a repeating pattern in the signal.

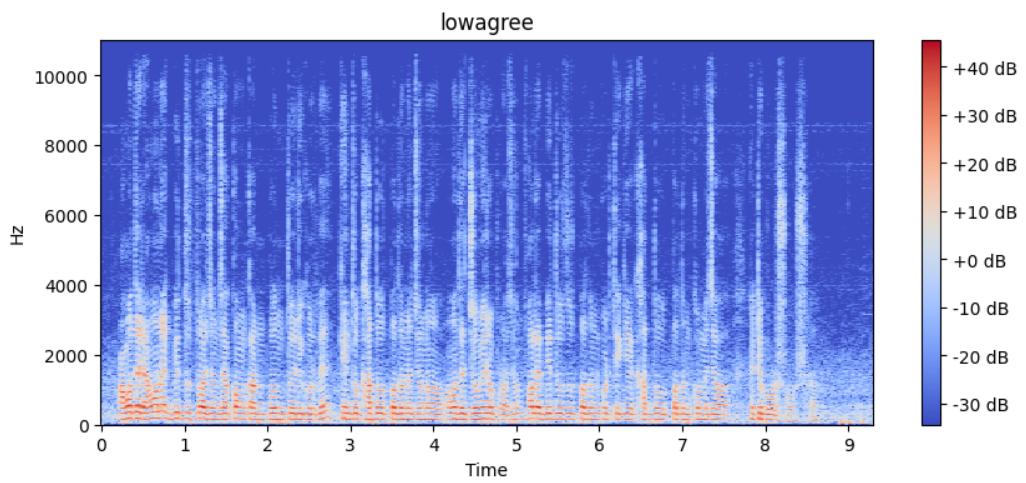


Figure 3.17: Lowagree Plot of Spectrogram

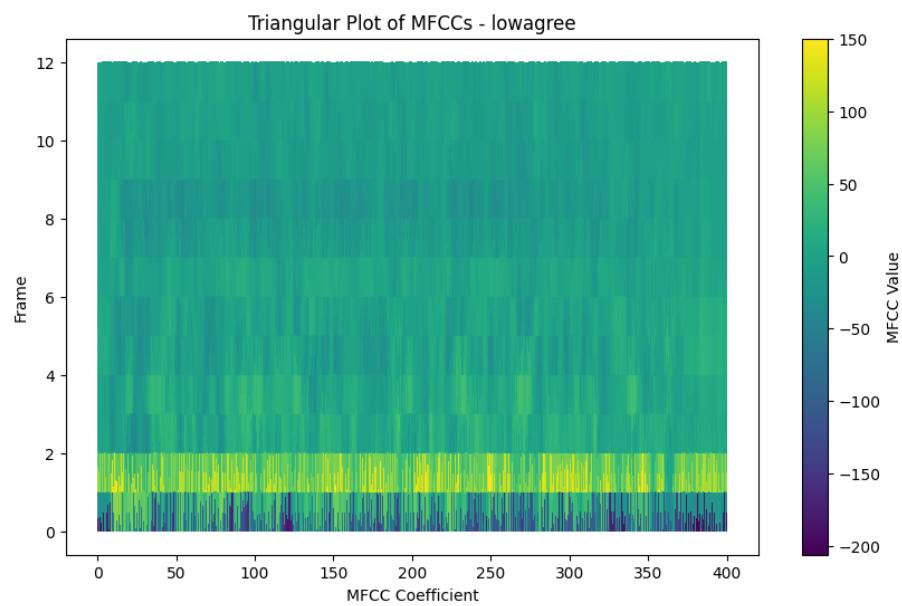


Figure 3.18: Lowagree Plot of MFCCs

In 3.18, the x-axis is labeled “MFCC Coefficient” and ranges from 0 to 350. The y-axis is labeled “Frame” and ranges from 0 to 12. The color scale is on the right side of the graph and ranges from -250 to 150.

We provide one HighConscientious audio sample’s waveplot, spectrogram, and mfcplot in Figure 4.7,3.20,4.3 respectively.

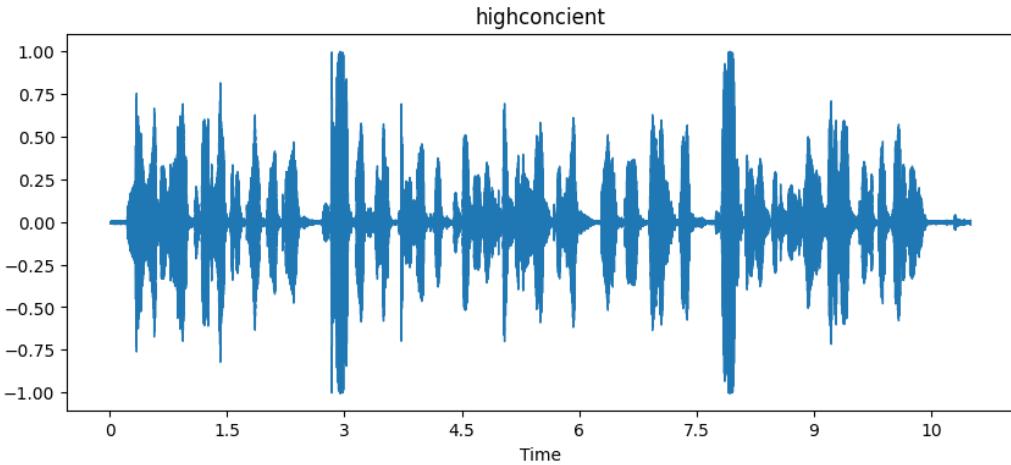


Figure 3.19: HighConscientious Waveplot

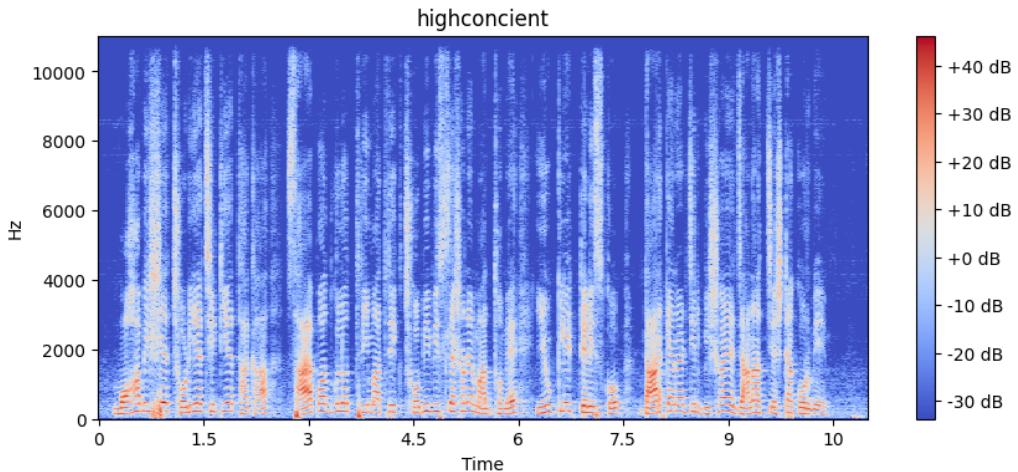


Figure 3.20: HighConscientious Plot of Spectrogram

In 4.13, the x-axis represents time in seconds, and the y-axis represents frequency in Hertz. The colors represent the amplitude of the signal at each frequency and time, with blue being the lowest amplitude and red being the highest amplitude. The graph shows a series of vertical lines, indicating a repeating pattern in the signal.

In 3.18, the x-axis is labeled “MFCC Coefficient” and ranges from 0 to 350. The y-axis is labeled “Frame” and ranges from 0 to 12. The color scale is on the right side of the graph and ranges from -250 to 150.

We provide one HighConscientious audio sample’s waveplot, spectrogram, and mfcplot in Figure 3.22,3.23,3.24 respectively.

In 3.23, the x-axis represents time in seconds, and the y-axis represents frequency in Hertz. The colors represent the amplitude of the signal at each frequency and

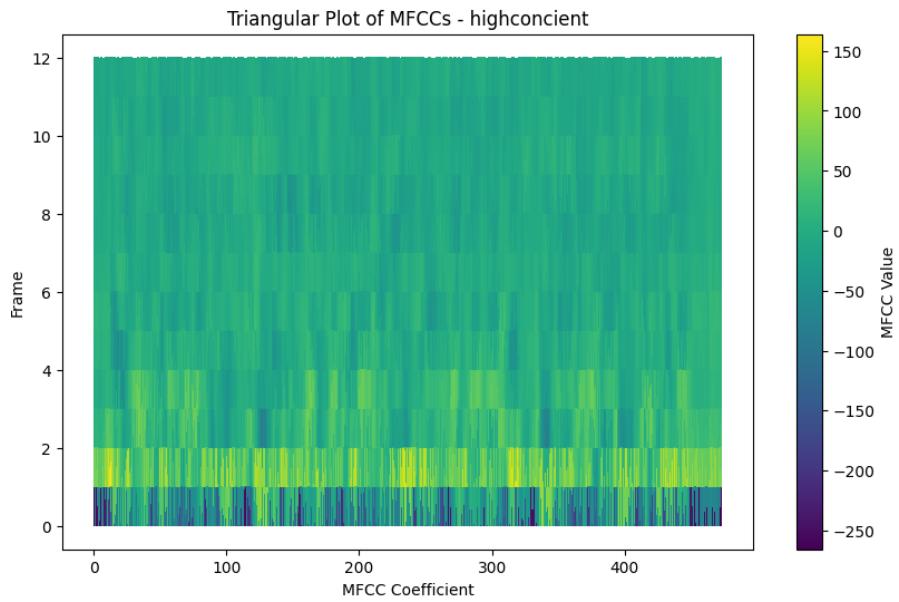


Figure 3.21: HighConscientious Plot of MFCCs

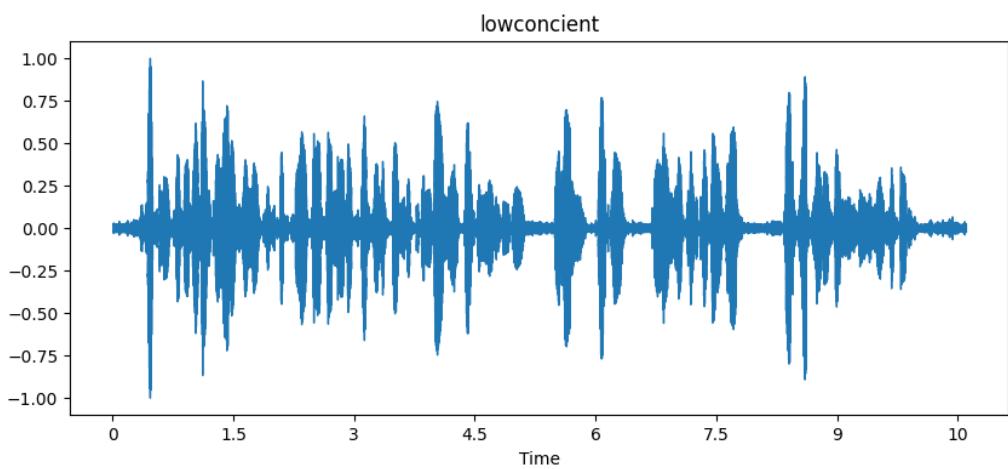


Figure 3.22: LowConscientious Waveplot

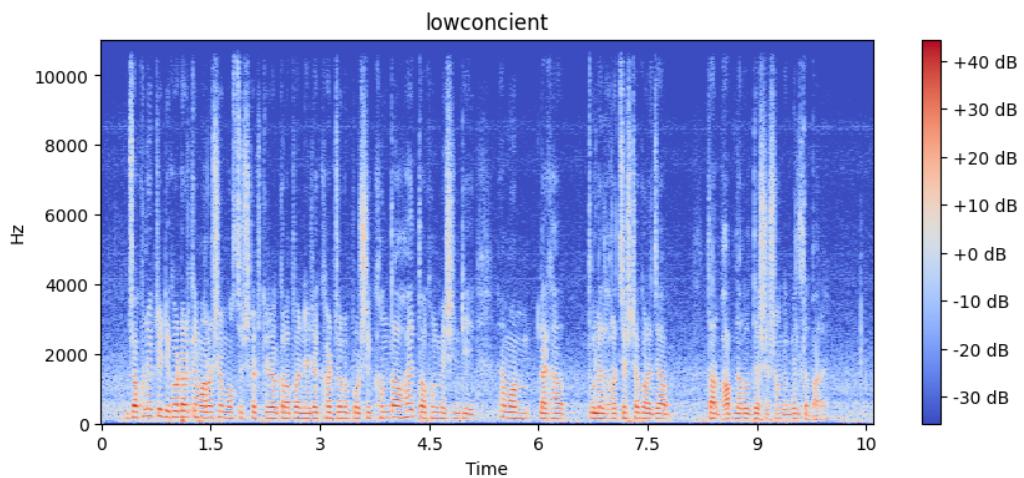


Figure 3.23: LowConscientious Plot of Spectrogram

time, with blue being the lowest amplitude and red being the highest amplitude. The graph shows a series of vertical lines, indicating a repeating pattern in the signal.

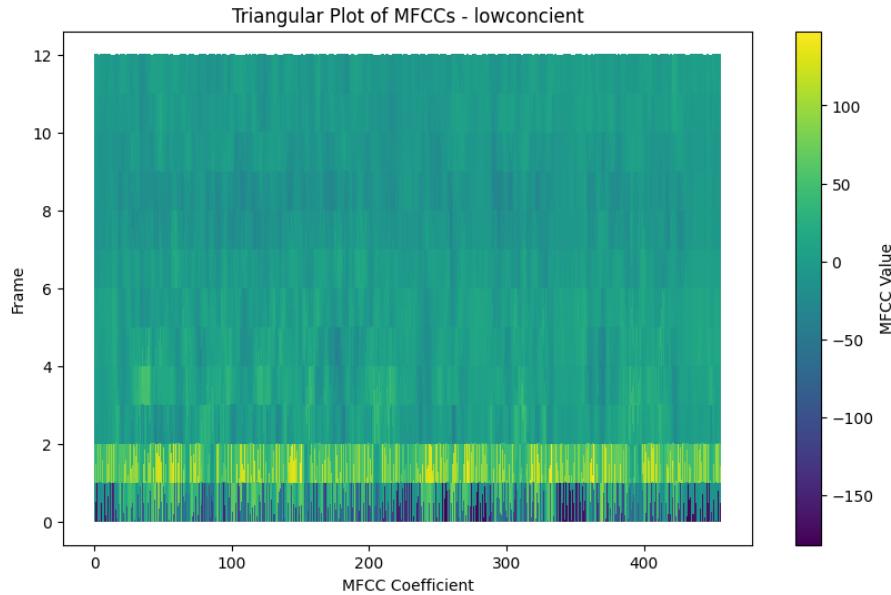


Figure 3.24: LowConscientious Plot of MFCCs

In 3.24, the x-axis is labeled “MFCC Coefficient” and ranges from 0 to 350. The y-axis is labeled “Frame” and ranges from 0 to 12. The color scale is on the right side of the graph and ranges from -250 to 150.

We provide one HighExtrover audio sample’s waveplot, spectrogram, and mfccplot in Figure 3.25,3.26,3.27 respectively.

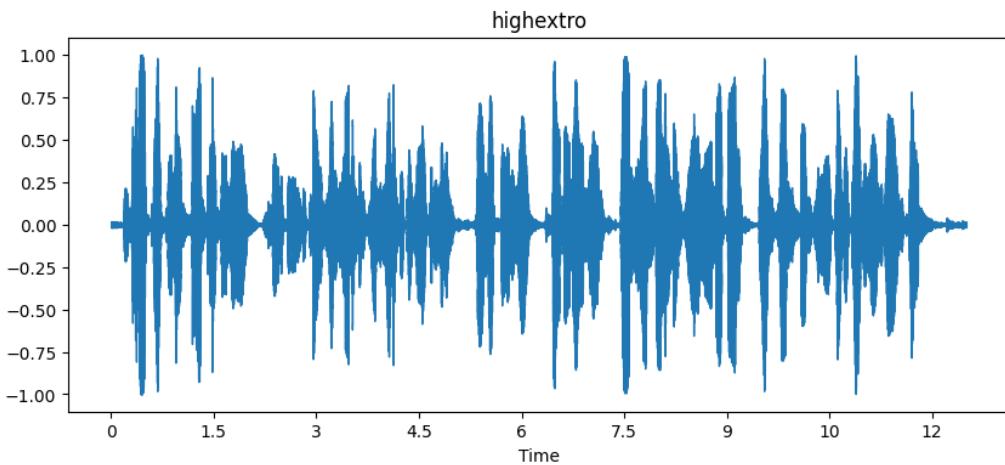


Figure 3.25: HighExtrover Waveplot

In 3.26, the x-axis represents time in seconds, and the y-axis represents frequency in Hertz. The colors represent the amplitude of the signal at each frequency and time, with blue being the lowest amplitude and red being the highest amplitude. The graph shows a series of vertical lines, indicating a repeating pattern in the signal.

