CARNEGIE MELLON UNIVERSITY

Computational Techniques for Voice Intelligence: Deducing Psychological Factors from Human Voice

Submitted in partial fulfillment for
the degree of
Doctor of Philosophy
in
Language Technologies Institute

by

Hira Dhamyal

Thesis committee:

Rita Singh	Carnegie Mellon University (Chair)
Bhiksha Ramakrishnan	Carnegie Mellon University
Richard M. Stern	Carnegie Mellon University
Helen Meng	The Chinese University of Hong Kong

February 2025

Acknowledgements

I am extremely grateful to my supervisors Dr. Rita Singh and Dr. Bhiksha Raj, for the opportunities and guidance they have given me over many years. I first met them in last semester of my undergraduate degree and was fortunate to take the Machine Learning course with them. They saw something in me at that time and took me on as a research assistant in their lab a year later. I was introduced to the world of research through their lab and continued to learn and grow, eventually becoming their PhD student. They have believed in me, pushed me, and taught me to do better. Their influence has not only been on my academic growth but also in my personal life. They have taught me to be a believer in the goodness within myself and others, to be resilient in the face of odds, to keep pushing regardless of setbacks. They have given me a home in this part of the world and I would always be grateful and look up to them for guidance.

My PhD journey would have been impossible without the support of my friends and collaborators. I owe big thanks to the entire MLSP group and the legacy they carry. I am grateful to Dr. Roshan Sharma for his friendship, and always lending an ear to my ideas and pushing me to speak up, to Dr. Benjamin Elizalde for always giving advice whenever I needed, to Ahmed Shah for his friendship and for being a guiding light, to Mark Lindsay, Ankit Shah, Joseph Konan, Soham Deshmukh, Shahan Ali Memon, Aqsa Kashaf, Xiang Li, Hao Chen, Raphael Olivier, Dareen Alharthi, Hazim Bukhari, Massa Baali, Satvik Dixit, Raymond Xia, David Bick, Sarthak Bisht, Yandong Wen, Wayne Zhao, Yang Gao, Abelino Jimenez, and to many others. I am grateful to the three internship experiences I had with Microsoft and Meta. My work at Microsoft under Dr. Benjamin Elizalde has directly impacted my thesis. My work at Meta under Dr. Leda Sari and Dr. Ehab AlBadawy were great learning experiences and have opened doors to new opportunities for me. I want to thank my friends in Pittsburgh with whom I cherish many chai sessions and gossips and who make my life brighter.

Thank you to the members of my thesis committee for accepting to be on my committee. I am grateful to Prof Richard Stern, who has provided valuable feedback to me and has been an inspiration throughout my PhD journey. I have heard his stories with awe. I am grateful to Prof Helen Meng who has given valuable feedback on my work. I hope to stay in touch and

collaborate in the future. I want to thank the various sponsors who have funded my research over the years. In chronological order, I would like to thank the Army project, DSTA, and AIVA.

I owe a lot to my family, their prayers, blessings, and belief in me. Firstly, to my father, M. Yasin Ahmed, and my mother, Naheed Kausar, who strived to get their children the best education possible and never compromised on it. My parents have been rule breakers and dreamers in their community for giving their children the opportunities they have. The reason I am here today is because of how they never said no to any academic related activity I wanted to pursue and always encouraged me. I left my home when I was 17 for higher education and they never held me back from my adventures. Their trust, belief and support means the world to me. Secondly, to my siblings, Saqib, Kiran, Saba, and Jaweria who in their own right are inspiring individuals and have accomplished careers. Their journeys have inspired me to continue striving harder in my own. Their guidance, love, and support are what I seek whenever faced with a difficult decision.

Alhamdulillah, I have been fortunate in life, because of the people I have met, the opportunities that came my way and the support system I have. These are not things I had any direct control over and yet I have been blessed with them, for which I am grateful. I want to dedicate this thesis to those who hope and desire to make something of their lives, who constantly strive to break the norms, who are constantly learning and don't let their shortcomings stand in the way. That is an honorable journey, worth living this life.

Abstract

Speech carries information about the speaker's psychological traits, including their behavioral tendencies, leadership, emotions, and personality. The process of communication of these traits through speech can be thought of as a combination of an encoding process and a decoding process. The speaker, who expresses these traits, encodes them into low-dimensional characteristics of speech signals, and the listener, who perceives them, decodes the speech signal to make inferences about the state of the speaker. The encoding of psychological traits into speech is influenced by multiple factors, such as the context of the utterance, the speaker's personality traits, the environment, and more. It is well known that psychological traits are encoded in different aspects of speech, i.e. not only in 'what' is said but also in 'how' it is said. Thus while decoding, the listener must consider not only the linguistic content, but also the acoustic, prosodic, and other non-linguistic cues in the speech.

In this thesis, we develop computational models that attempt to emulate the human decoding process for two types of psychological traits: emotion and personality.

This thesis is accordingly divided into two parts. **In the first part**, we focus on **emotion**. We study how emotions are encoded and decoded in speech, and how they are affected by factors such as context, the naturalness of expression including real (spontaneous and involuntary) vs. enacted (prompted and voluntary), and other nuances of speech production. In our study of emotion and its representation in speech, we also take the lexical and phonotactic content of speech into consideration. A speech signal contains words and phonemes, each of which can be thought of as a stream or modality of information that exhibits unique characteristics that encode emotion, including intensity, cadence, rhythm, and more. We hypothesize that in order to decode emotion from speech, computational models must capture the characteristic variations within each modality. In this thesis, we objectively show how important this approach is for the computational analysis of emotion.

In devising better methodologies for emotion detection, we also focus on the intra-emotion range and absolute intensity (or degree) of emotional expression. Humans are able to decode emotions at very fine granularities. However, state-of-the-art techniques for automatic emotion detection (or decoding) work with predefined sets of discrete emotions, extended into a three-dimensional continuous space denoting the valence, arousal, and dominance of each discrete emotion. Any decoding technique that is learned from data is clearly restricted by the discretization of labels assigned to emotions by human annotators. In this thesis, we work on techniques that are agnostic to such restrictions. We hypothesize that for efficient decoding, it would be effective to utilize discrete and continuous information simultaneously, in a hierarchical framework.

Expanding this work, we propose a second approach, which is inspired by the fact that humans inherently use natural language to describe the emotions that they perceive. We contend that while the labels that humans assign to each emotion are restricted by the descriptors available in their language, the diversity of emotions can nevertheless be captured by the flexibility that natural language provides – namely by the affective language that is often casually used to describe an emotion. Such affective language can often have measurable acoustic correlates. For example: an angry man 'shouting loudly' is describing the emotion by directly referring to the loudness or intensity of the speech. We show that when the labeling is done with natural language descriptions, guided by acoustic properties of speech, the computational process of emotion decoding is significantly improved. In an extension of this work, we add a learnable 'prompt', with the hope that more textual information is incorporated automatically by the model, that helps the model towards better emotion decoding.

In the second part of this thesis, we focus on **personality**. We study how personality is encoded in speech. We specifically explore utterance-level voice signal characteristics such as pitch, loudness, and many others., and many relevant utterance-level voice quality features that have been observed to correlate with personality traits in scientific literature.

In order to better understand personality decoding, we revisit the widely accepted OCEAN traits. OCEAN forms the 5 bases across which every individual is rated, and is the result of the psycholexical hypothesis, which assumes that our usage of language and words to describe humans would reveal the underlying bases of the personality. We hypothesize that these bases, or in fact different bases, would reveal themselves when the same language and words are analyzed with newer techniques, i.e. word representations learned from large language models. In this work, we show how the newer techniques reveal the most informative number of bases as two, and the next most informative as five, which in fact are on average aligned with the OCEAN traits.

In summary, this work fills the gap in our current understanding of computational ways to process psychological traits from speech signals. The potential uses of such technologies in applications that involve human-computer interaction are many. Such technologies can also aid in the assessment and monitoring of mental illnesses and psychological problems in humans, helping all involved – healthcare providers and affected people – in positive ways. They can also help predict the long-term susceptibility of individuals to specific types of work, social situations, and other factors. Looking into the future, we believe this work is important in the inevitable rise of speech-based Artificial Intelligence systems, which will carry their own emotions and personalities and also understand those of their human users.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Human Voice as a Biometric

1.1.1 Information carried by Human Voice

The goal of this section is to give a broad overview of the human voice production mechanism. In order to analyze how speech is influenced by different psychological factors, it is important to understand how voice is produced, and what body parts are involved.