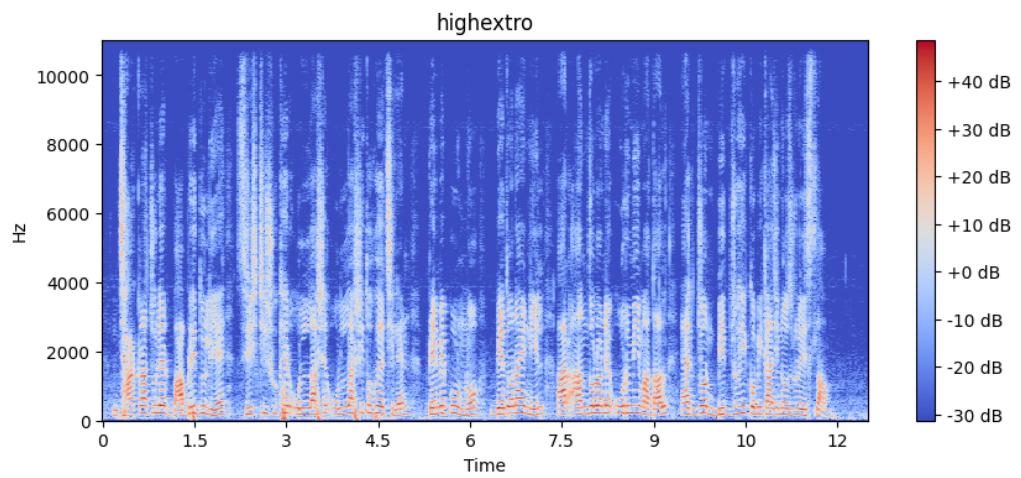


Figure 3.26: HighExtrover Plot of Spectrogram

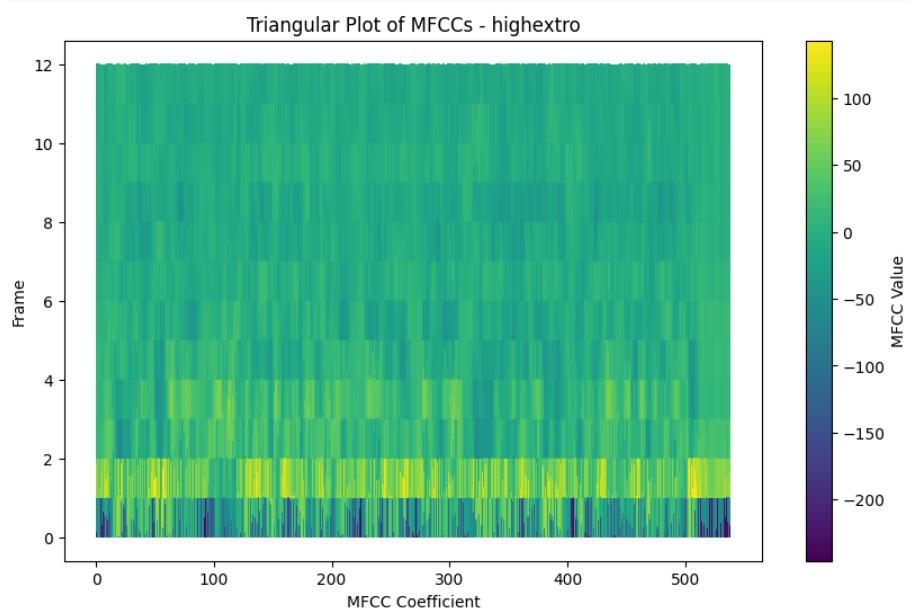


Figure 3.27: HighExtrover Plot of MFCCs

In 3.27, the x-axis is labeled “MFCC Coefficient” and ranges from 0 to 350. The y-axis is labeled “Frame” and ranges from 0 to 12. The color scale is on the right side of the graph and ranges from -250 to 150.

We provide one LowExtrover audio sample’s waveplot, spectrogram, and mfccplot in Figure 3.28,3.29,3.30 respectively.

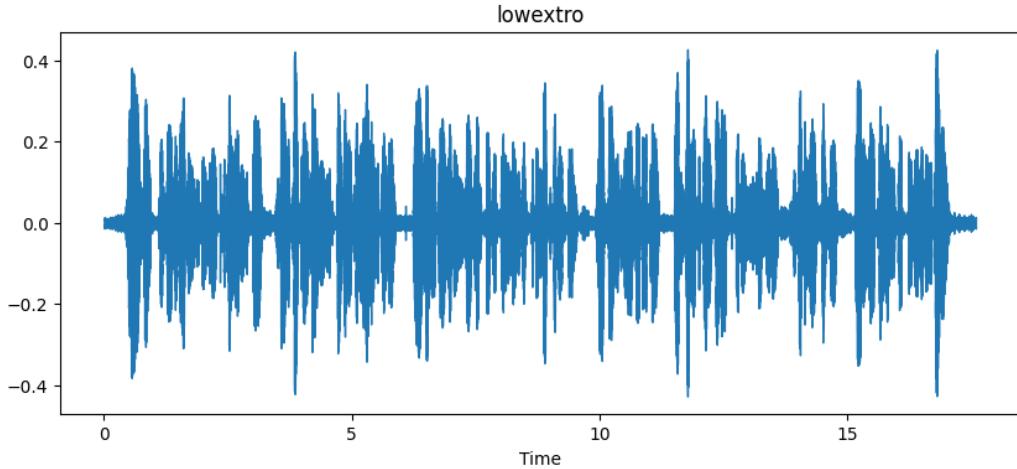


Figure 3.28: LowExtrover Waveplot

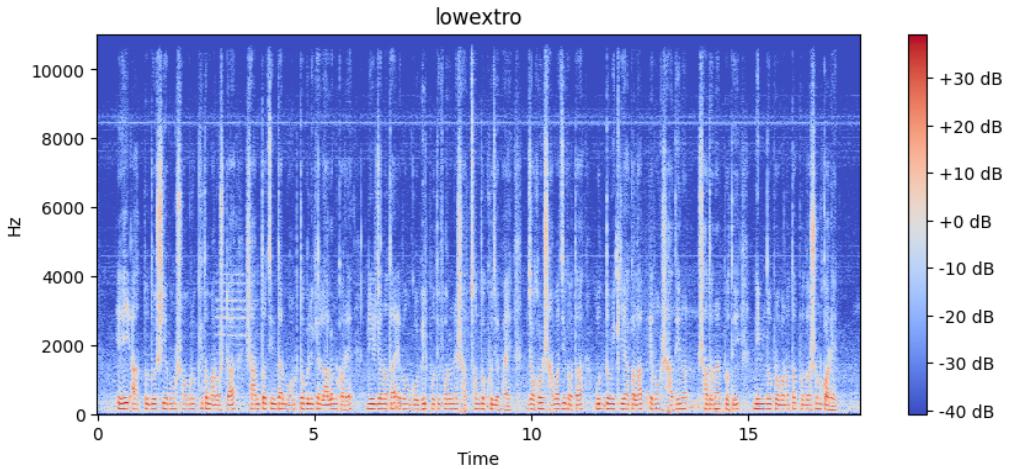


Figure 3.29: LowExtrover Plot of Spectrogram

In 3.29, the x-axis represents time in seconds, and the y-axis represents frequency in Hertz. The colors represent the amplitude of the signal at each frequency and time, with blue being the lowest amplitude and red being the highest amplitude. The graph shows a series of vertical lines, indicating a repeating pattern in the signal.

In 3.30, the x-axis is labeled “MFCC Coefficient” and ranges from 0 to 350. The y-axis is labeled “Frame” and ranges from 0 to 12. The color scale is on the right side of the graph and ranges from -250 to 150.

We provide one HighOpen audio sample’s waveplot, spectrogram, and mfccplot in Figure 3.31,3.32,3.33 respectively.

In 3.32, the x-axis represents time in seconds, and the y-axis represents frequency in Hertz. The colors represent the amplitude of the signal at each frequency and

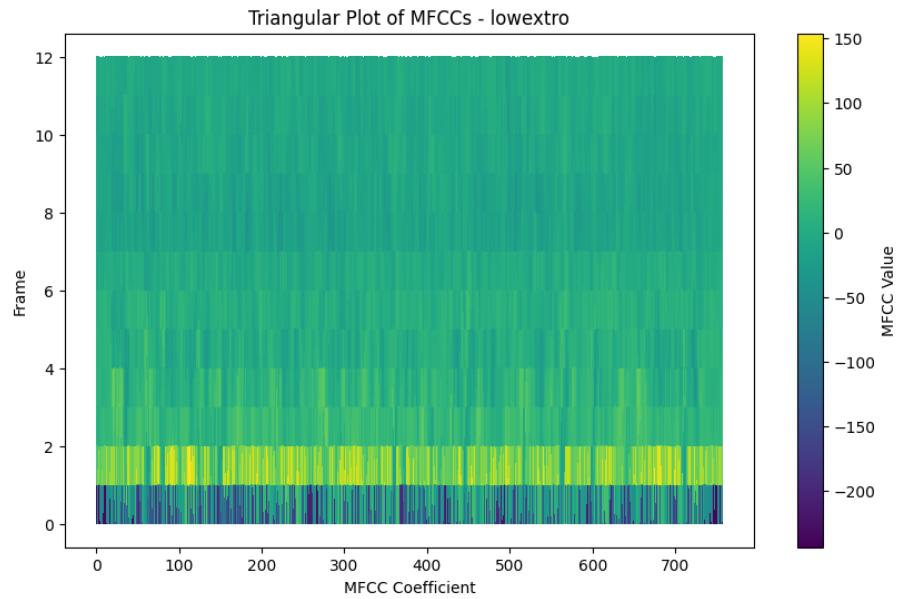


Figure 3.30: LowExtrover Plot of MFCCs

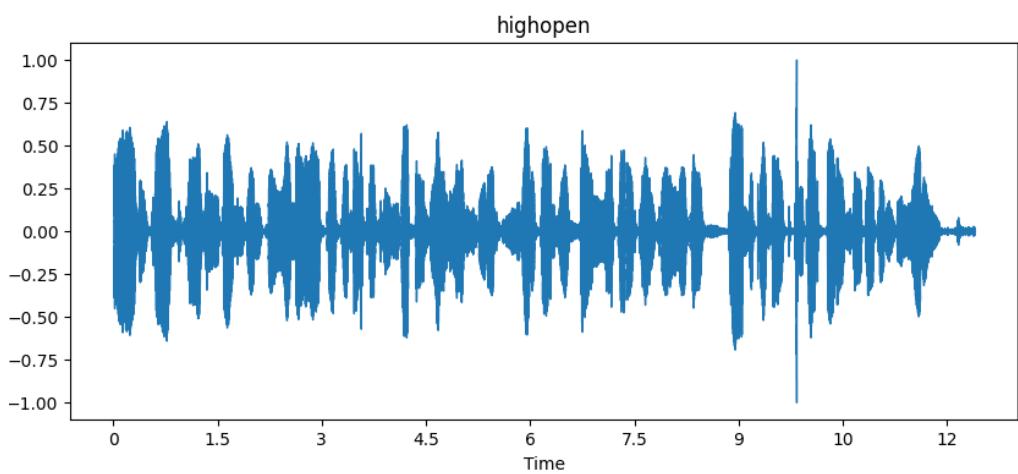


Figure 3.31: HighOpen Waveplot

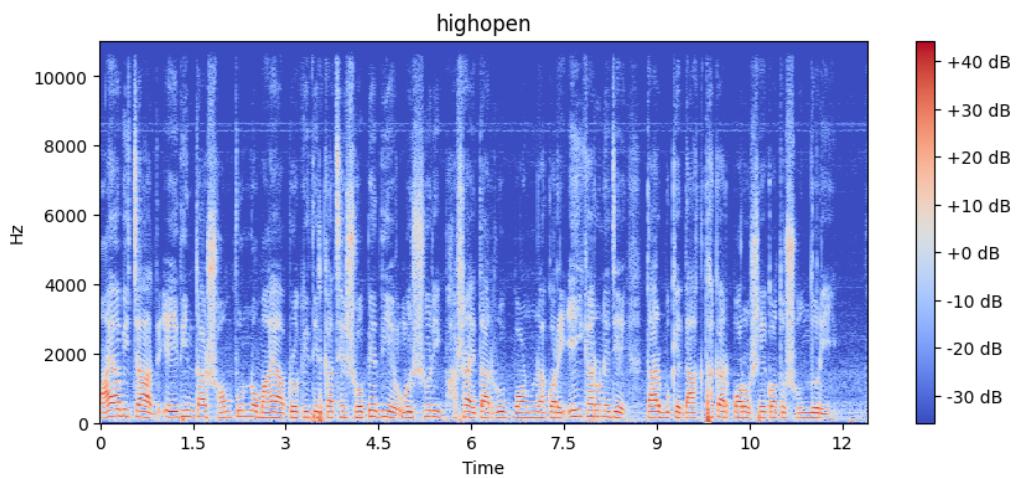


Figure 3.32: HighOpen Plot of Spectrogram

time, with blue being the lowest amplitude and red being the highest amplitude. The graph shows a series of vertical lines, indicating a repeating pattern in the signal.

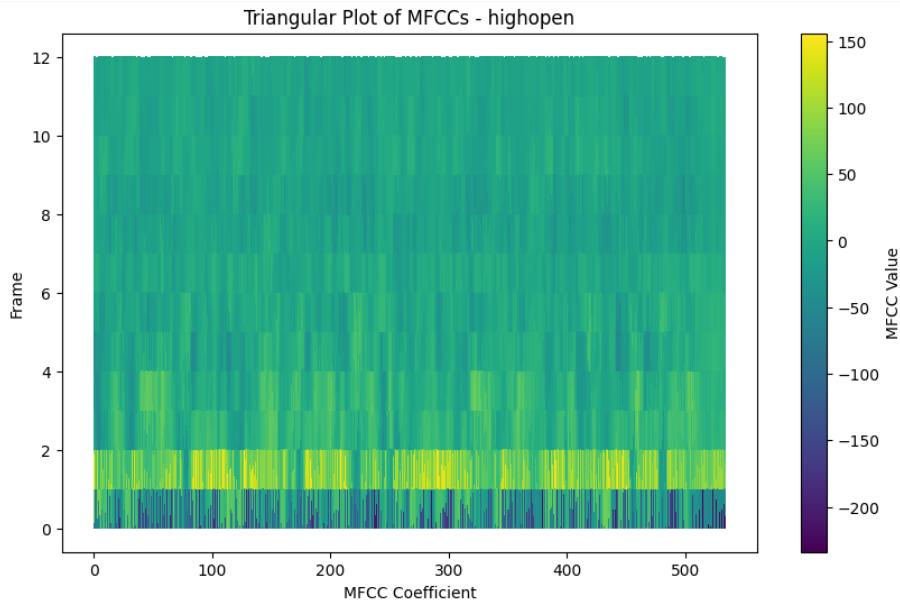


Figure 3.33: HighOpen Plot of MFCCs

In 3.33, the x-axis is labeled “MFCC Coefficient” and ranges from 0 to 350. The y-axis is labeled “Frame” and ranges from 0 to 12. The color scale is on the right side of the graph and ranges from -250 to 150.

We provide one LowOpen audio sample’s waveplot, spectrogram, and mfccplot in Figure 3.34,3.35,3.36 respectively.

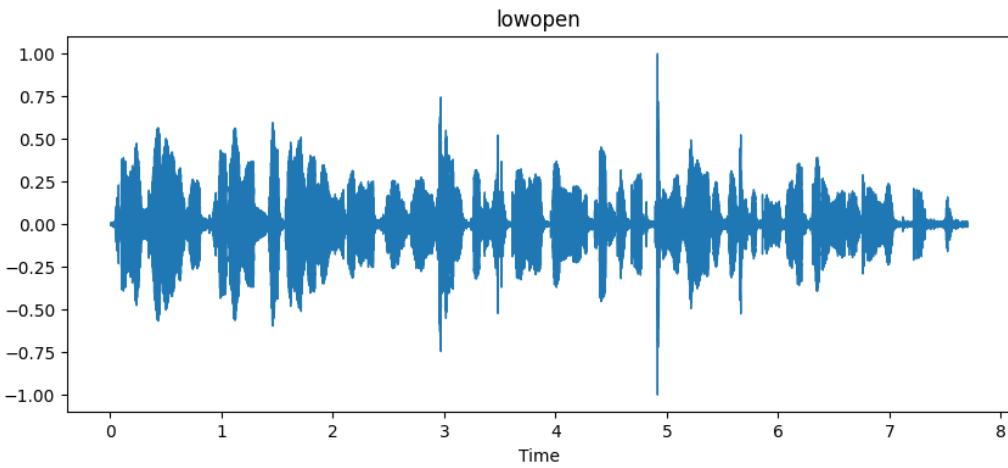


Figure 3.34: LowOpen Waveplot

In 3.35, the x-axis represents time in seconds, and the y-axis represents frequency in Hertz. The colors represent the amplitude of the signal at each frequency and time, with blue being the lowest amplitude and red being the highest amplitude. The graph shows a series of vertical lines, indicating a repeating pattern in the signal.

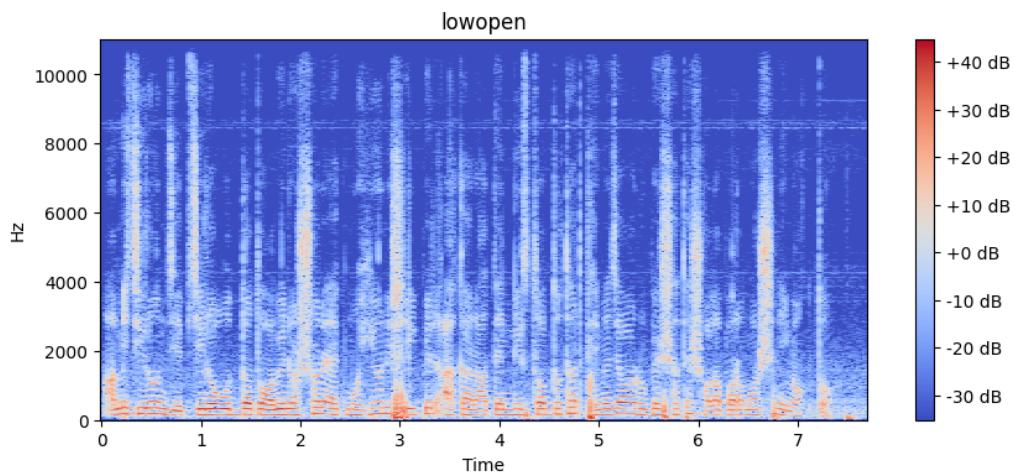


Figure 3.35: LowOpen Plot of Spectrogram

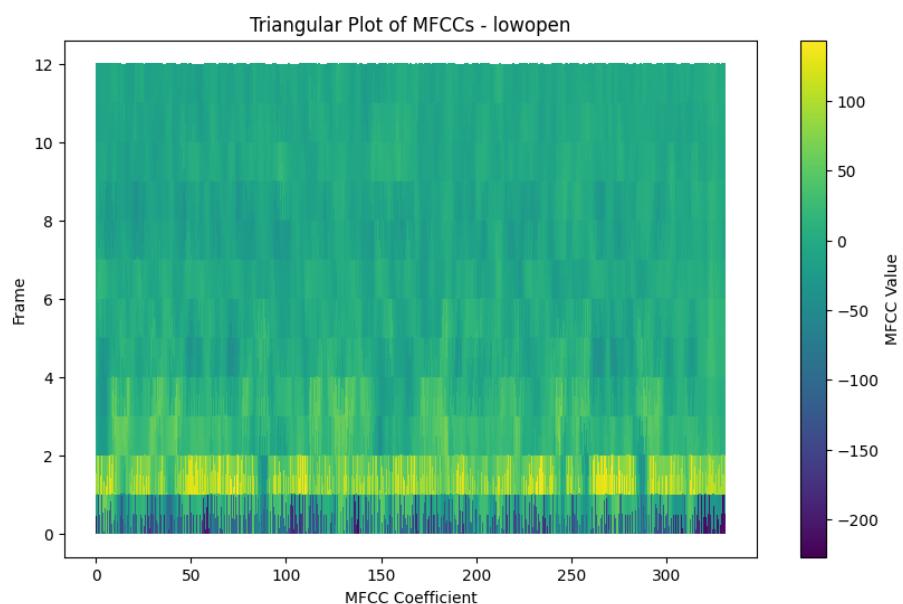


Figure 3.36: LowOpen Plot of MFCCs

In 3.36, the x-axis is labeled “MFCC Coefficient” and ranges from 0 to 350. The y-axis is labeled “Frame” and ranges from 0 to 12. The color scale is on the right side of the graph and ranges from -250 to 150.

3.4.2 Mel-Frequency Cepstral Coefficients with Linear Predictive Coding (MELP)

Besides MFCCs, we develop a feature extraction technique called MELP based on MFCCs and LPC (Linear Predictive Coding) to capture unique acoustic features in speech signals. LPC is try to calculate a set of coefficients that describe the filter and tries to predict the next sound sample based on previous samples [54]. From [32], we can define the LPC formula:

$$b_n = \log\left[\frac{1 - p_n}{1 + p_n}\right] \quad (3.3)$$

Now from equation (3.2) and equation (3.3), we get:

$$MELP = MFCCs \oplus LPC \quad (3.4)$$

$$MELP = \hat{a}_m \oplus b_n \quad (3.5)$$

The concatenation of MFCCs and LPC in the MELP technique offers a comprehensive representation of speech signals. While MFCCs capture spectral features and intensity-related information, LPC delves into the vocal tract’s characteristics. In terms of labels, we used one-hot encoding.

For visual understanding,

We provide one Lowneurotic audio sample’s lpcplot in Figure 3.37.

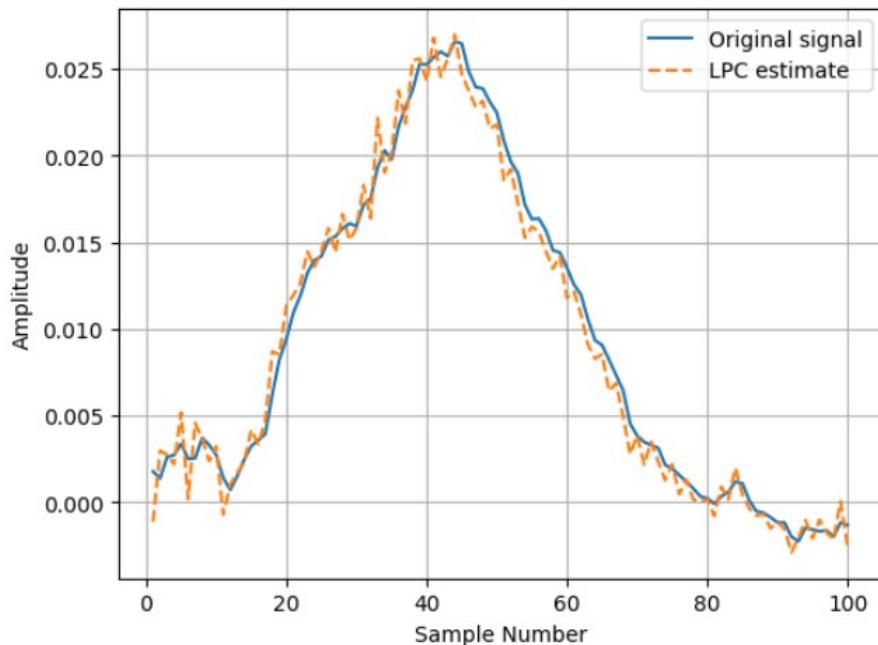


Figure 3.37: Lowneurotic Plot of LPC

In Figure 3.37, x-axis is labeled “Sample Number” and the y-axis is labeled “Amplitude”. The blue line is labeled “Original signal” and the orange line is labeled “LPC estimate”. The blue line has a higher amplitude than the orange line.

We provide one Highneurotic audio sample’s lpcplot in Figure 3.38.

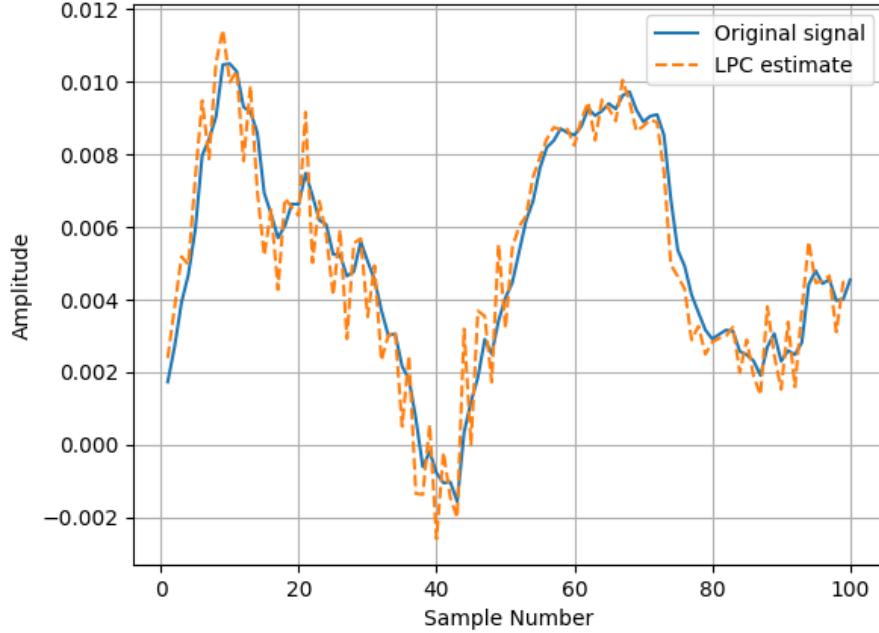


Figure 3.38: Highneurotic Plot of LPC

In Figure 3.38, x-axis is labeled “Sample Number” and the y-axis is labeled “Amplitude”. The blue line is labeled “Original signal” and the orange line is labeled “LPC estimate”. We see, the lines have a similar shape and follow a similar pattern, but the orange line is smoother and less jagged than the blue line.

We provide one Lowagree audio sample’s lpcplot in Figure 3.39.

In Figure 3.39, x-axis is labeled “Sample Number” and the y-axis is labeled “Amplitude”. The blue line is labeled “Original signal” and the orange line is labeled “LPC estimate”. We see, the blue line has a higher amplitude than the orange line and is more jagged. The orange line is smoother and has a lower amplitude than the blue line. The lines intersect at multiple points and follow a similar pattern.

We provide one Highagree audio sample’s lpcplot in Figure 3.40.

In Figure 3.40, x-axis is labeled “Sample Number” and the y-axis is labeled “Amplitude”. The blue line is labeled “Original signal” and the orange line is labeled “LPC estimate”. We see, the x-axis ranges from 0 to 100 and the y-axis ranges from -0.015 to 0.005. The blue line has a sharp dip around sample number 40 and then rises again. The orange line follows the blue line closely but is slightly smoother.

We provide one Lowconscientious audio sample’s lpcplot in Figure 3.41.

In Figure 3.41, x-axis is labeled “Sample Number” and the y-axis is labeled “Amplitude”. The blue line is labeled “Original signal” and the orange line is labeled “LPC estimate”. We see, the blue line has a jagged pattern, while the orange line

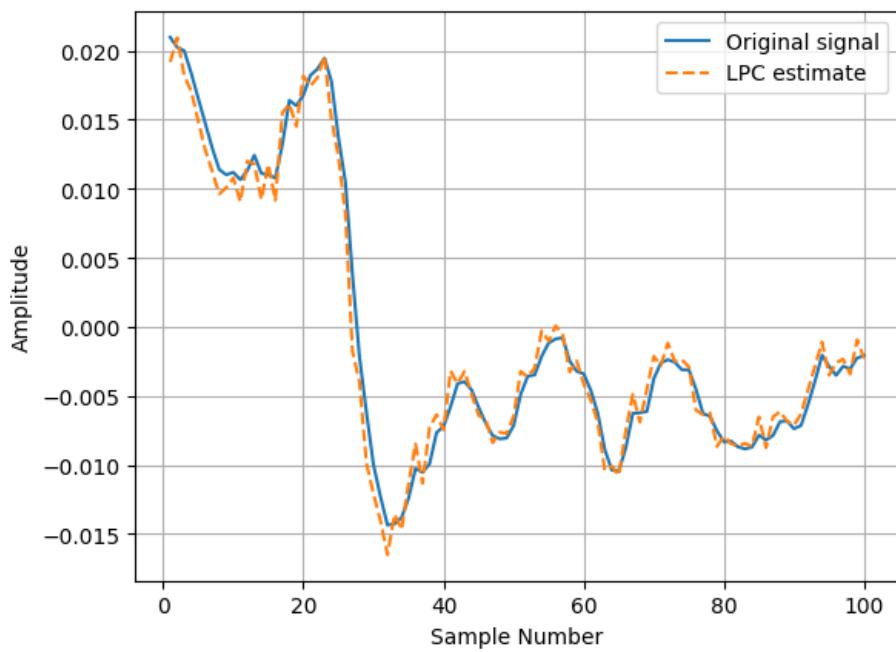


Figure 3.39: Lowagree Plot of LPC

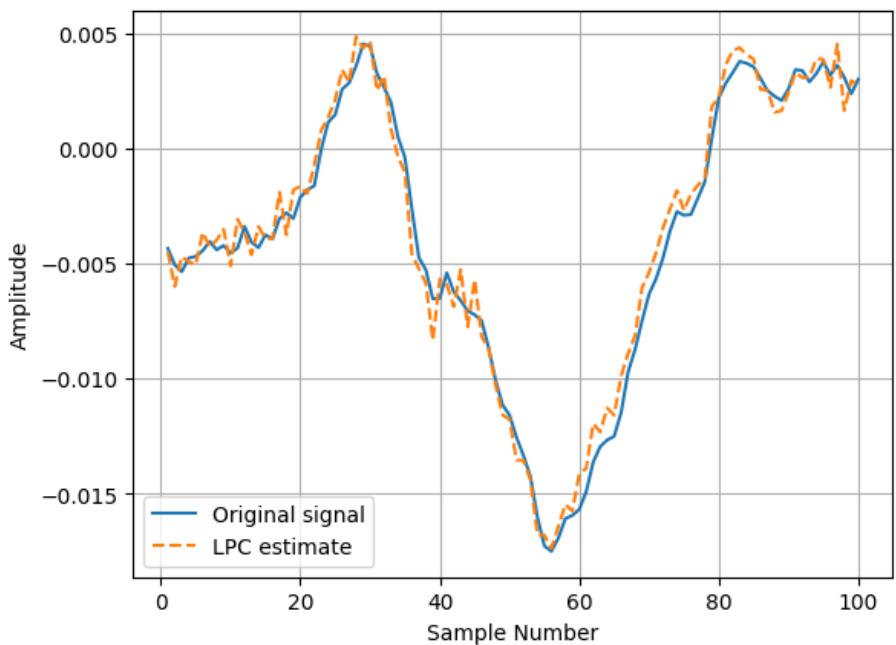


Figure 3.40: Highagree Plot of LPC

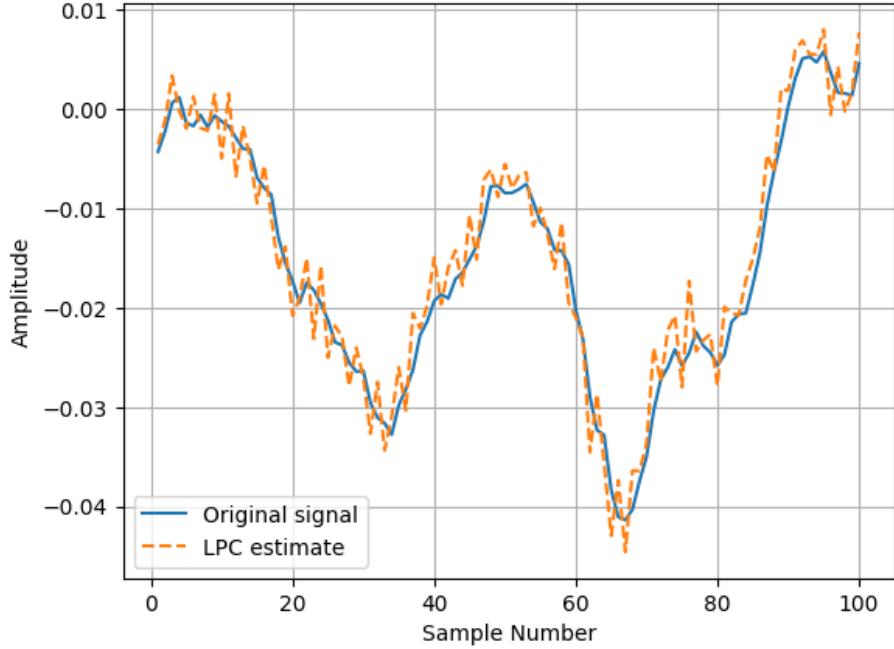


Figure 3.41: Lowconscious Plot of LPC

has a smoother pattern. The lines start at the same point on the left side of the graph and end at the same point on the right side of the graph. The lines intersect at multiple points throughout the graph.

We provide one Highconscious audio sample's lpcplot in Figure 3.42.

In Figure 3.42, x-axis is labeled "Sample Number" and the y-axis is labeled "Amplitude". The blue line is labeled "Original signal" and the orange line is labeled "LPC estimate". We see, the lines are jagged and do not follow a clear pattern. The lines intersect at multiple points.

We provide one Lowextrover audio sample's lpcplot in Figure 3.43.

In Figure 3.43, x-axis is labeled "Sample Number" and the y-axis is labeled "Amplitude". The blue line is labeled "Original signal" and the orange line is labeled "LPC estimate". We see, the blue line has a higher amplitude than the orange line and is more jagged. The orange line is smoother and has a lower amplitude than the blue line. The x-axis ranges from 0 to 100 and the y-axis ranges from -0.015 to 0.025.

We provide one Highextrover audio sample's lpcplot in Figure 3.44.