Voice is produced in the vocal tract. The vocal tract consists of numerous components including the trachea, vocal folds, larynx, glottis, and pharynx. The vocal folds are soft tissue structures contained within the larynx and act as the primary source of sound for vowels and as well as pressure controllers for many consonants. Speech sounds are produced when we inhale or exhale air; as air is expelled from the lungs and passes through the vocal folds, the vocal folds vibrate at about the rate of hundreds of vibrations per second. With these vibrations, the vocal folds convert the steady airflow from the lungs into a series of flow pulses by periodically opening and closing the air space between the vocal folds, providing the excitation of the vocal tract. The vocal tract including the oral and nasal cavities, the pharynx, and the larynx, then acts as a filter. This is the widely used linear system of voice production (source-filter model) where excitation signals from the vocal folds are modified by the vocal tract which acts as a filter.

Figure 1.1 shows the movement of the vocal folds; the first half of the complete cycle, where the vocal folds are initially closed and then they are opening to let the air out. The charts below the vocal folds represent the glottal area (the area of the vocal folds in contact with each other). The subsequent charts show the glottal flow (airflow from the vocal folds), and the derivative of the glottal flow; which is what is used in the source filter model to identify the glottal flow.

Given the voice production mechanism, we can identify innumerable factors that would influence the final signal that comes out of the mouth. The vocal folds are the source of the

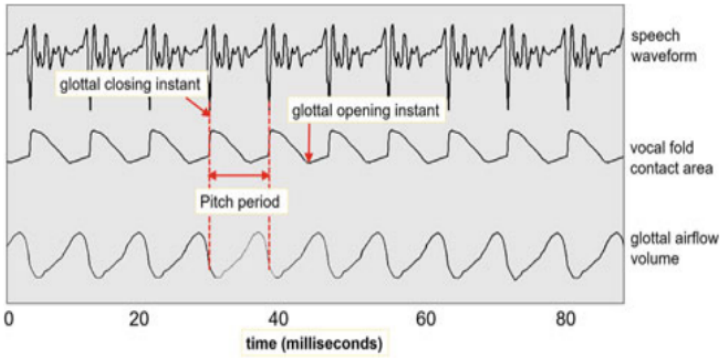


FIGURE 1.1: Top: Speech signal.
Middle: Contact between vocal folds as a function of time. When the.
Bottom: Glottal airflow volume as a function of time. When the vocal folds are closed, i.e. the vocal fold contact area is maximum, there is zero glottal airflow. When the vocal folds are open, i.e. the vocal fold contact is zero, there is maximum glottal airflow.

vibrations, and hence the voice; we can identify some of the factors that determine the workings of the vocal folds in an individual. Such factors include the vocal fold tension, vocal fold length, viscosity of the vocal folds, and the involved muscles' movements. All these factors in turn are dependent on many others for example vocal fold tension is affected by the age of the speaker (1) and the health of the speaker (2). The involved muscles' movements are affected by the age of the speaker (1), vocal fold length is affected by the gender of the speaker (3), and the viscosity of the vocal folds is affected by the health of the speaker (4). These are some among many others which are interconnected and affect the voice signal, tension, the mass of the vocal cords, and the sub-glottal pressure. It carries information regarding the prosody or rhythm, speaking style, accent, emotion, and personality among many other

We divide these factors into three different macro-factors, including (1) **bio-physical factors**, (2) **social and demographic factors**, and (3) **psychological factors**. Within each of these macro-factors, there are multiple other micro-factors. Figure 1.2 shows some of these micro-factors. We present this model as a multiple hub and spoke model in order to keep the diagram simple. The micro-factors are directly connected with each other as well for example, age under social and demographic factors directly affects the facial and body structure under the bio-physical factors. Instead, we connect all the micro-factors through genetics. Genetics connects all the macro-factors together and affects each of the micro-factors except for the social factors.

This thesis explores some of the micro-factors under the psychological factor that affect voice. Specifically, we aim to understand how emotion and personality affect voice. Let us look into details and background work that study how the voice production mechanism is affected by psychological traits.

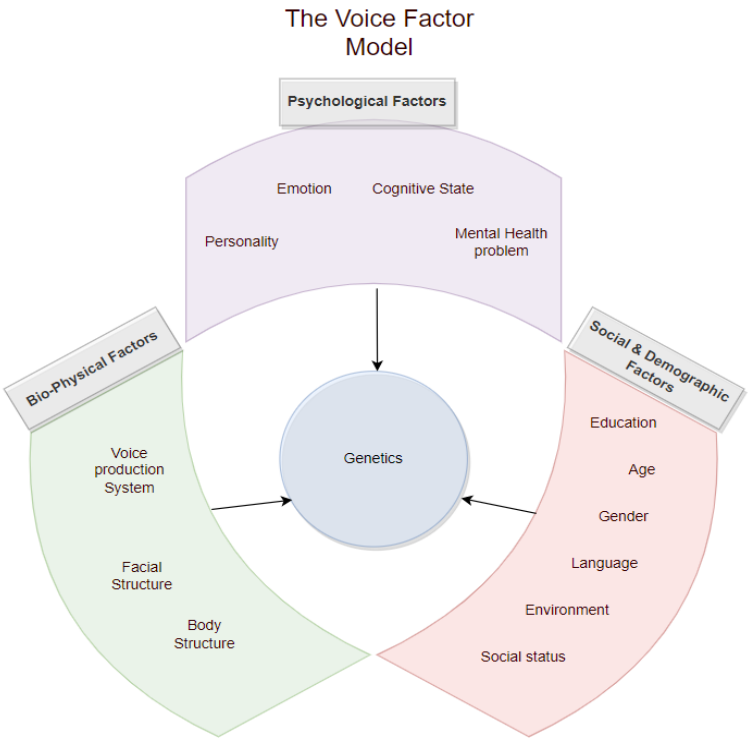


FIGURE 1.2: Multiple Hub and Spoke Model listing all the factors that affect the human voice.

1.1.1.1 Effect of Emotions on Speech

Each emotion individually has an effect on different components in the voice production mechanism, which ultimately show up in various aspects of speech. It has been found that the articulators act differently in emotions vs. neutral speech, i.e. (5) found that in emotional speech there are more peripheral or advanced tongue positions than in neutral speech articulation, specifically (6) found this for the vowel /i/. Parts of the vocal tract act differently under various emotions, i.e. for high arousal emotions like happiness and anger, (7) found that the vocal tract opening is greater compared to neutral speech. Studies like (5; 8) show that the vocal tract length tends to be shorter for happiness than for anger or sadness. (9) also finds that lip spreading and larynx elevation are important factors contributing to the decrease of the vocal tract length for happy speech. Furthermore, (10) found that lip spreading and laryngeal elevation often affect perceived emotional quality in the dominance dimension. (5) observed that the tongue tip, jaw, and lip positioning become more advanced when emotionally charged. In anger, the movement range of the jaw is larger compared to other emotions. (5) also found that angry speech was characterized by greater ranges of displacement and velocity, while it was the opposite for sad speech. Happy speech was comparable in articulation to neutral speech but showed the widest range of pitch variation.

Under different emotions, the different physiological changes have an impact on the voice being produced. For example, the shorter vocal tract length in happiness in comparison to that of anger or sadness will lead to higher pitch speech. Such acoustic and prosodic correlates of the

physiological changes in the body can be associated with all the changes listed earlier. These changes help humans express different emotions, and therefore, they ought to help distinguish between the different emotions as well.

1.1.1.2 Effects of Personality on Speech

Studies show that different personality traits have an impact on the various low-level speech features and are observable in voice quality, acoustic, and prosodic features. For example, extroverts speak with higher fundamental frequency (11), high loudness (11; 12; 13), higher speech rate (12), with higher frequency range than introverts (12; 14). Introversion is negatively related to loudness, low pitch, and resonance.

(15) found that there exists a positive association between submissiveness and rapid rate of speech. Dominance is negatively correlated with pitch and loudness (15) and is positively correlated with formant dispersion (16; 17) and not found correlated in others (18). In another study (12), it is also found that dominance is negatively correlated with intensity, intensity variation, speech rate, F0 variation, and sound silence ratio, and positively correlated with F0. Intensity is negatively correlated with energetic personality (12). Intensity variation is positively correlated with self-doubting personality (12) and negatively correlated with extraversion.

Observations have been made on various voice quality features that are perceived as certain personalities (19). For example, varying tempo and varying rhythm are indicative of high openness. Straight rhythm is indicative of low openness, and agreeableness. Many long pauses in speech are related to openness and neuroticism. Overall calm voice is associated with conscientiousness, low extroversion, and high agreeableness. Monotonous speech is associated with low agreeableness and high neuroticism. An extroverted voice is described as livelier.