In Figure 3.44, x-axis is labeled "Sample Number" and the y-axis is labeled "Amplitude". The blue line is labeled "Original signal" and the orange line is labeled "LPC estimate". We see, the x-axis ranges from 0 to 100 and the y-axis ranges from -0.006 to 0.004.

We provide one Lowopen audio sample's lpcplot in Figure 3.45.

In Figure 3.45, x-axis is labeled "Sample Number" and the y-axis is labeled "Amplitude". The blue line is labeled "Original signal" and the orange line is labeled "LPC estimate". We see, the blue line has a more jagged appearance, while the orange line is smoother. The x-axis ranges from 0 to 100 and the y-axis ranges from

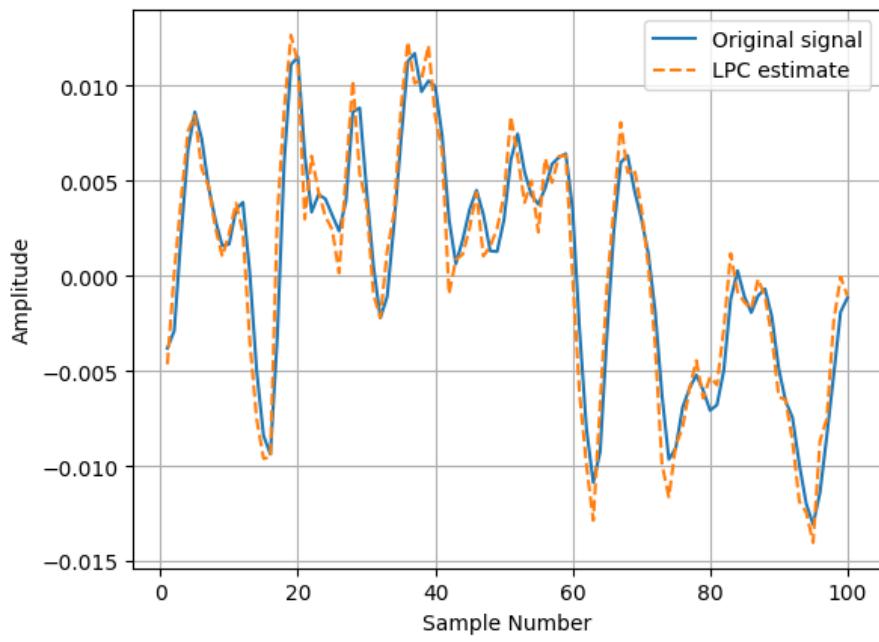


Figure 3.42: Highconscious Plot of LPC

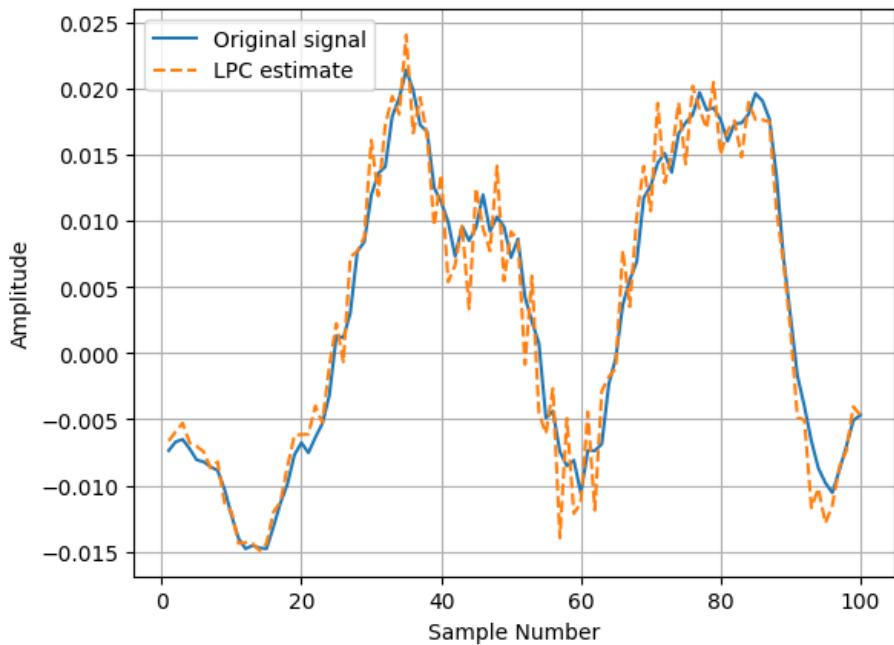


Figure 3.43: Lowextrover Plot of LPC

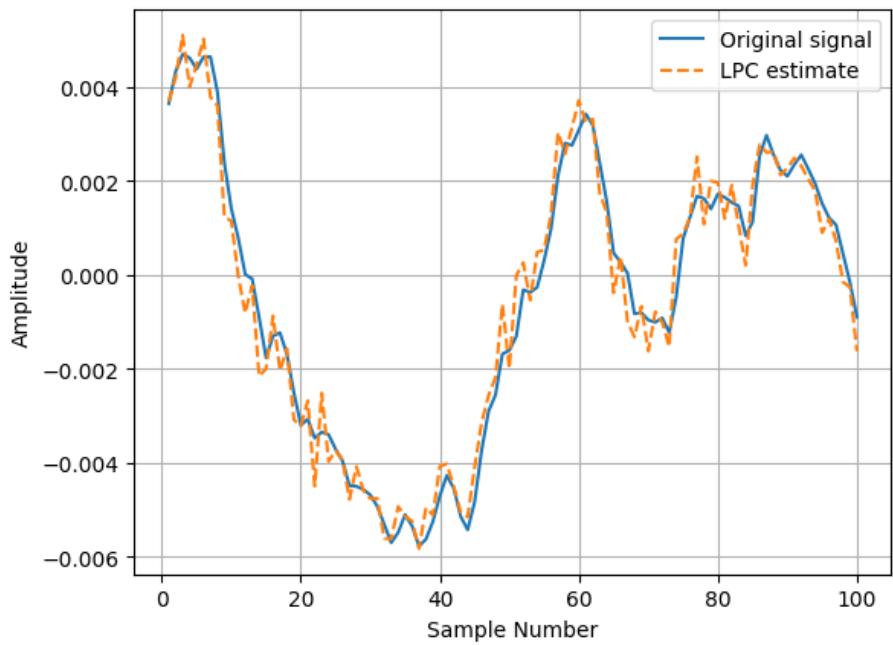


Figure 3.44: Highextrover Plot of LPC

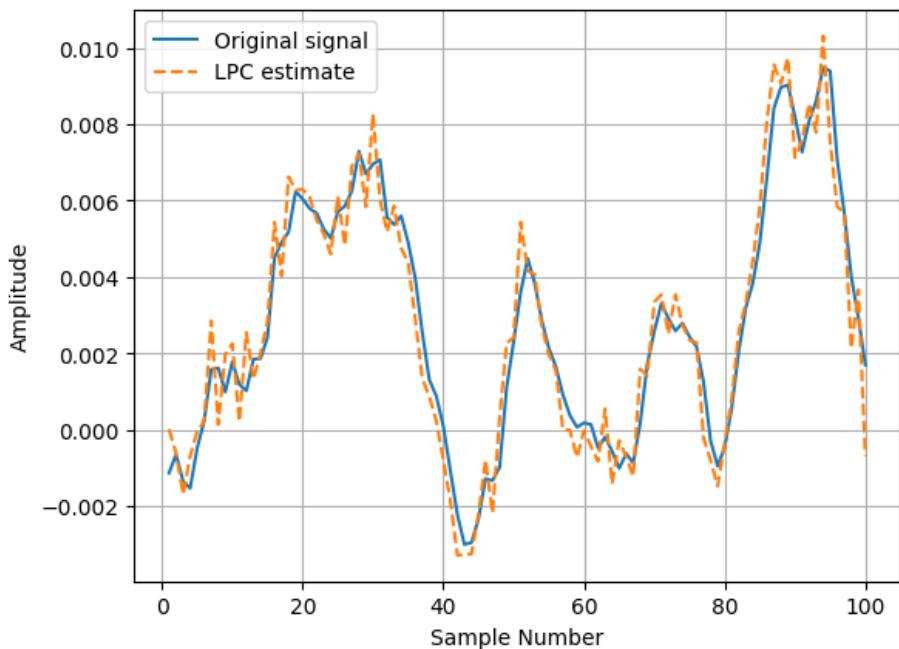


Figure 3.45: Lowopen Plot of LPC

-0.002 to 0.010.

We provide one Highopen audio sample's lpcplot in Figure 3.46.

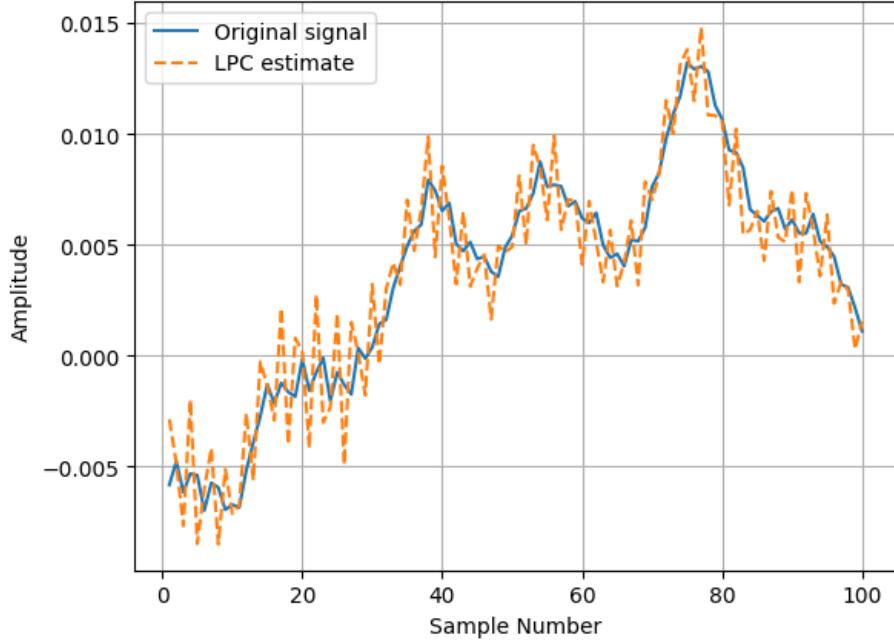


Figure 3.46: Highopen Plot of LPC

In Figure 3.46, x-axis is labeled “Sample Number” and the y-axis is labeled “Amplitude”. The blue line is labeled “Original signal” and the orange line is labeled “LPC estimate”. We see, the blue line has a higher amplitude than the orange line and is more jagged. The orange line is smoother and has a lower amplitude.

3.4.3 Mel-Frequency Cepstral Coefficients with Wiener Linear Predictive Coding (MEWLP)

The Wiener filter [10] is an adaptive, frequency domain linear filter designed to minimize mean square error, widely used for noise reduction in signals. For MEWLP feature extraction, we use Wiener filter to denoise an audio signal and then compute the LPC coefficients of the denoised signal. Furthermore, we concatenate WLPC coefficients of the denoised signal with MFCCs to obtain feature vector. It is defined as:

original audio signal = $y[n]$

$$LPC(y, order) = a_k; k = 1, 2, \dots, order = 40 \quad (3.6)$$

In eqaution (3.6), a_k is LPC coefficients.

$$ns[n] = y[n] + (nl * rn[n]) \quad (3.7)$$

In equation (3.7), we generate a noisy version of the original signal by adding some random noise. ns , nl , and rn represent *noisy_signal*, *noise_level*, and *random_noise* respectively.

$$W_k = \frac{|a_k|}{|a_k| + var(y - ns)} \quad (3.8)$$

In equation (3.8), the Wiener filter coefficients are computed where W_k is the Wiener filter coefficient for the $K - th$ LPC coefficient.

$$denoised_signal[n] = y[n] * W_n; n = 0, 1, \dots, len(y) - 1 \quad (3.9)$$

In equation (3.9), apply the Wiener filter coefficients to each sample of the original signal.

$$WLPC(denoised_signal, order) = a'_k \quad (3.10)$$

In equation (3.10), the WLPC coefficients are computed of the denoised signal where a'_k are the WLPC coefficients after denoising.

$$MFCC(y, n_mfcc) = \frac{1}{T} \sum_{t=1}^T MFCC_t \quad (3.11)$$

In equation (3.11), the Mel-Frequency Cepstral Coefficients (MFCC) are extracted where $MFCC_t$ is the vector of MFCC coefficients at time t and T is the number of frames.

$$cf = [MFCC_1, \dots, MFCC_{n_mfcc}, a'_1, \dots, a'_{order}] \quad (3.12)$$

In equation (3.12), the MFCC and WLPC coefficients concatenate to obtain the combined feature vector where cf represent *combined_features*.

3.4.4 Morlet-based Mel-frequency Cepstral Coefficients (MoMF)

For MoMF feature extraction, we use Morlet low pass filter that effectively suppresses the higher-frequency components of the Morlet wavelet [3], resulting in a filter that retains the low-frequency information in the analysis of time-frequency representations of signals. Since all audio samples in our dataset feature male voices, this approach effectively isolates the distinctive lower-frequency components inherent in male vocalizations that enhancing the analysis of key characteristics such as pitch and fundamental frequency. From [37], we can define the Morlet wavelet formula:

$$Morlet(t) = \cos(2\pi f_c t) \cdot e^{\frac{-t^2}{2B^2}} \quad (3.13)$$

In equation (3.13), f_c , B , t represent *center_frequency*, *bandwidth*, and *time* respectively. As we want to capture low-frequency information so we need convolution operation. The convolution operation:

$$fs[n] = \sum_{k=-\infty}^{\infty} signal[k].Morlet(n - k) \quad (3.14)$$

In equation (3.14), fs , k , n represent *filtered_signal*, *input_signal*, and *kernel* respectively. The convolution operation is used here to implement Morlet low-pass filter.

$$x[j] = |FFT(fs[n])| \quad (3.15)$$

In equation (3.15), we use Fast Fourier Transform (FFT) to compute the magnitude spectrum ($x[j]$) of the filtered signal.

$$ms[i] = \sum_j mf[i, j] \cdot x[j] \quad (3.16)$$

In equation (3.16), Mel filterbank is applied using matrix multiplication where $i = 0$ to `num_filters-1`, $j = 0$ to `len(magnitude_spectrum)`, ms and mf represent `mel_spectrum` and `mel_filters` respectively. `mel_spectrum` captures the energy distribution across different frequency bands, emphasizing regions that are more perceptually significant. Then the log compression is applied to mimic the human auditory system's sensitivity to differences in loudness:

$$lce[i] = \log(\epsilon + ms[i]) \quad (3.17)$$

In equation (3.17), applies a logarithmic compression to the Mel spectrum where lce represent `log_compressed_energies` and ϵ is a small constant to avoid taking logarithm zero.

$$fv[i] = \sum_{i=0}^{N-1} ice[i] \cdot \cos\left(\frac{\pi}{N}j(i + \frac{1}{2})\right) \quad (3.18)$$

In equation (3.18), DCT (Discrete Cosine Transform) is applied to the `log_compressed_energies` to obtain the feature vector where fv represent `feature_vector`, N represent the number of coefficients.

From equation (3.18) and equation (3.2), we get

$$MoMF = fv \oplus MFCCs \quad (3.19)$$

MFCC is a column vector of size $(M * 1)$ and Morlet is a column vecor of size $(N * 1)$. After concatenation, we get single vector shape $((M + N) * 1)$ for each signal.

3.5 DistilRo and BiG: Soft Voting Ensemble Models for Personality Classification in Speech, and Speech-to-Text Modalities

For the classification of Speech, and Speech-to-Text modalities, we use different techniques. Voting classifier is a technique in ensemble learning. Ensemble learning models leverage the decisions made by various baseline models to enhance overall performance. In the case of a soft voting ensemble model, it predicts the class label by considering the highest sum of predicted probabilities from the baseline models. For multi-class classification in the speech-to-text modality, we introduce DistilRo. DistilRo is a soft voting classifier that combines the strengths of DistilBERT [50] and RoBERTa [41] baseline models. They work together in a soft voting setup to figure out the personality traits and considering all semantic stuff in the text. In

the speech multi-class classification, we present BiG. BiG is another soft voting classifier, this time utilizing Bi-LSTM and GRU baseline models. These components work together to make accurate predictions while also considering the semantic subtle present in the spoken language. Now, we explain each model and then we explain parameter tuning that used to develop them.

3.5.1 Distilled Bidirectional Encoder Representations from Transformers (DistilBERT) and Robustly optimized BERT approach (RoBERTa)

DistilBERT [50] is a shorter version of BERT model. It uses the same architecture of BERT [38], which is a transformer model. Based on the model [50], multi-head self-attention of transformer allows to focus on different parts of the input sentence and learn the relationship between them. It is defined as:

$$g(Q, K, V) = c(h_1, \dots, h_t)w^0 \quad (3.20)$$

$$\text{where, } h_l = a(Qw_l^Q, Kw_l^K, Vw_l^V), a(Q, K, V) = s\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (3.21)$$

In equation (3.20) and (3.21) multihead, concat, attention, and softmax are considering as g, c, a, and s respectively.

Feed-forward neural network allows the model to apply non-linear transformations to the input and learn complex features. It is defined as:

$$f(a) = r(aw_1 + d_1)w_2 + d_2 \quad (3.22)$$

In equation (3.22), ReLU function is considering as r.

Model takes input as a sequence of tokens, which are words or subwords, and converts them into vector using word embeddings and position embeddings. Word embeddings capture the meaning of each token, and position embeddings capture the order of each token in the sequence [47]. Then the model sum word embeddings, and position embeddings for producing a final hidden state for each token [50]. The equation for the output of DistilBERT is:

$$\text{hidden}^l = \text{transformer}(e + p) \quad (3.23)$$

$$\text{transformer}(a) = f(g(a, a, a)) \quad (3.24)$$

In equation(3.23), e is the embedding matrix, p is the position embedding matrix, and transformer is the Transformer encoder with l layers. Final hidden state can be used for different tasks, in our case it's a classification task.

RoBERTa [41] uses the same architecture as BERT [38], which is based on the Transformer model with more layers, and parameters that make it larger, and powerful. The working procedure of DistilBERT and RoBERTa are same. In Figure 3.47, we show an overview of DistilBERT model where we use one openness text from our dataset.

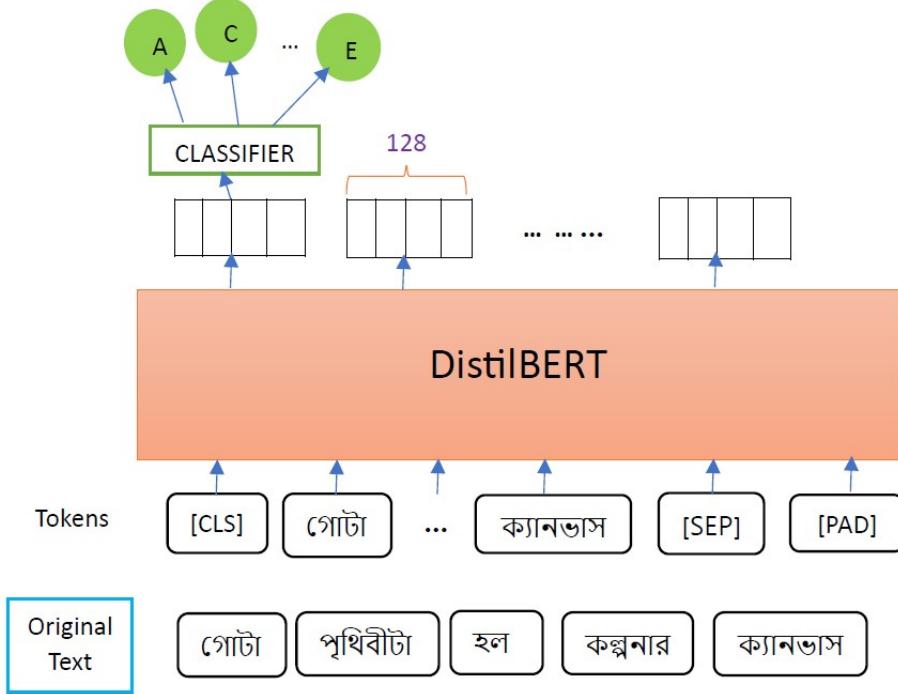


Figure 3.47: An overview of DistilBERT work flow

3.5.2 DistilRo

Our DistilRo model is a sophisticated soft voting classifier, bringing together the two powerful baseline models: DistilBERT [50] and RoBERTa [41]. We have fine-tuned each baseline model to optimize the performance of the soft voting classifier. In Table 3.5, we provide some specific parameters that have been employed in these baseline models.

Table 3.5: Baseline Models parameters of DistilRo

Baseline Models	Parameters
DistilBERT	train_batch_size = 16 eval_batch_size = 64 epochs = 11
RoBERTa	train_batch_size = 8 eval_batch_size = 32 gradient_accumulation = 4 epochs = 20

We've provided a visual representation of its structure in Figure 3.48. This unique combination of DistilBERT and RoBERTa, working together as a soft voting classifier, is designed to capture and utilize semantic information, ensuring accurate and reliable results in personality classification.

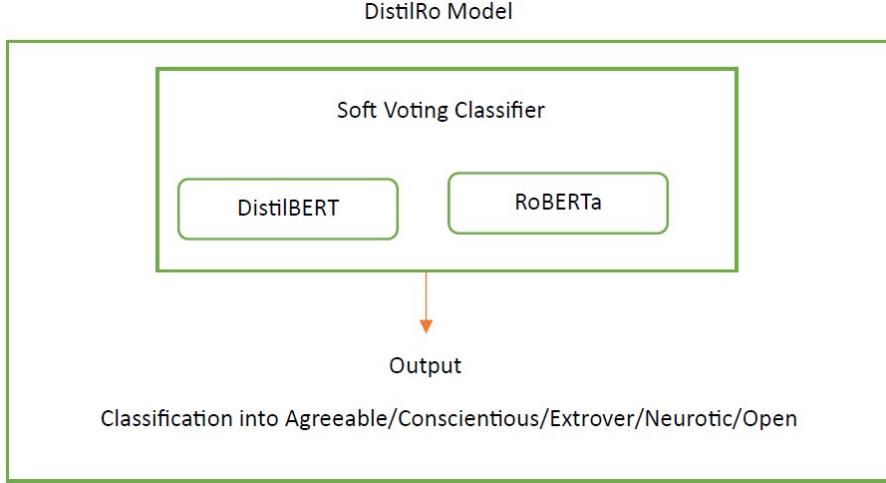


Figure 3.48: Structure of DistilRo Model

3.5.3 Bidirectional Long Short-Term Memory (Bi-LSTM)

Bi-LSTM [61] is one kind of LSTM [18] that can learn long-term dependencies from sequential data in both forward and backward directions. It combines both LSTM and bidirectional processing for sequence learning. Based on [18], for forward direction, it can be defined as:

$$p_k = \sigma(w_{ap}a_k + w_{lp}l_{k-1} + d_p) \quad (3.25)$$

$$q_k = \sigma(w_{aq}a_k + w_{lq}l_{k-1} + d_q) \quad (3.26)$$

$$r_k = \sigma(w_{ar}a_k + w_{lr}l_{k-1} + d_r) \quad (3.27)$$

$$\tilde{s}_k = \tanh(w_{as}a_k + w_{ls}l_{k-1} + d_s) \quad (3.28)$$

$$s_k = q_k \odot s_{k-1} + p_k \odot \tilde{s}_k \quad (3.29)$$

$$l_k = r_k \odot \tanh(s_k) \quad (3.30)$$

where σ is sigmoid function, \tanh is tan hyperbolic function, \odot is point-wise multiplication, a_k is input vector at time step k . p_k , q_k , r_k , and \tilde{s}_k are the input, forget, output, and cell gate respectively. s_k is cell state, and l_k is hidden state at time step k . w and d are the weight matrices and bias for each gate.

For backward direction:

$$\overleftarrow{p}_k = \sigma(\overleftarrow{w}_{ap}\overleftarrow{a}_k + \overleftarrow{w}_{lp}\overleftarrow{l}_{k-1} + \overleftarrow{d}_p) \quad (3.31)$$

$$\overleftarrow{q}_k = \sigma(\overleftarrow{w}_{aq}\overleftarrow{a}_k + \overleftarrow{w}_{lq}\overleftarrow{l}_{k-1} + \overleftarrow{d}_q) \quad (3.32)$$

$$\overleftarrow{r}_k = \sigma(\overleftarrow{w}_{ar}\overleftarrow{a}_k + \overleftarrow{w}_{lr}\overleftarrow{l}_{k-1} + \overleftarrow{d}_r) \quad (3.33)$$

$$\overleftarrow{\tilde{s}}_k = \tanh(\overleftarrow{w}_{as}\overleftarrow{a}_k + \overleftarrow{w}_{ls}\overleftarrow{l}_{k-1} + \overleftarrow{d}_s) \quad (3.34)$$

$$\overleftarrow{s}_k = \overleftarrow{q}_k \odot \overleftarrow{s}_{k-1} + \overleftarrow{p}_k \odot \overleftarrow{\tilde{s}}_k \quad (3.35)$$

$$\overleftarrow{l}_k = \overleftarrow{r}_k \odot \tanh(\overleftarrow{s}_k) \quad (3.36)$$

For all backward equation, the notation is similar to the forward direction, but with an overline to indicate the backward direction. Final hidden state concatenate forward and backward direction that is defined as:

$$l_k^* = [l_k \oplus \overleftarrow{l}_k] \quad (3.37)$$

where \oplus denotes the concatenation operation. Final hidden state can be used for different tasks, in our case it's a classification task.

3.5.4 Gated Recurrent Unit (GRU)

GRU [28] is a simplified version of LSTM to process a sequence of tokens and consists of two gates i.e reset gate and update gate. These two gates decide how much information to keep or discard from the previous and current states [31]. Based on [28], the equations can be defined as:

$$m_k = \sigma(w_{am}a_k + w_{lm}l_{k-1} + d_m) \quad (3.38)$$

$$n_k = \sigma(w_{an}a_k + w_{ln}l_{k-1} + d_n) \quad (3.39)$$

$$\tilde{l}_k = \tanh(w_{al}a_k + w_{ll}l_{k-1} + d_l) \quad (3.40)$$

$$l_k = (1 - n_k) \odot l_{k-1} + n_k \odot \tilde{l}_k \quad (3.41)$$

where σ is sigmoid function, \tanh is tan hyperbolic function, \odot is point-wise multiplication, a_k is input at time step k , m_k , and n_k are reset and update gates respectively. \tilde{l}_k is candidate hidden state, l_k is hidden state. w , and d are the weight matrices and bias for each gate.

3.5.5 BiG

Our BiG model is a sophisticated soft voting classifier, bringing together the two baseline models: Bi-LSTM and GRU. We have fine-tuned each baseline model to optimize the performance of the soft voting classifier. In Table 3.6, and Table 3.7, we provide specific parameters that we've employed in these baseline models based on MFCCs and MELP feature extractions technique respectively.

Table 3.6: Baseline Models parameters of BiG using MFCCs & MoMF

Baseline Model	Parameters
Bi-LSTM	regularizer = 0.01 dropout = 0.2 loss = categorical_crossentropy ephocs = 92 (MFCCs) ephocs = 95 (MoMF)
GRU	batch_size = 64 optimizer = adam ephocs = 100 (MFCCs) ephocs = 220 (MoMF)

Table 3.7: Baseline Models parameters of BiG using MELP & MEWLP

Baseline Model	Parameters
Bi-LSTM	input shape = (81,1) dropout = 0.2 loss = categorical_crossentropy ephocs = 114 (MELP) ephocs = 142 (MEWLP)
GRU	batch_size = 64 input shape = (81,1) optimizer = adam ephocs = 200 (MELP) ephocs = 186 (MEWLP)

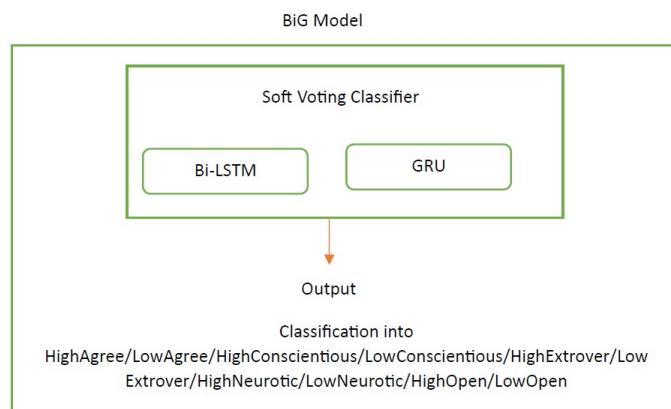


Figure 3.49: Structure of BiG Model

We've provided a visual representation of its structure in Figure 3.49. This combination of Bi-LSTM and GRU, working together as a soft voting classifier, is designed to capture and utilize semantic subtle present in spoken language, ensuring accurate and reliable results in personality classification.

Chapter 4

Results and Discussions

In this section, we're describing the validation results of our experiments and discuss about the models on our dataset of Bangla speeches.

The dataset was partitioned, with 80% reserved for training data and the remaining 20% allocated for validation data.

For training our baseline models, we use Google Colab Pro platform [36]. To gauge how well our models are doing, we used confusion metrics [26], precision, recall, and F-1 score [9]. These help us to understand how good our models for classifying personality traits.

4.0.1 Parameter Selection for Feature Extraction

In the context of MELP, each feature vector is composed of the concatenation of 40 default MFCCs features and 40 LPC features. One-hot encoding is employed for label representation across all feature extraction methods. For MEWLP, 40 LPC features are specified for each audio file, while the *noise_level* is fixed at 0.01. In the case of MFCCs, the number of frames, denoted as T , is set to 40. Concerning MoMF, parameters are configured as follows: $f_c = 1000$, $B = 5$, $t = 30$ in equation (3.13). A total of 30 mel filters (*mel_filters* = 30) are utilized in equation (3.16). To prevent issues associated with the logarithm of zero, ϵ is set to $1e^{-5}$. Furthermore, equation (3.18) involves a specification of $N = 30$, representing the number of coefficients.

4.1 Personality Classification using DistilRo

DistilRo uses two models called DistilBERT and RoBERTa as its foundation. These models are designed to work together to improve the performance of the DistilRo model.

Additionally, to prevent overfitting when dealing with smaller datasets, the models use an l2 regularizer mechanisms. DistilBERT and RoBERTa as baseline models make the DistilRo perform better in classifying the personality traits. In Table 4.1 and Table 4.2, we present the baseline performance of these models when it comes to categorizing individual personality traits.

In Figure 4.1 and Figure 4.2, which display the confusion matrices of our baseline models. Table 4.3 provides a detailed classification report for DistilRo, including its precision, recall, and F1-score for each personality trait. The model performs best

Table 4.1: RoBERTa classification results

Traits	Precision	Recall	F1-score
Agreeableness	0.83	0.88	0.85
Conscientiousness	0.73	0.77	0.75
Openness	0.98	0.84	0.91
Extroversion	0.93	0.8	0.86
Neroticism	0.73	0.88	0.79
Macro Average	0.84	0.83	0.84

Table 4.2: DistilBERT classification results

Traits	Precision	Recall	F1-score
Agreeableness	1.00	0.82	0.9
Conscientiousness	0.76	0.77	0.77
Extroversion	0.85	0.9	0.88
Neroticism	0.81	0.82	0.82
Openness	0.7	0.83	0.76
Macro Average	0.82	0.83	0.83

in classifying Agreeableness and Extroversion, while still achieving good results for Neuroticism, Conscientiousness, and Openness.

In Figure 4.3, which display the confusion matrices of DistilRo model. Table 4.4 provides a comparison between the performance of baseline models and the ensemble model. We can see that RoBERTa performs well for Extroversion and DistilRo performs well for all other traits. Overall model performance, DistilBERT achieve 83% F-1 score, RoBERTa achieve 84% F-1 score and DistilRo achieve 89% F-1 score in the speech-to-text modality. In Figure 4.4, we display the accuracy and F-1 score of each model.

Table 4.3: DistilRo classification results

Traits	Precision	Recall	F1-score
Agreeableness	0.95	0.9	0.92
Conscientiousness	0.9	0.85	0.87
Openness	0.89	0.92	0.9
Neroticism	1.00	0.8	0.89
Extroversion	0.71	0.97	0.82
Macro Average	0.89	0.88	0.89

4.2 Personality Classification using BiG

BiG uses two models called Bi-LSTM and GRU as its foundation. These models are designed to work together to improve the performance of the BiG model. Additionally, to prevent overfitting when dealing with smaller datasets, we use l2 regularizer and early stopping mechanisms. Bi-LSTM and GRU as baseline models make the BiG model perform better in classifying the personality traits.