Apart from the speech features, lexical usage also gives clues about personality traits. For example, agreeableness was negatively related to the use of swear words and positively correlated with the use of first-person singular pronouns (I, me, my) (20) Talking was considered an indicator of extroversion, cursing an expression of a lack of agreeableness, and laughing a sign of emotional stability. (20)

1.1.1.3 Connecting Emotion and Personality

Emotion and personality are two views of the human psychological state, but not independent views. Personalities are linked with certain kinds of emotions and the expression of emotion varies depending on the individual's personality. Many previous research studies have linked the two. Neuroticism is likened to high emotionality and is related to anxiety, depression, tension, and emotional characteristics, and high neuroticism people tend to be emotionally unstable, worried, or highly reactive to environmental stimuli (21). Individuals with higher extroversion are more energetic and social, characterized by active emotion coping styles,

more positive affect, and less anxiety, which possibly renders them with less negative feelings (22). Highly agreeable people are less likely to demonstrate high emotion (23). Extroverts are typically characterized as talkative, bold, and assertive (24). Those high in neuroticism tend to be discontented, emotional, and angry (24). These individuals do not adapt well to change in the workplace and tend to get emotional during conflicts (25).

The relationship between emotion and personality types leads to the belief that emotion expression varies based on the personality of the individual, and therefore this relationship can be exploited when computationally modeling emotions.

1.2 About This Thesis

Speech carries an immense amount of information about the speaker, including the speaker's psychological traits, i.e. emotion, and personality. These psychological traits are very complex and any computational analysis of speech for these traits poses a lot of challenges. We tackle some of the challenges for emotion and personality analysis from speech. For speech emotion analysis, some of the challenges arise from the manner in which the context affects the expression of the trait. Such context includes whether the emotion is expressed voluntarily or if it is acted out. In addition, there are multiple aspects of speech in which these traits are expressed and in order for any computational analysis, all these aspects need to be analyzed in conjunction with each other. Further, representations of emotion in ways that can be computationally processed are quite limited and therefore cause hindrances in computational analysis of emotions. We tackle these listed challenges, individually, in each chapter under part one of this thesis.

In part two of this thesis, we tackle some of the challenges in personality analysis from speech. One of the biggest challenges comes from data scarcity and the the noisy labels for personality. Instead of proposing data intensive modeling approaches, we present a more knowledge based approach. We study the low-level speech features and voice quality features that are correlated with the different personalities from prior scientific literature. We extract OCEAN traits using the proposed formulae and show how they perform better than data based methods. Furthermore, representation of personalities have focused heavily on the Big-5 OCEAN traits. These five traits are a result of psycholexical studies, done in the early 1900's. Using these same bases in today's computational methods proves to be challenging. We delve deeper into these listed challenges, individually, in each chapter under part two of this thesis.

This thesis aims to computationalize the analysis of the psychological traits from speech by explicitly taking into consideration the above-mentioned challenges.

1.3 Challenges Addressed

Below I elaborate on the challenges in the computationally decoding of the psychological traits from speech:

- What are Emotions and Personality and how can they be computationally represented: In the psychological literature, there is no single, definitive way of describing these psychological traits (26). This causes significant challenges when performing computational decoding of the traits. Computationally representing the traits requires quantifying them in some way. However, in the absence of a holistic understanding of these psychological traits, this process becomes suboptimal. We address this challenge by proposing methodologies to combine already existing representation strategies for emotions. Furthermore, we propose new ways of representing emotions using natural language descriptors, and delve deeper into understanding the personality representations that are widely accepted.
- Contextual subjectivity: The expression of emotion and personality is impacted by the context in which they are expressed. Taking the example of emotions: the way an emotion is acted out is quite different from when it is naturally expressed. This is because when naturally experienced, the traits cause physiological changes in the speaker, which ultimately manifest in the speech. In contrast, when emotions are acted, these physiological changes are absent, leading to differences in vocal characteristics. We quantify the differences between acted and natural expression of emotion.
- Multi-modality: Emotion and personality can be expressed through various modalities, including facial expression, body language, gestures, and vocal expression. Taking the example of emotions, sometimes they are better conveyed through facial movements than through speech. For instance, an individual's facial expression may convey anger even when the tone of voice is monotonous. Therefore, for a holistic understanding of emotion, all sources of information may need to be considered. This makes the task an even more challenging since oftentimes, the available modalities are limited. To tackle this challenge, we use the speech signal, and extract other modalities from the signal like the word sequence (transcript) and the phoneme sequence. We show that these modalities provide additional information beyond just the speech when modeled efficiently.
- Lack of labeled datasets for emotion and personality. One of the widely used benchmark datasets for emotion is IEMOCAP (27), which is 12 hours of data. Compared to other speech tasks, such as ASR, for which datasets contain thousands of hours of data, progress in ASR has steadily increased as a function of dataset sie. Similarly, for personality, there are very few datasets for speech personality analysis. For instance, one of the widely used datasets, SSP-Net (28), has only 1.7 hours of data, makes it difficult to make generalizable models. To address this challenge, we propose knowledge-intensive approaches that do not rely on labeled data. We also show how the approach performs on personality prediction task.

Part I

Chapter 2

Prior Work

2.1 Theoretical Emotion Models

'Emotions' have garnered a lot of interest from scientists from various fields, including neuroscience, psychology, philosophy, social and behavioral sciences, biology, and computer science. Emotions are described as reactions to events deemed relevant to the needs, goals, or concerns of an individual. Emotions encompass physiological, affective, behavioral, and cognitive components.

Throughout history, psychologists have been troubled with the definition of emotions and what exactly they are. There have been a number of monumental works to describe emotions as early as Darwin's *The Expression of the Emotions in Man and Animals*, published in 1872 (29). Many theoretical models of emotions have been developed henceforth, a summary of which can be found here (30).

Scholars traditionally define emotions as consisting of a set number of basic emotions and others that build over these basic ones. Basic emotions are defined as those emotions that fulfill the following requirements: (1) they are also exhibited by other animals, (2) they have a specific, innately determined biological basis in brain organization, (3) they develop very early in life, (4) they are irreducible, not composed of 2 or more simpler emotions, (5), they have distinctive neuro-muscular expressive pattern manifested in facial expression (29). Building on this definition, different scholars have come up with different lists of basic emotions. Therefore even the notion of what are basic emotions has not been agreed upon by scholars. In 1962, Tomkins (31) lists down **8** basic emotions including fear, anger, anguish, joy, surprise, interest, and shame. In 1980, Plutchik (32) also said that emotions can be boiled down to 8 basic ones, but different from Tomkin's list. Plutchik's list includes fear, anger, sorrow, joy, disgust, surprise, acceptance, and anticipation. In 1992, Paul Ekman (33) defined basic emotion as those emotions that have distinct facial emotions and said that there are only **6** basic emotions including fear, anger, sadness, happiness, disgust, and surprise.

Other scholars like Schlosbery (34) said that instead of basic emotions, emotions should be explained on continuous-valued dimensions. Initially, these dimensions only included valence and arousal; whereby valence means the positive or negative nature of the emotions, i.e. the sentiment, and arousal means the energy exerted in the expression of the emotion, e.g. happy is lower arousal than ecstatic. Later a third dimension was also introduced, i.e. dominance, which represents the degree of control over a social situation

The lack of knowledge on what is emotion, and how to quantify and represent them are hindrances for computationally processing emotions. However, this hasn't limited researchers to continuing research on emotions, specifically speech emotion research. Below we present some of the background work in the field of speech emotion.

2.2 Computational Emotion Models

The task of computationally detecting emotion from small speech samples, which often have expressions of only one emotion, has been a popular task. The performance of the task has been shown to depend heavily on the kind of features that are extracted from speech. We will only list speech-based features, but there is a plethora of work on other modalities as well like visual, facial, and textual features. Some works directly use pitch (35; 36; 37), energy (35; 38; 39), ZCR (zero crossing rate) (39), spectral features like MFCC, LPCC (39; 40; 41; 42; 43; 38; 44), prosodic features (37; 45).

Earlier works have used a variety of models for computationally detecting emotions including models like SVM (39; 46; 47; 40; 48; 41), GMMs (36; 49; 50), HMM (51; 52; 53), MLPs (54; 55; 37). A whole of research on the topic has been done using Deep Neural Networks, specifically using models like CNN (56; 57; 58; 59; 60), recurrent neural networks like LSTM, RNN (61; 62), encoder-decode style models (63; 64; 65), models taking context into consideration (66). There have been comprehensive reviews done on the background work on speech emotion, some of the recent ones are (67; 68).