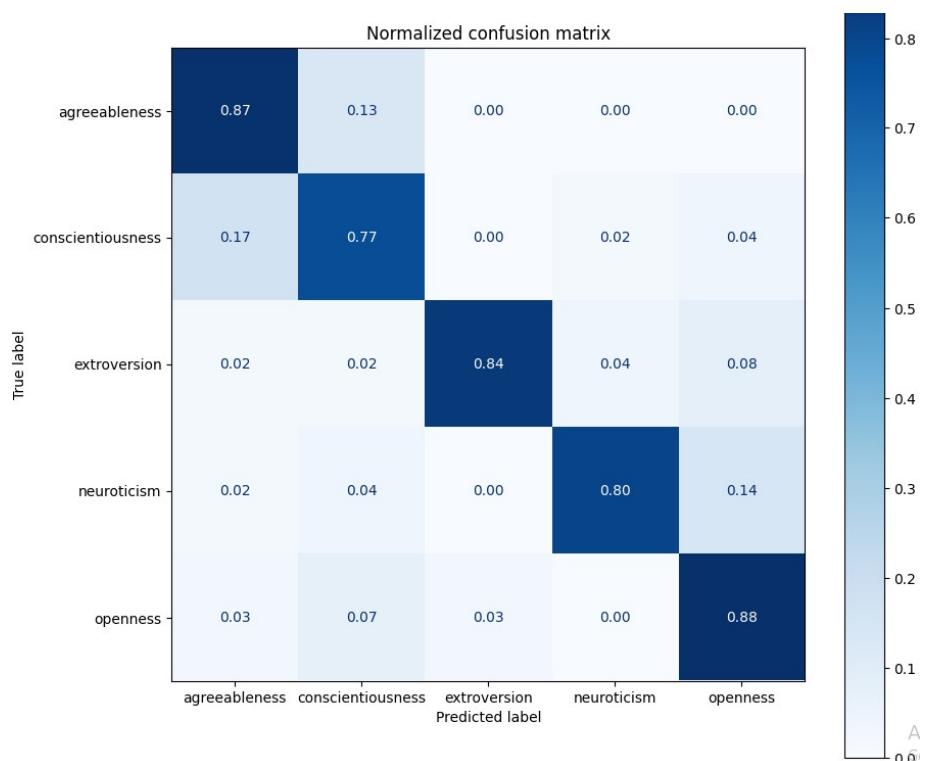


Figure 4.1: Confusion matrix of RoBERTa. RoBERTa effectively captures the Agreeable and Open personality traits. However, a discernible inconsistency of 17% in data alignment with Agreeable is observed within the Conscientious trait. Additionally, a 14% data mismatch is identified between Neurotic and Open traits.

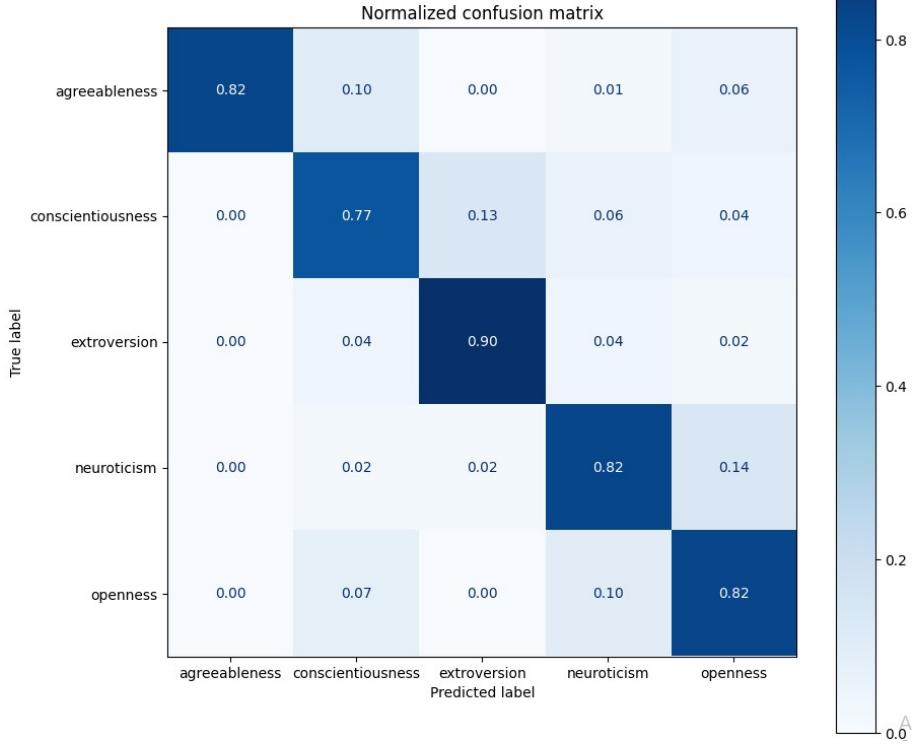


Figure 4.2: Confusion matrix of DistilBERT. DistilBERT effectively captures the Extrover personality traits. However, 13% data mismatch is identified between Conscientious and Extrover traits. Additionally, a 14% data mismatch is identified between Neurotic and Open traits.

Table 4.4: Comparing F-1 score of three models

Traits	DistilBERT	RoBERTa	DistilRo
Agreeableness	0.9	0.85	0.92
Conscientiousness	0.77	0.75	0.87
Extroversion	0.88	0.91	0.9
Neroticism	0.82	0.86	0.89
Openness	0.76	0.79	0.82

4.2.1 MFCC base findings

In Table 4.5 and Table 4.6, we present the baseline performance of these models when it comes to categorizing individual personality traits. In Table 4.5, we see that HC, HE, LC, and LE traits are highly identified by Bi-LSTM based on MFCCs and promising at HA, LA, HN, LN and HO traits but struggle with LO trait. Furthermore, In Table 4.6, we see that HC, HE, and LE traits are highly identified by GRU based on MFCCs and promising at HA, LC, and LN traits but struggle with HN, LA, and LO traits.

In Figure 4.5 and Figure 4.6, which display the confusion matrices of our baseline models. Table 4.7 provides a detailed classification results for BiG, including its precision, recall, and F1-score for each personality traits where HighAgree, LowAgree, HighExtrover, LowExtrover, HighOpen, LowOpen, HighNeurotic, LowNeurotic, HighConscientious, LowConscientious are considering as HA, LA, HE, LE,

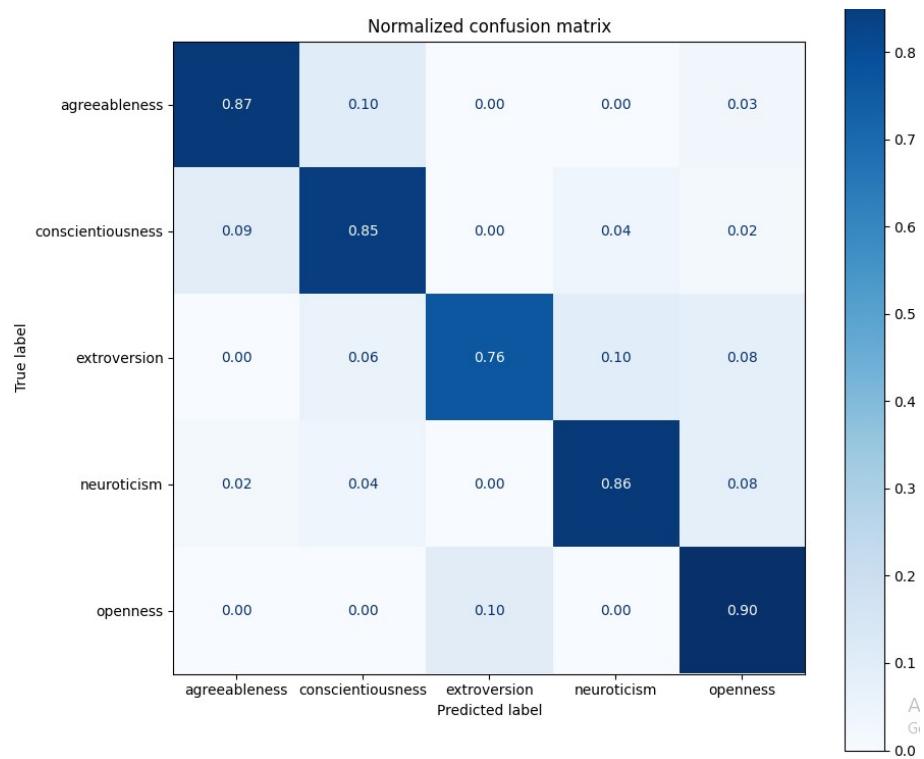


Figure 4.3: Confusion matrix of DistilRo. DistilRo highly captures the Agreeable, Conscientious, Neurotic, and Open personality traits. It encounters challenges in accurately representing the Extrover trait, particularly in establishing distinctions with Neurotic and Open personality traits.

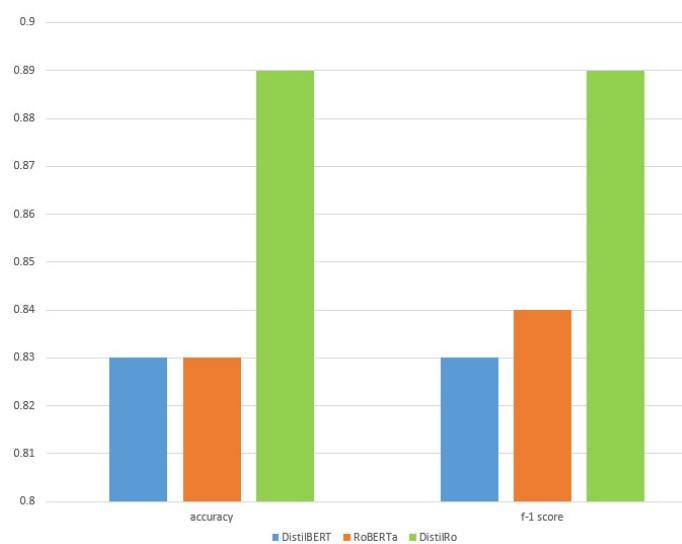


Figure 4.4: Each Model Accuracy and F-1 score of Speech-to-text Modality

Table 4.5: Bi-LSTM classification results based on MFCCs

Traits	Precision	Recall	F1-score
HighAgree	0.73	0.84	0.76
HighConscientious	0.92	0.90	0.91
HighExtrover	0.97	0.94	0.96
HighNeurotic	0.71	0.81	0.75
HighOpen	0.76	0.68	0.70
LowAgree	0.86	0.68	0.77
LowConscientious	0.83	0.95	0.88
LowExtrover	0.97	0.94	0.95
LowNeurotic	0.74	0.85	0.79
LowOpen	0.49	0.53	0.50
Macro Average	0.80	0.79	0.79

Table 4.6: GRU classification results based on MFCCs

Traits	Precision	Recall	F1-score
HighAgree	0.59	0.93	0.72
HighConscientious	0.97	0.90	0.93
HighExtrover	0.92	0.98	0.95
HighNeurotic	0.55	0.84	0.67
HighOpen	0.55	0.31	0.40
LowAgree	0.97	0.36	0.53
LowConscientious	0.72	0.90	0.80
LowExtrover	0.88	0.98	0.93
LowNeurotic	0.72	0.74	0.73
LowOpen	0.67	0.39	0.49
Macro Average	0.75	0.73	0.73

Table 4.7: BiG classification results based on MFCCs

Traits	Precision	Recall	F1-score
HighAgree	0.77	1.00	0.87
HighConscientious	0.94	1.00	0.97
HighExtrover	1.00	1.00	1.00
HighNeurotic	0.74	0.87	0.80
HighOpen	0.73	0.50	0.59
LowAgree	0.94	0.62	0.75
LowConscientious	0.68	0.94	0.79
LowExtrover	0.97	0.97	0.97
LowNeurotic	0.87	0.87	0.87
LowOpen	0.56	0.50	0.53
Macro Average	0.82	0.83	0.81

HO, LO, HN, LN, HC, LC respectively. The model BiG performs best in classifying HC, HE, LE, while still achieving good results for HA, HN, LN, LC, LA and struggle with HO and LO. In Figure 4.7, which display the confusion matrices of BiG model. Bi-LSTM performs well for HO, LA, LC traits and BiG performs well for all other

traits. Overall model performance, Bi-LSTM achieve 79% F-1 score, GRU achieve 73% F-1 score and BiG achieve 81% F-1 score in the speech modality.

	HA	HC	HE	HN	HO	LC	LE	LN	LO	SP
True Labels	20	0	0	0	0	0	0	0	0	3
HA	0	7	0	0	0	0	0	0	0	0
HC	0	3	24	0	0	0	0	0	0	0
HE	0	0	0	17	0	0	0	0	0	0
HN	0	0	0	3	13	0	0	0	4	0
HO	0	0	0	1	12	3	0	0	0	1
LC	0	0	0	0	0	1	25	0	0	0
LE	0	0	0	0	0	0	0	18	0	0
LN	0	0	0	0	1	0	0	4	21	0
LO	0	0	0	2	8	0	0	0	0	9
SP										

Figure 4.5: Confusion matrix of Bi-LSTM based on MFCCs. Bi-LSTM effectively capture all the personality traits.

4.2.2 MoMF base findings

In Table 4.8 and Table 4.9, we present the baseline performance of these models when it comes to categorizing individual personality traits. In Table 4.8, we see that HA, HC, HE, HN, and LC traits are highly identified by Bi-LSTM based on MoMF and promising at HO, LA, LE, LN, and LO traits. Furthermore, In Table 4.9, we see that HA, HC, HE, LE traits are highly identified by GRU based on MoMF and promising at LC, LN, LO traits but struggle with HN, HO, and LA.

In Figure 4.8 and Figure 4.9, which display the confusion matrices of our baseline models. Table 4.10 provides a detailed classification report for BiG, including its precision, recall, and F1-score for each personality traits. The model BiG performs best in classifying HA, HC, HE, LE, and LN, while still achieving good results for HN, LA, LC, LO but struggle with HO. In Figure 4.10, which display the confusion matrices of BiG model. Bi-LSTM performs well for HN, HO, and LC traits. GRU performs well for HC, and LO traits and BiG performs well for all other traits. Overall model performance, Bi-LSTM achieve 80% F-1 score, GRU achieve 79% F-1 score and BiG achieve 84% F-1 score in the speech modality.

4.2.3 MELP base findings

In Table 4.11 and Table 4.12, we present the baseline performance of these models when it comes to categorizing individual personality traits. In Table 4.11, we see

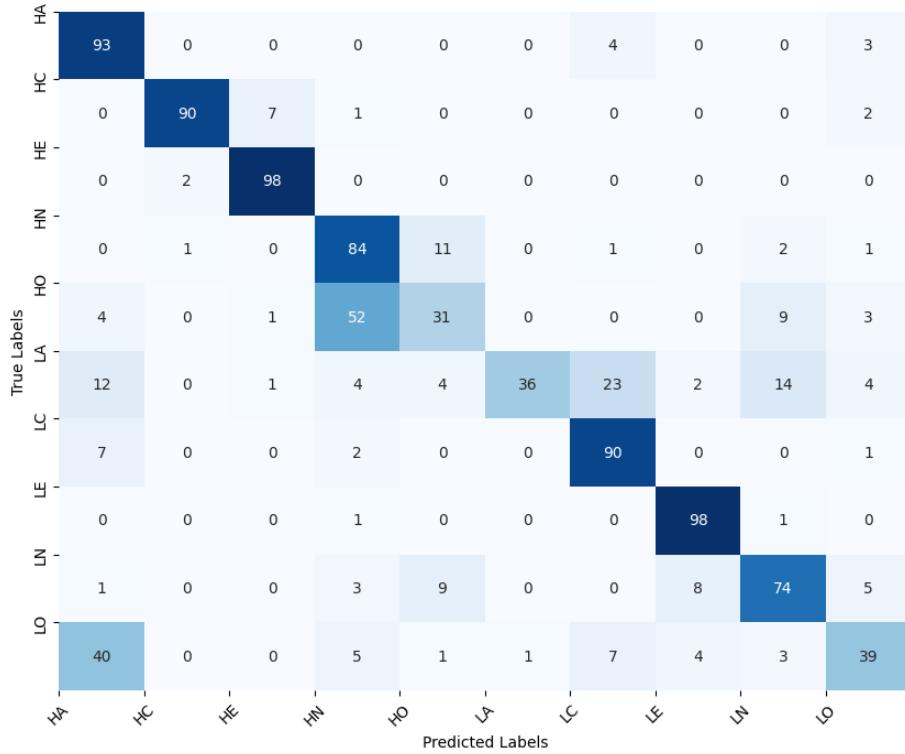


Figure 4.6: Confusion matrix of GRU based on MFCCs. GRU effectively capture all the personality traits except LA, HO and LO. However, most data of LA are mismatch with LC, LN, and HA traits. Additionally, 52 and 40 data of HO and LO traits are considering as HN and HA traits respectively.

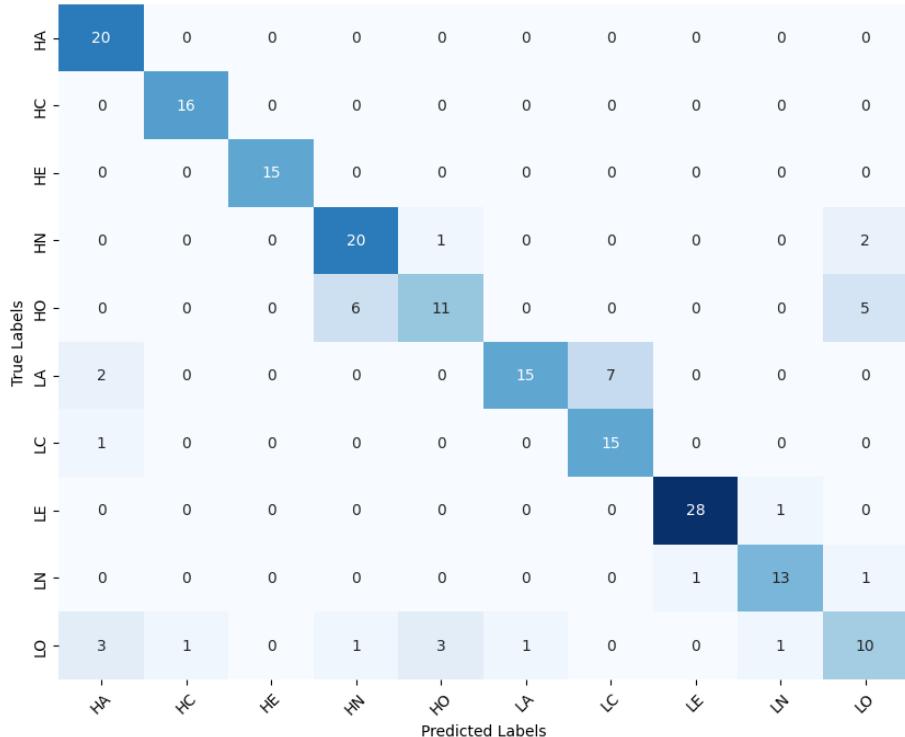


Figure 4.7: Confusion matrix of BiG based on MFCCs. BiG effectively capture all the personality traits except HO and LO.