Below we list prior work directly related to the work we have completed and also highlight the missing science.

2.2.1 Natural vs Acted Emotion

Differences in emotion expression under natural and acted emotion have been studied before, with contradicting findings. Studies like (69), which analyze acted and natural emotional speech with the help of human listeners conclude that the listeners are not able to distinguish between the two categories. The problem is also studied in the domain of false expression, where human listeners are asked to identify the truthfulness of the speech. It also reaches the conclusion that humans are less likely to differentiate between the two. On the other hand, studies like (70), which also use human listeners, conclude that about 78% of listeners were able to differentiate

between the natural and acted emotion with only audio clues, and even more could differentiate when provided with audio-visual cues. Studies have observed that acted and natural expression of speech innately differ based on voice quality (71) and on prosodic properties (72). Acted speech is considered to be delivered in a more emotionally intense fashion and acted speech affects the vocal expression in a more general way, without the nuances of the changes caused by the natural emotion (73).

Challenges Remaining: The above studies primarily analyze the effect of acted emotion on the listener, however, a listener is not always a valid discriminator between the two types of emotional expression. We hypothesize that instead the factors that comprise natural and acted emotion differ significantly at a more basic level which a listener may not be able to identify and perceive. Therefore, a computational framework is needed to study the innate differences between the two types of expression, which can lead to a more comprehensive understanding of the differences between the two different expressions of emotions.

2.2.2 Cadence in Emotion

Emotion expression is manifested in multiple modalities, which include the speech signal, the spoken words, and the phoneme sequence in addition to the facial expression and video modality. Several works have explored using computational models to combine multiple modalities for the task of emotion recognition; the most recent ones being transformer architecture (74; 75). Vocal expression of emotions has a rhythm unique to each emotion, and unique to each modality as well. Rhythm in speech refers to a number of quantities like the time taken for speaking a unit of sound, the intensity of the unit of sound, the voice quality metrics for the unit of sound, etc. Psychological literature has identified the uniqueness of rhythm in each emotion (76; 77), and how important it is for the expression of emotional speech.

Challenges Remaining Although speech cadence is important for the task of emotion expression. The quantities that refer to cadence, like the time, intensity, pitch, voice quality per unit sound, etc, are difficult to quantify and use in the automatic analysis of speech emotion. Therefore, no computational way has been proposed so far to capture and utilize cadence for emotion expression. We propose a methodology using deep neural models like transformers and show the importance of the cadence of different modalities for speech emotions.

2.2.3 Representing Emotions: Unifying Discrete and Continuous Views

Automatic methods of emotion detection from speech require emotions to be represented in a way that can be computationally processed. Traditionally, emotions have been represented as a discrete set, consisting of a handful of emotion classes or they have been represented as a continuous vector in multiple dimensions (78) comprising of Valence (V), Arousal (A), and Dominance (D) (from Russell (79)). Translating between these two forms of emotion representations has been of interest (80; 81; 82; 83), both with audio and textual input. Though

there appear to be conflicting views (81; 83), many works have established different extents of correlations between the continuous and discrete emotions- like anger has a lower valence that is high in arousal (84; 85).

Challenges Remaining Traditionally, for the task of speech emotion recognition, researchers have used either discrete (86) or continuous emotions (87) for the automatic detection of emotions. Some have exploited the relationship between discrete and continuous emotion representations, by combining them together, however, it has not been explored in depth. We propose a model that uses both these representations simultaneously and exhibits the usefulness and the relationship between the two representations.

2.2.4 Representing Emotions: Natural Language Acoustic Prompts

Emotions are very complex and require an equally complex representation strategy. In the previous section, we looked at how the currently used emotion representations could be utilized together for the automatic detection of the task. In the current scenario, we want to expand to new ways of representing emotions. We take inspiration from computer vision, where there has been work using natural language descriptions of images in a deep learning framework. For example, CLIP (88) is trained on image-text pairs where these pairs are collected from the internet. The learned model associates image captions (textual descriptions) with the images. Applying the learned model to various downstream tasks like zero-shot image classification and text retrieval, results in SoTA performance. A similar approach has recently been applied to the audio-text pairs in CLAP (89). This model is trained on audio captioning datasets where descriptions for each audio are already present. AudioClip (90) is an extension of CLIP, where three modalities are used, audio in addition to image and text. This model is trained on AudioSet (91) (contains images, audio, and text) where the labels of the audio are used as the textual descriptions. UrbanSound8k (92) and ESC50 (93) are used to train the audio and text models. SimEmotion (94) performs image emotion recognition, in a similar fashion to CLIP. The textual descriptions for the images are made by extracting emotion and entity-level information. The entities in the images are recognized using the Detectron2 library and labelled emotion for the images are expanded using the Plutchik chart.

Challenges Remaining Although textual descriptions have been used in relation to image tasks and audio tasks, they have not yet been explored for speech tasks like speech emotion. The challenge is how to form the textual descriptions of the speech, whether to use the content of the speech or to use the acoustic, prosodic qualities of speech. In our work, we hypothesize that acoustic properties of speech carry information about emotions and would be helpful in learning more fine-grained emotion representations. Therefore, we use the acoustic properties of speech to form textual descriptions.

Chapter 3

Natural vs Enacted Emotions

3.1 Introduction

Can vocal emotions be emulated? This question has led to long-standing debates in the speech community regarding *natural* versus *acted* emotions, in the context of emotion classification and emotion categorization tasks. To conduct any speech-based emotion research, an important factor is the nature of the speech samples or the vocal stimuli, and whether those samples are representative of natural emotions. Natural emotions can best be defined as emotions that are spontaneous and involuntary. Acted emotions, on the other hand, are prompted and voluntary. Because acted emotions are volitional, researchers argue that the physiological and psychological responses that natural emotions induce are absent from acted emotions (95; 96). Nevertheless, research on emotion perception uses acted emotions as convenient proxies for natural emotions. While many past studies have focused on presenting perception tests for natural versus acted emotions with mixed conclusions (96; 97; 70), there is a lack of an at-scale systematic framework to study the differences and similarities in those classes.

The distinction is, perhaps, best illustrated by the study, "The President's Speech", narrated by well-known neurologist Oliver Sacks (98), in which two types of patients, *aphasic*, and *tonal agnostic*, both find a presidential candidate's posturing on TV to be screamingly implausible, although normal viewers have no problem with it. *Aphasic* patients are highly sensitive to expression and tone, but cannot interpret the words. On the other hand, *tonal agnostic* patients lack any sense of expression and tone, and pay attention to exactness of words and word use to capture the emotion. The skilled actor, in this case, the presidential candidate, attempts to convey feelings and emotions through a combination of affect that is given a pass by normal viewers. Neither of these patients can grasp the totality of the natural emotion. However, both types of patients who, unlike normal people, only perceive some of these factors, find them sufficiently implausible as to cause them distress. Simply by their inability to consider the totality of affect that normal people can perceive, the patients cannot be lied to or deceived.

3.2 Proposed Theoretical NAO Model

In order to develop a systematic framework to study the differences and similarities in acted and natural emotions, we must recognize that there are, in fact *three entities* to be considered. The communication of vocal emotions is, at its essence, a combination of an encoding and a decoding process. The subject *expressing* the emotion encodes their emotional state into the low-dimensional speech signal. The subject *perceiving* the signal decodes it to make inferences about the state of the speaker. We will distinguish between two types of encoders: the *non-actor* who actually experiences the emotion, and the *actor* who may not. In all cases, the decoder is an *observer*, whose only cue in terms of vocal emotions is the vocal stimuli. Based on these, we propose the *non-actor, actor, and observer (NAO)* model (Figure **??**), which represents all three entities and the relation between them. The actor aims to encode synthetic emotion in a manner that the observer cannot distinguish from the genuine emotion encoded by the non-actor. This enables us to formulate a hypothesis that can be formally tested – that there nevertheless remain identifiable fundamental differences in the encoded signals in the two cases. If the test fails, that would mean natural emotions *can* be emulated, and that acted emotions can be used as proxies for natural emotions. If the test passes, however, that would signal towards dichotomy between acted and natural emotions, leading to a low validity and value in using acted stimuli. We note that the non-actor and the actor differ in their encoding of emotions. Because natural emotions are, for the most part, involuntary, they include physiological and psychological responses as concomitants, such as heart rate, breathing rate, muscle tension, and mood. These physiological changes manifest in the voice by changing the spectro-temporal structure of individual sounds (99; 100). For example, (101; 102) argue that vowels and consonants produced in *fear* are often more precisely articulated than they are in *neutral* situations. The physiological changes along with the psychological factors also define the choice of phonemes (i.e. the lexical content) and prosodic cues (103; 104). As an example, words of aggressive nature are more likely to be used by an individual in an aggressive mood (105).

The encoding of emotion is hence an aggregate, or *perceptual sum* of these acoustic, phonetic, and linguistic influences, or *factors*. The observer decodes the perceptual sum of these factors to make an inference about the emotional state of the speaker. In emulating an emotion, the actor attempts to produce a somewhat similar *aggregate* of these factors as the non-actor, *i.e.* a combination of factors that he expects the observer to decode into a near identical perceptual sum. If the actor succeeds, he conveys the target emotion to the observer. We hypothesize that in doing so, the *individual factors* that the actor produces will, however, still be incorrect or even implausible, even though the perceptual sum may be plausible.

Vocal (or indeed any) expression of emotion is, of course, a complex phenomenon, and the complete set of acoustic, phonetic, linguistic and prosodic factors used in expressing it is still not fully understood. To test our hypothesis, we must nevertheless identify one or more of these factors that can be statistically quantified. As mentioned earlier, physiological and psychological changes concomitant with emotion are known to affect the *choice of phonemes* and their *manner*

of delivery. We will refer to these as the *phonetic bases of vocal emotions*. By our hypothesis, there will be a statistically measurable difference in these between the actor and non-actor.

To verify our hypothesis, we require a mechanism to quantifiably extract these bases from the speech signal. To do so, we train a neural network model for emotion classification tasks on two datasets, one of natural speech and the other of acted emotional speech, using an *attention* mechanism. The attention aims to identify the most important phonemes in an utterance in order to classify its emotion. We compare the statistical patterns of the most *attented* phonemes across actor and non-actor. As we will see in the final sections of our paper, these factors do indeed differ in a statistically significant manner, bringing the validity of conclusions drawn from acted emotional speech as a proxy for natural emotion into question.

3.3 Experimental Validation

3.3.1 Dataset

We use two types of datasets; acted and natural. We run our experiments for only four emotions: angry, happy, sad and neutral.

3.3.2 Acted Data

The acted dataset used in the experiment is IEMOCAP (27). It consists of ten sessions, each of which is a conversation between two actors. The conversations are divided into labeled sentences. We implement a 10-fold cross-validation training setup. In each fold, data from 9 speakers is used for training the model, and data from 1 speaker is used for testing. The data consists of 1103 angry, 1636 happy, 1708 neutral and 1084 sad utterances. The average duration of utterances in this dataset is 4 seconds.

3.3.3 Natural Data

For natural speech, we used the CMU-SER data (106). This dataset has been collected from NPR podcasts (107), and television programs hosted by the Internet Archive (108). The dataset is annotated using the Amazon Mechanical Turk (109). It has 6000 utterances in the training set and 2571 utterances in the test set, with a total of 1099 angry, 3028 happy, 1262 neutral, and 611 sad utterances. The average duration of utterances in this dataset is 5 seconds. Further details of the CMU-SER dataset can be found in (106).

3.3.4 Alleviating speaker-dependent bias

Because we compare acted versus natural emotions based on the two datasets with differences in speakers, it is possible for our phonemic content and, hence, phoneme distributions and the attended phonemes to be influenced by the word choices of different speakers. To ensure that our analyses only reflect the differences in the emotional content rather than the differences in speakers, we eliminate speaker dependencies at the time of training our model. Because the natural dataset is collected from a diverse set of online sources, it is reasonable to assume that there are fewer cases of a speaker represented more than once in the data. On the other hand, the acted dataset consists of 10 speakers only. Hence, we perform leave-one-out cross-validation to alleviate speaker dependencies in the results. These steps ensure that our models and analyses are robust to the difference of speakers.

3.3.5 Characterizing content-dependent bias

It is possible for our analyses to be influenced by the differences in the content of the two datasets. To ensure that the difference in content is not a confounding factor, we study the phoneme distributions of the two datasets. Figure 3.1 presents the phoneme distributions of both datasets. To determine if the difference between the distributions is statistically significant, we run a *Wilcoxon rank test* (110). The Wilcoxon rank test is a non-parametric test, used to compare two related samples. In this case, the null hypothesis $\mathcal{H}_0$ is that there is no difference in the distributions of the phonemes under the two datasets, and the alternative hypothesis $\mathcal{H}_1$ is that there *is* a difference between the distributions of the two datasets. We obtain a p-value of .3, therefore with $\alpha = .05$, we fail to reject the $\mathcal{H}_0$. This ensures that the difference in the phonemic content of the two datasets is unlikely to affect the distribution of the attended-phonemes.

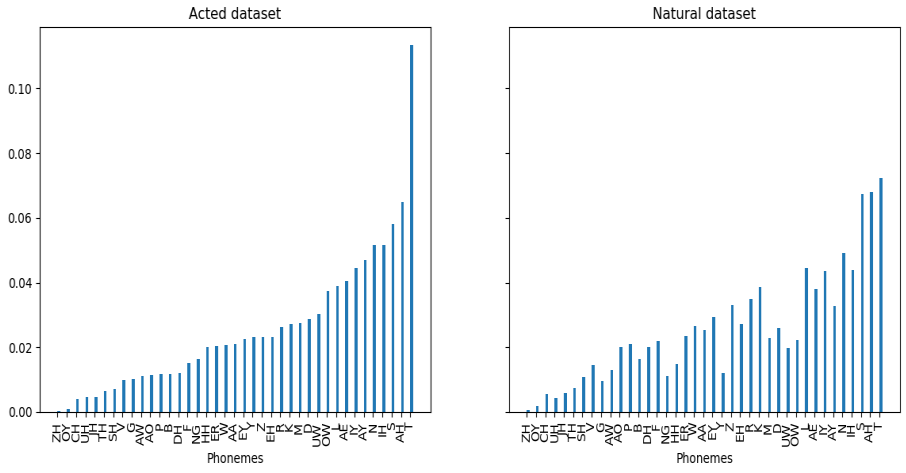


FIGURE 3.1: Total phoneme distributions under both natural and acted dataset

3.3.6 Neural Model

In order to extract the phonetic bases of vocal emotion, we propose a neural network model. We design the model to take into consideration both the lexical and acoustic aspects of the utterance and also the relationship between the two, to capture the phonetic bases of emotion. The linguistics should guide the model about the important parts of the acoustic. To create a vector representation of the linguistic part of the input, we pass it through an LSTM which captures the contextual information of the linguistics. This forms a context-sensitive lexical vector.

To capture the relationship between the two modalities, we utilize an attention-based mechanism. This enables the context-sensitive lexical vector to put attention on some parts of the audio, forming importance weights. The weights, when applied back to the input audio, make the output high in parts that the lexical vector points to and others become low in value. A feature vector is created from this weighted output, which thereafter goes into the classification layer for emotion. Training the model maximizes the classification accuracy, but in doing so, it teaches the model to create feature vectors which would be differentiable for the emotion classes. This, in turn, optimizes the attention mechanism, thereby allowing the lexical vector to focus only on those parts of the audio which would lead to the highest classification accuracy. This lays the basis of the model we have used, as shown in Figure 3.2.

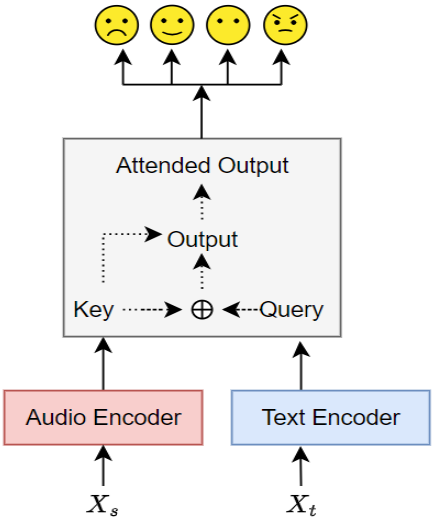


FIGURE 3.2: Neural network model used for the emotion classification with attention mechanism

Since we need both the acoustic and lexical content of an utterance, to train the model, we require the transcription of the recording. To get the transcription, the recording is passed through ASR. We used Google API (111) to extract the transcription. Each word in the transcript is represented as a BERT (112) contextualized word embedding. These embeddings are passed through an LSTM layer (shown in blue). For the attention layer we represent the keys as the output of the convolutional layer; a 3-dimensional output. The query is the last hidden state of the LSTM passed through a linear projection.