	HA	HC	HE	HN	HO	LA	LC	LE	LN	LO	
True Labels	22	0	0	0	0	0	0	0	0	0	1
HA	-	0	7	0	0	0	0	0	0	0	0
HC	-	0	9	18	0	0	0	0	0	0	0
HE	-	0	0	0	0	0	0	0	0	0	0
HN	-	0	0	0	17	0	0	0	0	0	0
HO	-	0	1	0	3	13	0	0	0	2	1
LA	-	0	2	0	1	0	7	7	0	0	0
LC	-	0	0	0	1	0	3	22	0	0	0
LE	-	1	0	0	0	0	0	0	17	0	0
LN	-	0	0	0	0	4	0	0	7	15	0
LO	-	1	0	0	0	8	0	0	0	0	10
	HA	HC	HE	HN	HO	LA	LC	LE	LN	LO	

Figure 4.8: Confusion matrix of Bi-LSTM based on MoMF. Bi-LSTM effectively capture all the personality traits but a discernible inconsistency of LA is observed alignment with LC.

	HA	HC	HE	HN	HO	LA	LC	LE	LN	LO	
True Labels	13	0	0	0	0	1	6	0	0	0	0
HA	-	0	16	0	0	0	0	0	0	0	0
HC	-	0	0	15	0	0	0	0	0	0	0
HE	-	0	0	0	0	0	0	0	0	0	0
HN	-	0	0	0	12	9	0	0	0	1	1
HO	-	0	0	0	9	9	0	0	0	1	3
LA	-	0	0	0	0	0	11	13	0	0	0
LC	-	1	0	0	0	0	1	14	0	0	0
LE	-	0	0	0	0	0	0	24	5	0	0
LN	-	0	0	0	0	1	0	0	1	9	4
LO	-	0	0	0	1	1	0	0	1	2	15
	HA	HC	HE	HN	HO	LA	LC	LE	LN	LO	

Figure 4.9: Confusion matrix of GRU based on MoMF. GRU effectively capture all the personality traits except LA and HO. However, most data of LA are identifying as LC. Additionally, for HO trait data, model don't separate HO and HN traits adequately.

Table 4.8: Bi-LSTM classification results based on MoMF

Traits	Precision	Recall	F1-score
HighAgree	0.89	0.91	0.90
HighConscientious	0.74	0.97	0.84
HighExtrover	0.96	0.76	0.85
HighNeurotic	0.86	0.96	0.91
HighOpen	0.67	0.75	0.71
LowAgree	0.89	0.59	0.71
LowConscientious	0.74	0.89	0.81
LowExtrover	0.73	0.87	0.79
LowNeurotic	0.77	0.67	0.72
LowOpen	0.86	0.62	0.72
Macro Average	0.81	0.80	0.80

Table 4.9: GRU classification results based on MoMF

Traits	Precision	Recall	F1-score
HighAgree	0.95	0.86	0.90
HighConscientious	0.95	0.95	0.95
HighExtrover	0.99	0.97	0.98
HighNeurotic	0.66	0.67	0.67
HighOpen	0.66	0.54	0.59
LowAgree	0.89	0.56	0.69
LowConscientious	0.62	0.86	0.72
LowExtrover	0.90	0.94	0.92
LowNeurotic	0.79	0.69	0.74
LowOpen	0.66	0.89	0.76
Macro Average	0.80	0.79	0.79

Table 4.10: BiG classification results based on MoMF

Traits	Precision	Recall	F1-score
HighAgree	0.95	1.00	0.98
HighConscientious	0.89	1.00	0.94
HighExtrover	1.00	1.00	1.00
HighNeurotic	0.69	0.96	0.80
HighOpen	0.71	0.45	0.56
LowAgree	1.00	0.62	0.77
LowConscientious	0.64	1.00	0.78
LowExtrover	0.91	1.00	0.95
LowNeurotic	0.93	0.87	0.90
LowOpen	0.86	0.60	0.71
Macro Average	0.86	0.85	0.84

that HA, HC, HE, HN, LA, LC, and LE traits are highly identified by Bi-LSTM based on MELP and promising at HO, LN, and Lo traits. Furthermore, In Table 4.12, we see that HC, HE, LC, and LE traits are highly identified by GRU based on MELP and promising at HA, HN, LA, and LN traits but struggle with HO, and

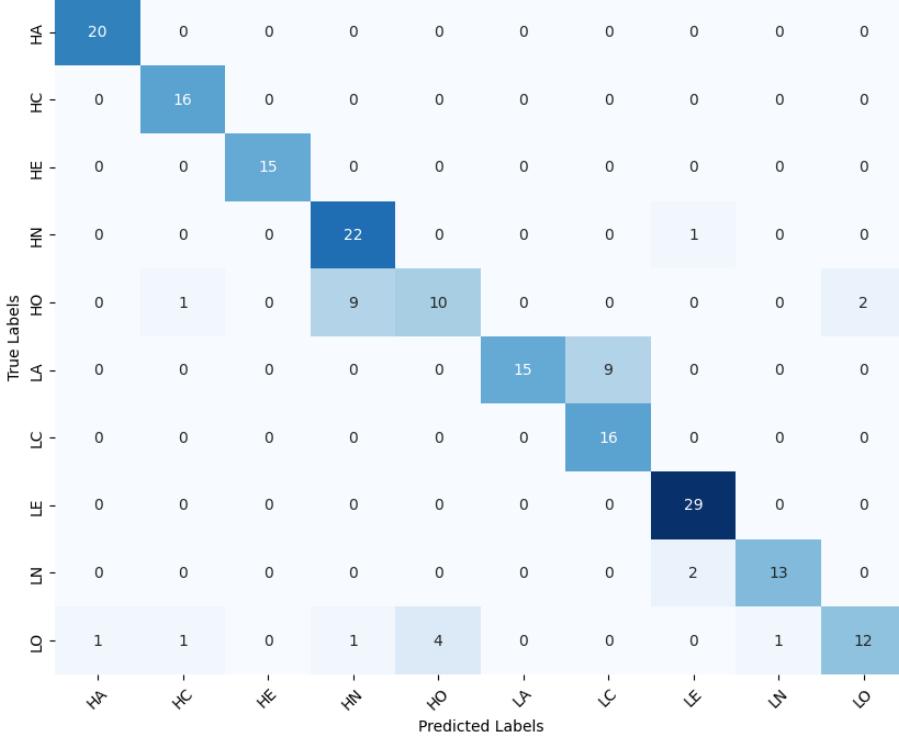


Figure 4.10: Confusion matrix of BiG based on MoMF. BiG effectively capture all the personality traits but model struggle with HO trait data.

LO traits.

Table 4.11: Bi-LSTM classification results based on MELP

Traits	Precision	Recall	F1-score
HighAgree	0.89	0.90	0.89
HighConscientious	0.91	0.90	0.91
HighExtrover	0.94	0.93	0.92
HighNeurotic	0.91	0.96	0.94
HighOpen	0.70	0.73	0.71
LowAgree	0.94	0.81	0.87
LowConscientious	0.86	0.91	0.88
LowExtrover	0.89	0.90	0.89
LowNeurotic	0.79	0.81	0.80
LowOpen	0.90	0.63	0.73
Macro Average	0.87	0.85	0.85

In Figure 4.11 and Figure 4.12, which display the confusion matrices of our baseline models. Table 4.13 provides a detailed classification results for BiG, including its precision, recall, and F1-score for each personality traits. The model BiG performs best in classifying HA, HC, HE, HN and LE, while still achieving good results for LA, HO, LC, LN and LO. In Figure 4.13, which display the confusion matrices of BiG model. Bi-LSTM perfoms well for HN traits and GRU performs well for HC, LE, LA, LC, LE traits and BiG performs well for all other traits. Overall model performance, Bi-LSTM achieve 85% F-1 score, GRU achieve 82% F-1 score and BiG achieve 88% F-1 score in the speech modality.

	HA	HC	HE	HN	HO	LA	LC	LE	LN	W	LO
True Labels	18	0	0	0	0	0	0	0	0	1	0
HA											
HC	0	17	0	0	1	0	0	0	0	0	0
HE	0	1	21	0	0	0	0	0	0	0	0
HN	0	0	0	17	0	0	0	0	0	0	0
HO	0	0	0	3	14	0	0	0	3	1	0
LA	1	0	0	0	0	21	2	1	0	0	0
LC	1	0	0	0	0	0	18	0	0	0	0
LE	0	0	0	0	0	0	0	14	0	0	0
LN	0	0	0	0	1	0	0	0	3	25	0
LO	0	0	0	0	5	0	0	0	1	0	10

Figure 4.11: Confusion matrix of Bi-LSTM based on MELP. Bi-LSTM effectively capture all the personality traits.

	HA	HC	HE	HN	HO	LA	LC	LE	LN	W	LO
True Labels	18	0	0	0	0	0	0	0	1	0	0
HA											
HC	1	15	1	0	0	0	0	0	0	0	1
HE	0	0	22	0	0	0	0	0	0	0	0
HN	0	0	0	11	4	1	0	0	1	0	0
HO	0	0	0	3	15	0	0	0	2	1	0
LA	1	0	0	0	0	19	3	1	1	1	0
LC	1	0	0	0	0	0	18	0	0	0	0
LE	0	0	0	0	0	0	0	14	0	0	0
LN	2	0	0	1	0	0	0	4	21	1	0
LO	6	0	0	0	4	0	0	1	0	5	0

Figure 4.12: Confusion matrix of GRU based on MELP. GRU effectively capture all the personality traits except LO trait. Additionally, most of the data of LO trait are considering as HA trait.

Table 4.12: GRU classification results based on MELP

Traits	Precision	Recall	F1-score
HighAgree	0.65	0.97	0.78
HighConscientious	0.97	0.91	0.94
HighExtrover	0.97	0.98	0.98
HighNeurotic	0.72	0.84	0.78
HighOpen	0.68	0.63	0.65
LowAgree	0.96	0.81	0.88
LowConscientious	0.91	0.96	0.94
LowExtrover	0.90	0.99	0.94
LowNeurotic	0.82	0.79	0.81
LowOpen	0.79	0.41	0.54
Macro Average	0.84	0.83	0.82

Table 4.13: BiG classification results based on MELP

Traits	Precision	Recall	F1-score
HighAgree	1.00	1.00	1.00
HighConscientious	0.94	0.89	0.91
HighExtrover	0.96	1.00	0.98
HighNeurotic	0.84	0.94	0.89
HighOpen	0.94	0.71	0.81
LowAgree	1.00	0.64	0.78
LowConscientious	0.68	1.00	0.81
LowExtrover	0.88	1.00	0.93
LowNeurotic	0.89	0.86	0.88
LowOpen	0.78	0.88	0.82
Macro Average	0.89	0.89	0.88

4.2.4 MEWLP base findings

In Table 4.14 and Table 4.15, we present the baseline performance of these models when it comes to categorizing individual personality traits. In Table 4.14, we see that HA, HC, HE, HN, LA, LC, and LE traits are highly identified by Bi-LSTM based on MEWLP and promising at HO, LN, and LO traits. Furthermore, In Table 4.15, we see that HC, and HE traits are highly identified by GRU based on MEWLP and promising at LE, and HN traits but struggle with HA, HO, LA, LC, LN, and LO traits.

In Figure 4.14 and Figure 4.15, which display the confusion matrices of our baseline models. Table 4.16 provides a detailed classification results for BiG, including its precision, recall, and F1-score for each personality traits. The model BiG performs best in classifying HA, HC, HE, HN, LE, and LN traits while still achieving good results for HO, LA, LC, and LO traits. In Figure 4.16, which display the confusion matrices of BiG model. Bi-LSTM performs well for HA, LA, LC, and LO traits and BiG performs well for all other traits. Overall model performance, Bi-LSTM achieve 87% F-1 score, GRU achieve 57% F-1 score and BiG achieve 90% F-1 score in the speech modality.

	HA	HC	HE	HN	HO	LA	LC	LE	LN	LO	LO
True Labels	19	0	0	0	0	0	0	0	0	0	0
HA											
HC	0	15	2	0	1	0	0	0	0	0	0
HE	0	0	22	0	0	0	0	0	0	0	0
HN	0	1	0	16	0	0	0	0	0	0	0
HO	0	0	0	3	17	0	0	0	1	0	0
LA	1	0	0	0	0	13	11	0	0	0	0
LC	0	0	0	0	0	0	19	0	0	0	0
LE	0	0	0	0	0	0	0	14	0	0	0
LN	0	0	0	0	1	0	0	0	7	21	0
LO	1	0	0	0	5	0	0	0	1	9	0
HA											
HC											
HE											
HN											
HO											
LA											
LC											
LE											
LN											
LO											

Figure 4.13: Confusion matrix of BiG based on MELP. BiG correctly capture all the personality traits but struggle with LA trait. However, to identify LA trait it sometime consider LC trait.

	HA	HC	HE	HN	HO	LA	LC	LE	LN	LO	LO
True Labels	20	0	0	0	0	0	0	0	0	0	0
HA											
HC	0	16	0	0	0	0	0	0	0	0	0
HE	0	1	14	0	0	0	0	0	0	0	0
HN	0	0	0	19	3	0	0	1	0	0	0
HO	0	0	0	3	18	0	0	0	1	0	0
LA	1	0	0	0	0	18	5	0	0	0	0
LC	0	0	0	0	0	2	14	0	0	0	0
LE	0	0	0	0	0	0	0	25	4	0	0
LN	0	0	0	0	0	0	0	1	14	0	0
LO	0	0	0	0	4	0	1	1	2	12	0
HA											
HC											
HE											
HN											
HO											
LA											
LC											
LE											
LN											
LO											

Figure 4.14: Confusion matrix of Bi-LSTM based on MEWLP. Bi-LSTM effectively capture all the personality traits.

	HA	HC	HE	HN	HO	LA	LC	LE	LN	LO
True Labels	4	0	0	0	1	1	2	0	10	2
HA										
HC	0	16	0	0	0	0	0	0	0	0
HE	0	0	15	0	0	0	0	0	0	0
HN	0	0	0	19	1	0	0	0	2	1
HO	0	0	0	17	4	0	0	0	1	0
LA	0	0	0	1	2	9	5	3	3	1
LC	0	0	0	0	0	1	8	6	1	0
LE	0	0	0	0	0	0	0	29	0	0
LN	0	0	0	0	0	0	0	0	13	2
LO	0	0	0	5	4	1	1	0	6	3
	HA	HC	HE	HN	HO	LA	LC	LE	LN	LO

Figure 4.15: Confusion matrix of GRU based on MEWLP. GRU can't capture all the personality traits. However HA, HO, LA, LC, and LO are mismatch with LN, HN, LC, LE, and LN traits respectively.

	HA	HC	HE	HN	HO	LA	LC	LE	LN	LO
True Labels	20	0	0	0	0	0	0	0	0	0
HA										
HC	0	16	0	0	0	0	0	0	0	0
HE	0	0	15	0	0	0	0	0	0	0
HN	0	0	0	22	1	0	0	0	0	0
HO	0	0	0	3	17	0	0	0	1	1
LA	0	0	0	0	0	17	7	0	0	0
LC	0	0	0	0	0	1	15	0	0	0
LE	0	0	0	0	0	0	0	29	0	0
LN	0	0	0	0	0	0	0	0	15	0
LO	3	0	0	0	3	0	0	1	0	13
	HA	HC	HE	HN	HO	LA	LC	LE	LN	LO

Figure 4.16: Confusion matrix of BiG based on MEWLP. BiG effectively capture all the personality traits except LO.