The network is optimized using the Cross-Entropy loss, with weights for individual labels due to the class imbalance in the datasets. The ASR based transcript of the utterance is segmented into different phonemes using an HMM based phoneme segmentor (113). Once the model is trained, the attended-phoneme outputs are inspected.

3.4 Result

We base our results on the output of the attention mechanism from the neural model described earlier. Since the model is trained for emotion classification over four emotions, it is useful to note the classification accuracies achieved by the model over the two datasets. On the acted data, we achieve a classification accuracy of 72% and on the natural data, we achieve a classification accuracy of 52.4%.

To perform analysis on the attended phonemes, for each emotion we aggregate the phonemes with the highest attention output. We normalize their frequencies by the total frequency of the phoneme in the data. Figures 3.3 and 3.4 show the distributions of these attended-phonemes for acted and natural conditions (from the corresponding datasets) respectively, for each emotion. We note several differences between the two distributions. The frequency of fricatives and stops is higher in natural speech than in acted speech.

We also observe that the frequency of vowels is higher in acted speech than in natural speech. Specifically, the phonemes /AA/, /B/ and /IH/ occur more frequently in acted speech. Moreover, an overall higher percentage of nasal phonemes occurs in natural speech.

We also study the attended-phoneme distribution under different test subsets created for the 10-fold cross-validation procedure to ensure consistency of the attended-phoneme distribution within the dataset. Variations of the phoneme frequency from the 10 different cross-validation results are shown in the box plots in figure 3.5 for both datasets. It can be observed that the results have lower variation in natural speech than in acted speech. In general, the same variation trends hold for other phonemes as well.

The box plots also illustrate the difference in the frequencies of each phoneme in the two datasets. In particular, we observe the frequencies of the vowels like /IY/, nasal phonemes /M/, /N/, stop phonemes /T/, and fricatives like /DH/ (figure 3.5) to be different among the two datasets. To calculate the significance of these differences for all emotions, we run a standard t-test. We find a **statistically significant result for all four emotions** $(p < .05)$.

Therefore, in the context of natural versus acted classes, our analysis concludes that there are significant differences between the phonetic bases of the two classes. Consequently, we conclude moderate to low validity and value in using acted emotions as proxies for natural emotions, suggesting that researchers should be wary of arriving at conclusions about natural emotions using acted emotion datasets.

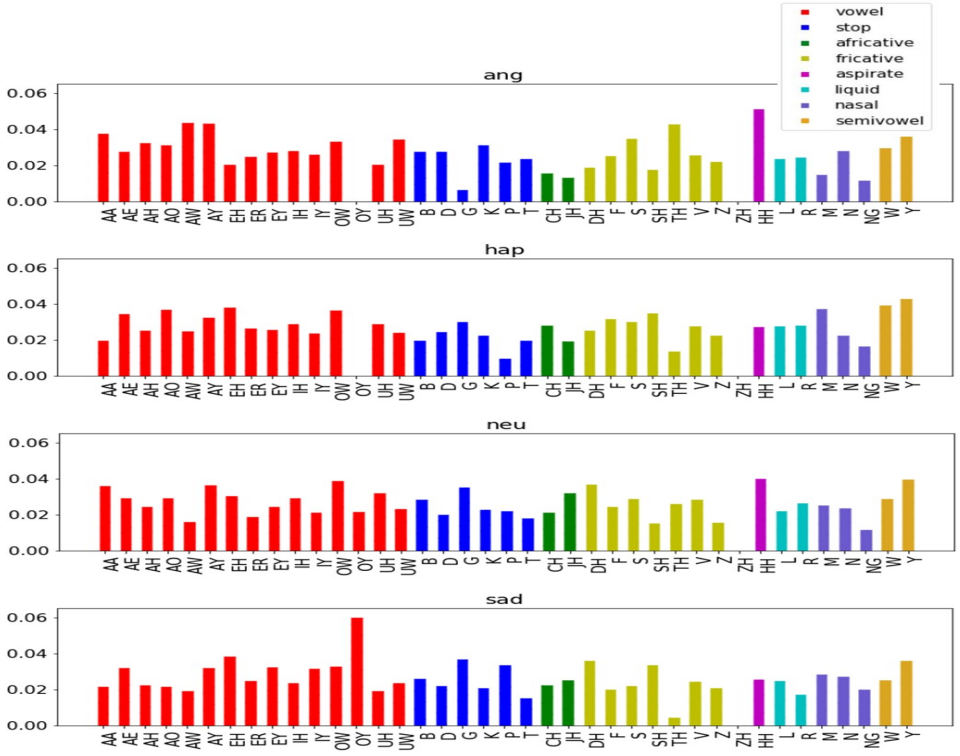


FIGURE 3.3: Distribution of attended-phonemes (and phonetic groupings) across four emotion classes on acted data

We would like to note that this study has only inspected English language data. However, the framework provided can easily be applied to any other language. We leave the investigation of the phonetic correlates of emotion in other languages, and its comparison with the conclusions provided in this study, as a possible future work.

One limitation of this study is the lack of the same set of observers across the acted and natural datasets. While we have no control over this within our analysis, given the diversity of observers for the two datasets, we expect little statistical observer bias. However, this remains to be verified by future studies.

3.5 Conclusion

In this section, we study how the emotion expression differs when it is acted out vs when it is naturally expressed. We perform this study by focusing on the phonetic bases of the vocal expression. Phonetic bases of emotion comprise 'what' phonemes are used and the 'manner' they are delivered in to express the emotion.

To run a quantifiable test, we model the task as an attention-based emotion classification problem. The attention mechanism aims to capture the "attended" phonemes in order to get the correct classification. We then calculate the distribution of these attended phonemes and examine how their distribution varies between natural versus acted emotions. We observe several differences,

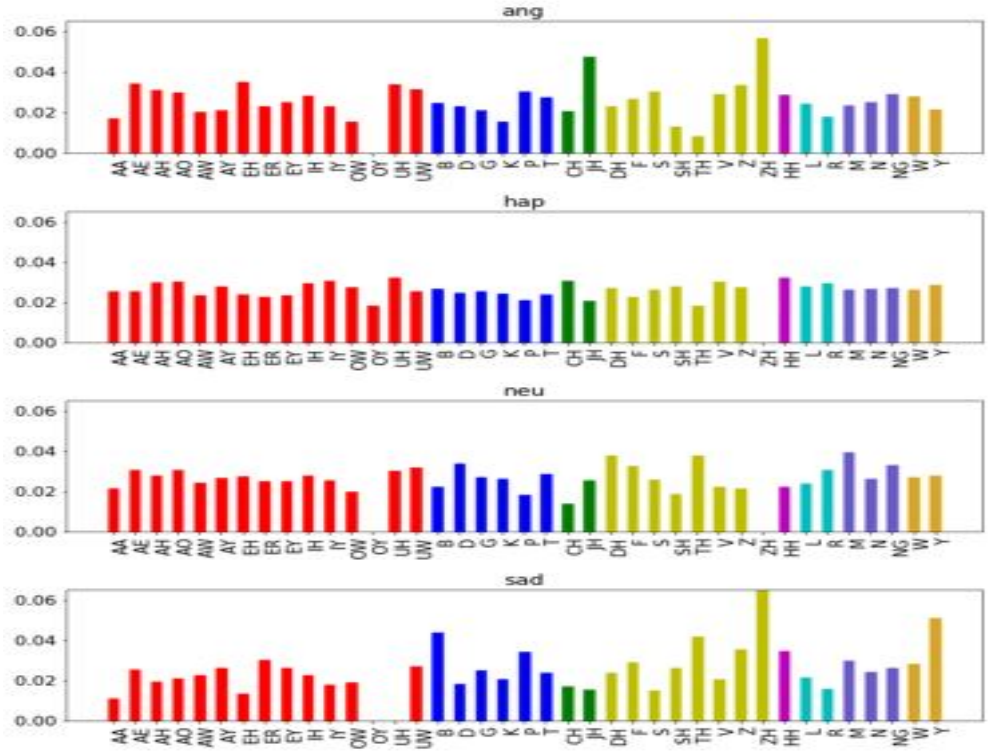


FIGURE 3.4: Distribution of attended-phonemes (and phonetic groupings) across four emotion classes on natural data

for example, a higher occurrence of fricatives and stops in natural speech than in acted speech. We obtain statistically significant differences in the attended-phoneme distribution among natural and acted emotions. Therefore, our hypothesis stands true. The differences in phonetic bases signal a dichotomy between natural and acted emotions. This study has applications in speech emotion recognition, emotional speech synthesis, and human-computer interaction.

Moving forward, we want to experiment with the bases of vocal expression from other modalities, in addition to the phonemes. In the next chapter, we will focus on the different modalities that carry emotion expression. These modalities include the speech signal, the phoneme, and the word sequences. We specifically try to model the speech cadence in each modality in addition to other features important for speech emotion.

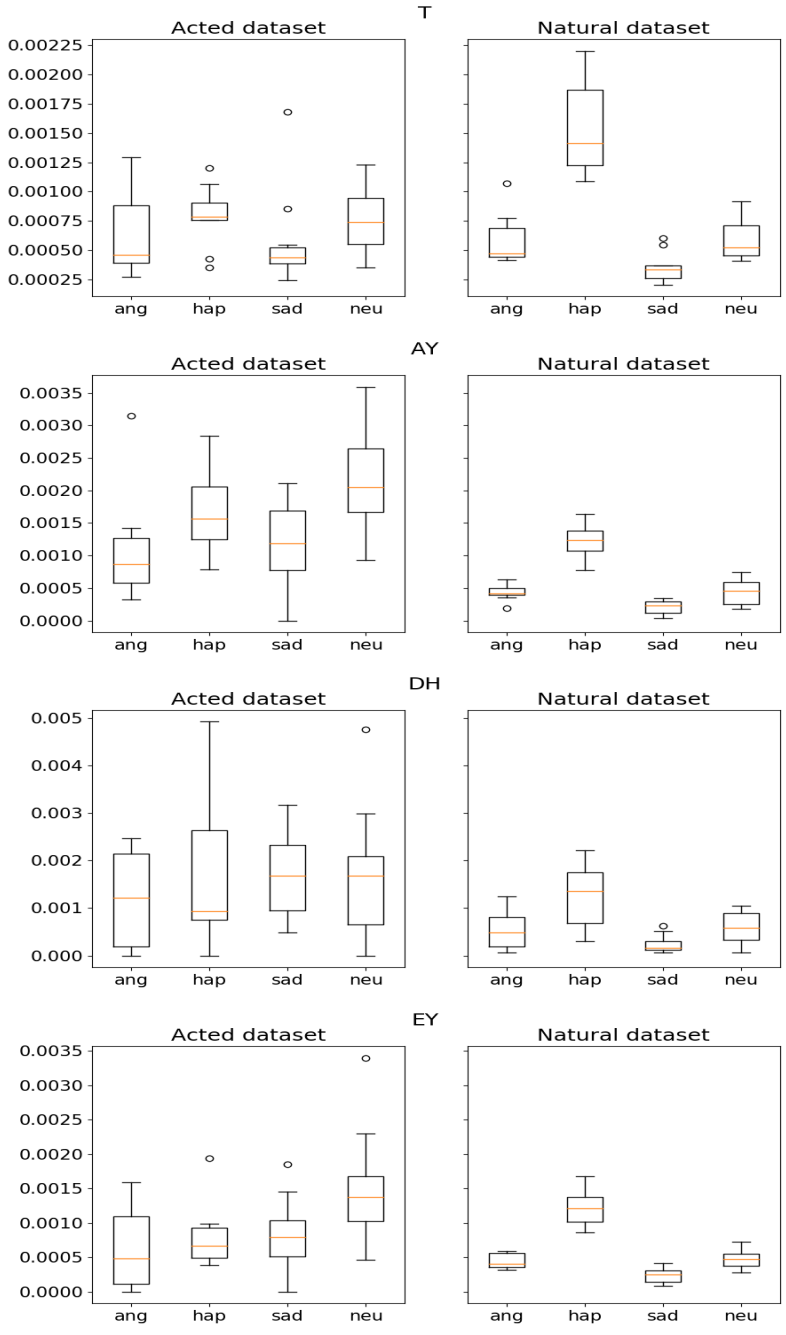


FIGURE 3.5: Box plot of attented phoneme /T/, /AY/, /DH/, /EY/ in natural versus acted dataset. The figure highlights the frequency difference among the two datasets. Note the median values for the box plots are very different for both datasets for a given emotion.

Chapter 4

Cadence in Emotion

In the previous chapter, we looked at how acted and natural emotions differ from each other, especially 'what' phonemes were used in each emotion and 'how' they were expressed. We extend this analysis to other modalities, including phonemes, words, and the speech signal itself. We will focus on how these modalities carry the emotion, which can help in the automatic decoding of the emotion expression.

4.1 Introduction

Systems that perform automatic detection of emotion have been shown to perform better when using multiple modalities than when using a single one (114). Using multiple modalities requires an understanding of how speech enunciation, word usage, etc changes under any emotional state. These may involve changes in the acoustic properties, phonetic changes, prosodic changes, changes in the usage of words and their delivery, etc.

In the case of phonetic changes, for example, in anger, phonemes are more distinctly articulated than they are in other emotions (115) and there are more distinct opening and closing movements for vowels (116). In addition to this, the choice of phonemes is also affected by the emotion, for example, pleasant emotion is correlated with higher use of more approximate consonants, more back vowels, and more tense vowels (105). Similarly, in the textual modality, previous studies have shown that the choice of words is dependent on the expressed emotion. For example (117) shows there is a significant effect on word choice under high emotional valence, e.g. the general use of pronouns, like the word "I" varies when expressing a negative versus a positive emotion.

We note here that changes at these finer levels in any modality are not independent of each other. For example, higher usage of first-person pronouns in anger changes the phonetic and acoustic properties of the signal as well. Also, we note that for each modality, individual change rates are different. For example, the phenomenon of coarticulation in phonemes affects a window of neighboring phonemes – thus the prosodic changes that happen for a specific emotion state

affect the nearby phonemes as well. In a similar way, changes in a word affected by an emotional state cause modifications in other words in the sentence, and such resultant changes can span multiple adjacent words. Formant movements in the audio signal vary at different rates as well. Therefore each of these modalities has its own natural rate, i.e. its natural 'cadence'. The cadences of the modalities are related, as explained earlier, but they are also distinct and local for each modality. Individually their cadence is important for emotion, because, as explained earlier, the choice of phonemes and words is related to the emotional state. Thus modelling the intra-modality cadence as well as the local cadence separately is important for the task of emotion detection.

In this chapter, we explore the best strategy to capture the natural cadence in each modality: whether their cadences must be individually modeled, or if they must all follow the natural cadence of "time" as captured in the audio signal; whether it is better to align the modalities with the audio signal in either case or if it is better for the modalities to remain unaligned. We also explore *how* the individual modality cadence is best captured. For this task, we use the transformer architecture (118). Since the model does not have a sequence-2-sequence (seq2seq) component, it cannot account for the order in the input stream. To capture the concept of order in the input stream, it uses 'positional encoding'.

4.2 Proposed Model

To explore the questions put forth, we introduce the notion of *'cadence retaining alignment'* (CRA) which is modeled by first adding the positional encoding for each input modality separately and then aligning the streams. In this case, each modality has a separate positional encoding. Secondly, we introduce the notion of *'cadence normalizing alignment'* (CNA) where the phoneme and word sequence ids are first stretched out to align with the audio features, and then a positional encoding is used. In this case, there is one shared positional encoding between the modalities. Thirdly, we introduce 'unaligned local cadence' (ULC) where the modalities are not aligned and have their individual positional encodings to capture the local cadence of the modalities.

In order to test the hypothesis we use the IEMOCAP dataset (27) and the CMU-MOSI dataset (119). Both datasets consist of aligned phonemes and words for each utterance.

In order to test our hypothesis, we need to perform analysis on the aligned and unaligned modalities with and without individual PEs. In our case, audio features act as the base for aligning the other two modalities. The phoneme ids and words are aligned based on the extracted HuBert features from the audio.

Given an audio signal A, let $\mathcal{H}_A = [H_1, H_2, \cdots, H_{N_A}]$ represent the sequence of *audio* feature vectors derived from it, where N_A is the length of the vector sequence and $H_i \in \mathbb{R}^{d_A \times 1}$, where d_A is the dimensionality of the individual feature vectors. As is the norm, the vectors in $\mathcal{H}$

represent features derived from uniformly spaced windows of the signal A and correspond to ticks of time.

Let $\mathcal{P}_A = [P_1, P_2, \cdots, P_{N_P}]$ represent embeddings of the sequence of phonemes in A, where N_P represents the number of phonemes in the recording, and $P_i \in \mathbb{R}^{d_P \times 1}$ is actually a d_P dimensional *embedding* of the phoneme that is learned. Let $b_1, b_2, \cdots, b_{N_P}$ represent time *boundaries* of phonemes in $\mathcal{H}$.