Table 4.14: Bi-LSTM classification results based on MEWLP

Traits	Precision	Recall	F1-score
HighAgree	0.93	0.96	0.94
HighConscientious	0.90	0.94	0.92
HighExtrover	1.00	0.93	0.95
HighNeurotic	0.88	0.91	0.89
HighOpen	0.75	0.80	0.77
LowAgree	0.93	0.81	0.86
LowConscientious	0.80	0.96	0.88
LowExtrover	0.91	0.86	0.87
LowNeurotic	0.77	0.95	0.85
LowOpen	0.90	0.70	0.79
Macro Average	0.88	0.88	0.87

Table 4.15: GRU classification results based on MEWLP

Traits	Precision	Recall	F1-score
HighAgree	0.53	0.20	0.29
HighConscientious	0.96	0.97	0.97
HighExtrover	0.97	0.98	0.98
HighNeurotic	0.48	0.83	0.61
HighOpen	0.31	0.16	0.21
LowAgree	0.77	0.37	0.50
LowConscientious	0.57	0.47	0.51
LowExtrover	0.64	0.96	0.77
LowNeurotic	0.41	0.90	0.56
LowOpen	0.50	0.17	0.25
Macro Average	0.61	0.60	0.57

Table 4.16: BiG classification results based on MEWLP

Traits	Precision	Recall	F1-score
HighAgree	0.87	1.00	0.93
HighConscientious	1.00	1.00	1.00
HighExtrover	1.00	1.00	1.00
HighNeurotic	0.88	0.96	0.92
HighOpen	0.81	0.77	0.79
LowAgree	0.94	0.71	0.81
LowConscientious	0.68	0.94	0.79
LowExtrover	0.97	1.00	0.98
LowNeurotic	0.94	1.00	0.97
LowOpen	0.93	0.65	0.76
Macro Average	0.91	0.90	0.90

Chapter 5

Conclusion

Our primary research focus centered around the classification of personality traits using Bangla speech, conducted in two distinct modalities: speech-to-text and speech analysis.

We created our own dataset, comprising 1,750 speeches for the speech-to-text modality and 1,000 acted speeches for the speech modality. Dataset will be made available to public to support progress of the research in this area. In the former, we categorized the data into five distinct personality trait classes, while the latter featured ten classes. Our ensemble models, DistilRo and BiG, delivered impressive results in accurately classifying personality traits across both phases. DistilRo achieved an outstanding 89% F-1 score in the speech-to-text modality, while BiG achieved an impressive 81% F-1 score based on MFCCs, 84% F-1 score based on MoMF, 88% F-1 score based on MELP, and 90% F-1 score based on MEWLP in the speech modality.

Total training time energy consumed is 50.40 kWh and carbon emissions 30.24 Kg.

One of the main obstacles we faced during our research was related to creating datasets. We recorded an experimental dataset featuring a non-professional speaker. As our speaker lacked professional acting skills, so it was a high challenge for him to act for each personality traits. Our assessors were unfamiliar with the speaker and were only given the speech to annotate. The short textual data in our speeches presented a challenge for the classification task. We extracted signal-based features that captured acoustic speech properties. We observed that in the DistilRo model, DistilBERT struggled somewhat in classifying conscientiousness and openness.

Similarly, in the BiG model, the GRU faced confusion when distinguishing between high and low agreeableness (HA, LA) and high and low openness (HO, LO). In most cases, machine consider openness trait data as neuroticism trait data. These difficulties stem from the limited amount of data available for each personality trait and we also notice that in data reliability openness trait Gaussian curves mostly overlap that means distinguishing between high and low trait are less straightforward. However, the models performed well in discriminating between high and low extroversion in both phases.

For future research, we intend to focus on factor reduction, particularly exploring

the correlations between NEO-FFI factors and prosodic and acoustic signal-based features. While our present dataset was collected from a single non-professional speaker, our future experiments will encompass both professional and non-professional speakers, as well as a more diverse range of text materials. Additionally, we see potential in investigating the identification of emotions and exploring the relationship between emotions and personality traits within speech, offering promising avenues for further research in this field.

Bibliography

- [1] L. R. Goldberg, “The structure of phenotypic personality traits.,” *American psychologist*, vol. 48, no. 1, pp. 26–34, 1993.
- [2] T. G. Dietterich, “Ensemble methods in machine learning,” in *International workshop on multiple classifier systems*, Springer, 2000, pp. 1–15.
- [3] J. Lin and L. Qu, “Feature extraction based on morlet wavelet and its application for mechanical fault diagnosis,” *Journal of sound and vibration*, vol. 234, no. 1, pp. 135–148, 2000.
- [4] A. H. Gunatilaka and B. A. Baertlein, “Feature-level and decision-level fusion of noncoincidentally sampled sensors for land mine detection,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 23, no. 6, pp. 577–589, 2001.
- [5] P. Baker, A. Hardie, T. McEnery, H. Cunningham, and R. J. Gaizauskas, “Emille, a 67-million word corpus of indic languages: Data collection, mark-up and harmonisation.,” in *LREC*, 2002.
- [6] G. Matthews, I. J. Deary, and M. C. Whiteman, *Personality traits*. Cambridge University Press, 2003.
- [7] L. A. Pervin, *The science of personality*. Oxford university press, 2003.
- [8] R. R. McCrae and P. T. Costa Jr, “A contemplated revision of the neo five-factor inventory,” *Personality and individual differences*, vol. 36, no. 3, pp. 587–596, 2004.
- [9] C. Goutte and E. Gaussier, “A probabilistic interpretation of precision, recall and f-score, with implication for evaluation,” in *European conference on information retrieval*, Springer, 2005, pp. 345–359.
- [10] J. Chen, J. Benesty, Y. Huang, and S. Doclo, “New insights into the noise reduction wiener filter,” *IEEE Transactions on audio, speech, and language processing*, vol. 14, no. 4, pp. 1218–1234, 2006.
- [11] R. Ober, “Kapati time: Storytelling as a data collection method in indigenous research,” *Mystery Train*, 2007.
- [12] P. T. Costa and R. R. McCrae, “The revised neo personality inventory (neo-*pi-r*),” *The SAGE handbook of personality theory and assessment*, vol. 2, no. 2, pp. 179–198, 2008.
- [13] R. W. Black, “Online fan fiction, global identities, and imagination,” *Research in the Teaching of English*, pp. 397–425, 2009.

- [14] T. Polzehl, S. Möller, and F. Metze, “Automatically assessing personality from speech,” in *2010 IEEE fourth international conference on semantic computing*, IEEE, 2010, pp. 134–140.
- [15] M. Tavakol and R. Dennick, “Making sense of cronbach’s alpha,” *International journal of medical education*, vol. 2, p. 53, 2011.
- [16] I. Trabelsi and D. B. Ayed, “On the use of different feature extraction methods for linear and non linear kernels,” in *2012 6th international conference on sciences of electronics, technologies of information and telecommunications (SETIT)*, IEEE, 2012, pp. 797–802.
- [17] F. Alam and G. Riccardi, “Comparative study of speaker personality traits recognition in conversational and broadcast news speech.,” in *INTERSPEECH*, 2013, pp. 2851–2855.
- [18] A. Graves, A.-r. Mohamed, and G. Hinton, “Speech recognition with deep recurrent neural networks,” in *2013 IEEE international conference on acoustics, speech and signal processing*, Ieee, 2013, pp. 6645–6649.
- [19] J. D. Greer and D. Mensing, “The evolution of online newspapers: A longitudinal content analysis, 1997–2003,” in *Internet Newspapers*, Routledge, 2013, pp. 13–32.
- [20] H. A. Schwartz, J. C. Eichstaedt, M. L. Kern, *et al.*, “Personality, gender, and age in the language of social media: The open-vocabulary approach,” *PloS one*, vol. 8, no. 9, e73791, 2013.
- [21] B. Verhoeven, W. Daelemans, and T. De Smedt, “Ensemble methods for personality recognition,” in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 7, 2013, pp. 35–38.
- [22] J. M. Burger, *Personality*. Cengage Learning, 2014.
- [23] R. Cotterell, A. Renduchintala, N. Saphra, and C. Callison-Burch, “An algerian arabic-french code-switched corpus,” in *Workshop on free/open-source arabic corpora and corpora processing tools workshop programme*, 2014, p. 34.
- [24] D. G. Bonett and T. A. Wright, “Cronbach’s alpha reliability: Interval estimation, hypothesis testing, and sample size planning,” *Journal of organizational behavior*, vol. 36, no. 1, pp. 3–15, 2015.
- [25] T. Polzehl, “Personality in speech,” *Assessment and automatic classification*, 2015.
- [26] X. Deng, Q. Liu, Y. Deng, and S. Mahadevan, “An improved method to construct basic probability assignment based on the confusion matrix for classification problem,” *Information Sciences*, vol. 340, pp. 250–261, 2016.
- [27] S. Basu, J. Chakraborty, A. Bag, and M. Aftabuddin, “A review on emotion recognition using speech,” in *2017 International conference on inventive communication and computational technologies (ICICCT)*, IEEE, 2017, pp. 109–114.
- [28] R. Dey and F. M. Salem, “Gate-variants of gated recurrent unit (gru) neural networks,” in *2017 IEEE 60th international midwest symposium on circuits and systems (MWSCAS)*, IEEE, 2017, pp. 1597–1600.

- [29] P. Valenzuela, M. J. Domínguez-Cuesta, M. A. M. García, and M. Jiménez-Sánchez, “A spatio-temporal landslide inventory for the nw of spain: Bapa database,” *Geomorphology*, vol. 293, pp. 11–23, 2017.
- [30] J. Yu and K. Markov, “Deep learning based personality recognition from facebook status updates,” in *2017 IEEE 8th international conference on awareness science and technology (iCAST)*, IEEE, 2017, pp. 383–387.
- [31] J. Yu and K. Markov, “Deep learning based personality recognition from facebook status updates,” in *2017 IEEE 8th international conference on awareness science and technology (iCAST)*, IEEE, 2017, pp. 383–387.
- [32] S. A. Alim and N. K. A. Rashid, *Some commonly used speech feature extraction algorithms*. IntechOpen London, UK: 2018.
- [33] J. R. Saurav, S. Amin, S. Kibria, and M. S. Rahman, “Bangla speech recognition for voice search,” in *2018 international conference on Bangla speech and language processing (ICBSLP)*, IEEE, 2018, pp. 1–4.
- [34] N. I. Tripto and M. E. Ali, “Detecting multilabel sentiment and emotions from bangla youtube comments,” in *2018 International Conference on Bangla Speech and Language Processing (ICBSLP)*, IEEE, 2018, pp. 1–6.
- [35] S. Ahammed, M. Rahman, M. H. Niloy, and S. M. H. Chowdhury, “Implementation of machine learning to detect hate speech in bangla language,” in *2019 8th International Conference System Modeling and Advancement in Research Trends (SMART)*, IEEE, 2019, pp. 317–320.
- [36] E. Bisong *et al.*, *Building machine learning and deep learning models on Google cloud platform*. Springer, 2019.
- [37] M. X. Cohen, “A better way to define and describe morlet wavelets for time-frequency analysis,” *NeuroImage*, vol. 199, pp. 81–86, 2019.
- [38] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, *Bert: Pre-training of deep bidirectional transformers for language understanding*, 2019. arXiv: 1810.04805 [cs.CL].
- [39] T. Ebert, J. E. Gebauer, T. Brenner, *et al.*, “Are regional differences in personality and their correlates robust? applying spatial analysis techniques to examine regional variation in personality across the us and germany,” Working papers on Innovation and Space, Tech. Rep., 2019.
- [40] N. Islam *et al.*, “The big five model of personality in bangladesh: Examining the ten-item personality inventory,” *psihologija*, vol. 52, no. 4, pp. 395–412, 2019.
- [41] Y. Liu, M. Ott, N. Goyal, *et al.*, “Roberta: A robustly optimized bert pre-training approach,” *arXiv preprint arXiv:1907.11692*, 2019.
- [42] H. Mochahary, “Translation literature in bodo language,” *Translation Literature*, 2019.
- [43] K. Pisanski and G. A. Bryant, “The evolution of voice perception,” *The Oxford handbook of voice studies*, pp. 269–300, 2019.
- [44] F. Rahman, H. Khan, Z. Hossain, *et al.*, “An annotated bangla sentiment analysis corpus,” in *2019 International Conference on Bangla Speech and Language Processing (ICBSLP)*, IEEE, 2019, pp. 1–5.

- [45] M. Shuvo, S. A. Shahriyar, and M. Akhand, “Bangla numeral recognition from speech signal using convolutional neural network,” in *2019 International Conference on Bangla Speech and Language Processing (ICBSLP)*, IEEE, 2019, pp. 1–4.
- [46] A. F. Adoma, N.-M. Henry, and W. Chen, “Comparative analyses of bert, roberta, distilbert, and xlnet for text-based emotion recognition,” in *2020 17th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)*, IEEE, 2020, pp. 117–121.
- [47] A. F. Adoma, N.-M. Henry, and W. Chen, “Comparative analyses of bert, roberta, distilbert, and xlnet for text-based emotion recognition,” in *2020 17th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)*, IEEE, 2020, pp. 117–121.
- [48] R. S. Camati and F. Enembreck, “Text-based automatic personality recognition: A projective approach,” in *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, IEEE, 2020, pp. 218–225.
- [49] U. Rudra, A. N. Chy, and M. H. Seddiqi, “Personality traits detection in bangla: A benchmark dataset with comparative performance analysis of state-of-the-art methods,” in *2020 23rd International Conference on Computer and Information Technology (ICCIT)*, IEEE, 2020, pp. 1–6.
- [50] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, *Distilbert, a distilled version of bert: Smaller, faster, cheaper and lighter*, 2020. arXiv: 1910.01108 [cs.CL].
- [51] R. N. Swarna, “Bangla broadcast speech recognition using support vector machine,” in *2020 Emerging Technology in Computing, Communication and Electronics (ETCCE)*, IEEE, 2020, pp. 1–6.
- [52] Y. Arslan, K. Allix, L. Veiber, *et al.*, “A comparison of pre-trained language models for multi-class text classification in the financial domain,” in *Companion Proceedings of the Web Conference 2021*, 2021, pp. 260–268.
- [53] L. Fu, P. Liang, X. Li, and C. Yang, “A machine learning based ensemble method for automatic multiclass classification of decisions,” in *Evaluation and Assessment in Software Engineering*, 2021, pp. 40–49.
- [54] M. Labied and A. Belangour, “Automatic speech recognition features extraction techniques: A multi-criteria comparison,” *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 8, 2021.
- [55] S. Sultana, M. Z. Iqbal, M. R. Selim, M. M. Rashid, and M. S. Rahman, “Bangla speech emotion recognition and cross-lingual study using deep cnn and blstm networks,” *IEEE Access*, vol. 10, pp. 564–578, 2021.
- [56] S. Ullah and Z. Halim, “Imagined character recognition through eeg signals using deep convolutional neural network,” *Medical & Biological Engineering & Computing*, vol. 59, no. 5, pp. 1167–1183, 2021.
- [57] M. D. Kamalesh and B. Bharathi, “Personality prediction model for social media using machine learning technique,” *Computers and Electrical Engineering*, vol. 100, p. 107852, 2022.
- [58] M. Ramezani, M.-R. Feizi-Derakhshi, M.-A. Balafar, *et al.*, “Automatic personality prediction: An enhanced method using ensemble modeling,” *Neural Computing and Applications*, vol. 34, no. 21, pp. 18369–18389, 2022.

- [59] M. Sarker, M. F. Hossain, F. R. Liza, S. N. Sakib, and A. Al Farooq, “A machine learning approach to classify anti-social bengali comments on social media,” in *2022 International Conference on Advancement in Electrical and Electronic Engineering (ICAEEE)*, IEEE, 2022, pp. 1–6.
- [60] O. Sen, M. Fuad, M. N. Islam, *et al.*, “Bangla natural language processing: A comprehensive analysis of classical, machine learning, and deep learning-based methods,” *IEEE Access*, vol. 10, pp. 38 999–39 044, 2022.
- [61] L. Zhou, Z. Zhang, L. Zhao, and P. Yang, “Attention-based bilstm models for personality recognition from user-generated content,” *Information Sciences*, vol. 596, pp. 460–471, 2022.
- [62] S. K. Bitew, V. Schelstraete, K. Zaporojets, K. Van Nieuwenhove, R. Meganck, and C. Develder, “Personality style recognition via machine learning: Identifying anaclitic and introjective personality styles from patients’ speech,” *arXiv preprint arXiv:2311.04088*, 2023.
- [63] A. A. Efat, A. Atiq, A. S. Abeed, A. Momin, and M. G. R. Alam, “Empoliticon: Nlp and ml based approach for context and emotion classification of political speeches from transcripts,” *IEEE Access*, 2023.
- [64] A. R. Julianda, W. Maharani, *et al.*, “Personality detection on reddit using distilbert,” *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 7, no. 5, pp. 1140–1146, 2023.
- [65] M. S. I. Malik, A. Nazarova, M. M. Jamjoom, and D. I. Ignatov, “Multilingual hope speech detection: A robust framework using transfer learning of fine-tuning roberta model,” *Journal of King Saud University-Computer and Information Sciences*, vol. 35, no. 8, p. 101 736, 2023.
- [66] F. Safari and A. Chalechale, “Emotion and personality analysis and detection using natural language processing, advances, challenges and future scope,” *Artificial Intelligence Review*, pp. 1–25, 2023.
- [67] N. Sujatha, S. Pramod, S. Bhatla, T. Thulasimani, R. Kant, and A. Chauhan, “Efficient method for personality prediction using hybrid method of convolutional neural network and lstm,” in *2023 4th International Conference on Electronics and Sustainable Communication Systems (ICESC)*, IEEE, 2023, pp. 959–964.