An *alignment* $\mathcal{P}_A^*$ of $\mathcal{P}$ to $\mathcal{H}$ is obtained by repeating each P_i b_i times:

$$\begin{aligned} \mathcal{P}_A^* &= [P_1, P_1, \cdots, P_1(b_1\text{times}), P_2, P_2, \cdots, P_2(b_2\text{times}), \\ &\quad \cdots, P_{N_P}, P_{N_P}, \cdots, P_{N_P}(b_{N_P}\text{times})] \end{aligned} \tag{4.1}$$

Note that $\mathcal{P}^*$ and $\mathcal{H}$ are now identical in length. Here, and below, the "*" superscript in $\mathcal{P}_A^*$ is used to indicate an alignment of $\mathcal{P}_A$ to $\mathcal{H}_A$.

Similarly, let $\mathcal{B}_A = [B_1, B_2, \cdots, B_{N_T}]$ represent the sequence of ($\mathbb{R}^{d_T \times 1}$ embeddings of) *words* in A and $q_t, q_2, \cdots, q_{N_T}$ be their boundaries in $\mathcal{H}_A$. In our work we have used the Bert model (112) to obtain the embeddings B_i. We can define $\mathcal{B}_A^*$, the alignment of $\mathcal{B}_A$ to $\mathcal{H}_A$ similarly to Eqn. 4.1.

We can now define our three models.

CRA: The CRA model locally positionally encodes the individual streams, aligns and concatenates them and uses them as features to the classifier. There are several proposed ways of using PE, but in our work we simply add the PE into the encoded inputs. We define the *locally positionally encoded* version of $\mathcal{H}_A$ as $\hat{\mathcal{H}}_A = \mathcal{H}_A + PE_{N_A}^{d_A}$. We can similarly define $\hat{\mathcal{P}}_A = \mathcal{P}_A + PE_{N_P}^{d_P}$ and $\hat{\mathcal{B}}_A = \mathcal{B}_A + PE_{N_T}^{d_T}$.

The *aligned* versions of $\hat{\mathcal{P}}_A^*$ and $\hat{\mathcal{B}}_A^*$ of $\hat{\mathcal{P}}_A$ and $\hat{\mathcal{B}}_A$ are obtained by aligning $\hat{\mathcal{P}}_A$ and $\hat{\mathcal{B}}_A$ respectively to $\mathcal{H}_A$ as in Equation 4.1. The final feature sequence $\mathcal{F}_{CRA} \in \mathbb{R}^{(d_A+d_P+d_T)\times N_A}$ for classification is obtained by vertical concatenation: $\mathcal{F}_{CRA} = \hat{\mathcal{H}}_A \odot \hat{\mathcal{P}}_A^* \odot \hat{\mathcal{B}}_A^*$, where $\odot$ represents the concatenation operator.

The final classification is performed using a softmax layer on $\Omega(\mathcal{F}_{CRA})$, i.e. the fixed-dimensional output of a Conformer encoder (120). $\Omega(\cdot)$ that operates on $\mathcal{F}_{CRA}$.

CNA: The CNA model concatenates aligned feature vector sequences before positionally encoding them. Thus $\mathcal{F}_{CNA} = \mathcal{H}_{CNA} + PE_{N_A}^{(d_A+d_P+d_T)}$, where $\mathcal{H}_{CNA}$ is the vertical concatenation of $\mathcal{H}_A, \mathcal{P}_A^*$ and $\mathcal{B}_A^*$. The final classification is performed using a softmax layer on $\Omega(\mathcal{F}_{CNA})$.

ULC: The ULC model independently processes $\mathcal{H}_A$, $\hat{P}_A$ and $\hat{T}_A$, each with their individual PE and concatenates their final outputs for classification. Thus, classification is performed by applying a softmax layer on $\Omega(\mathcal{H}_A) \odot \Omega(\hat{P}_A) \odot \Omega(\hat{T}_A)$.

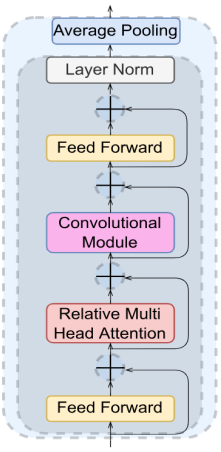


FIGURE 4.1: Conformer Encoder, with average pooling used in our experiments. In the later images, this model is denoted by Ω

After having described the basic three architectures that we want to experiment with, we devised six different models, shown in figure 4.2. Model A uses only the audio modality (devised for comparison purposes). It applies a softmax classifier directly to $\Omega(\mathcal{H}_A)$. In our experiments, the audio features $\mathcal{H}_A$ were derived from a HuBert model (121). Model B is the CRA model. Model C is the CNA model. Model D is the ULC model.

Models E and F are *double-fusion* models, where the individual streams are both concatenated to form a joint feature and indvidually processed. Thus model E is obtained by applying a softmax classifier to $\Omega(\mathcal{F}_{CNA}) \odot \Omega(\hat{P}_A) \odot \Omega(\hat{T}_A)$. Model F applies a softmax classifier to $\Omega(\mathcal{F}_{CRA}) \odot \Omega(\hat{P}_A) \odot \Omega(\hat{T}_A)$.

4.3 Experimental Validation

4.3.1 Dataest

We use IEMOCAP dataset, which is acted conversational speech, consisting of 10 speakers, 5 male and 5 female. The data consists of Anger, Happy, Sad, Neutral, Excitement, Frustration, Fear, Surprise, and others. Following prior work (122) however, since the dataset is not balanced we only use the most frequent four classes including Happy, Sad, Anger, and Excitement. Each emotion class per speaker can be seen in Table 4.1.

We also perform experiments on CMU-MOSI dataset (119), which is a collection of speakers reviewing products on YouTube. The data is labelled on a sentiment scale of -3 to 3, going from negative to position emotions. In our experiments we divide the classes into two, positive (≤ 0) and negative (> 0), following prior work (114). We use the standard train, test split published with the dataset. Table 4.3 shows the number of utterances in training and testing. Both of the datasets already contain the forced aligned phonemes and words.

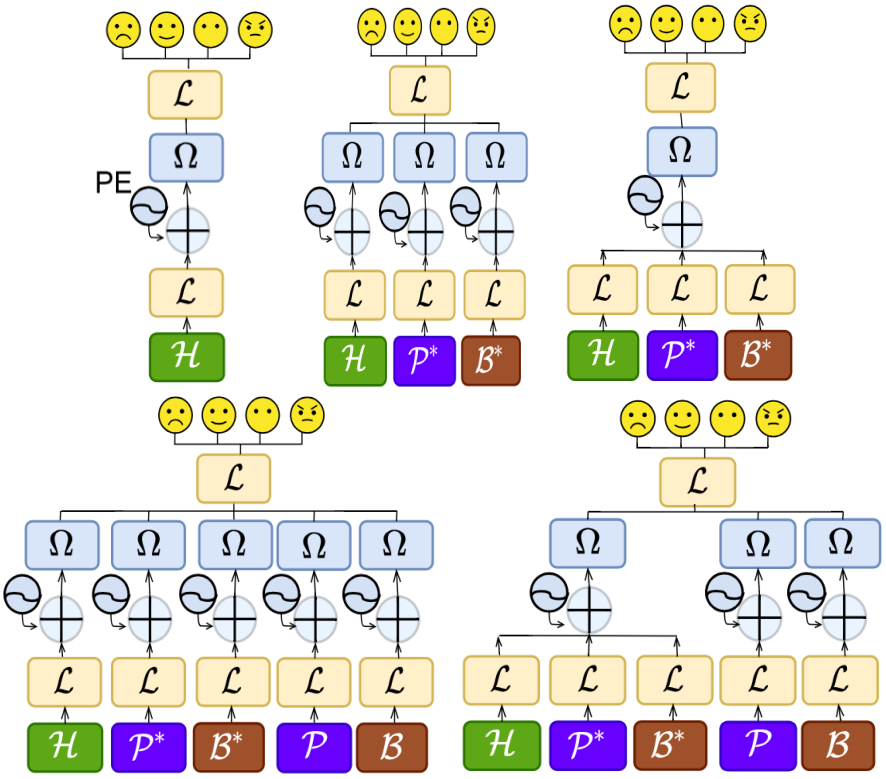


FIGURE 4.2: The 6 models used in our experiments (One is not shown). In each figure, Ω is the conformer encoder shown in Fig. 4.1, $\mathcal{L}$ is a linear layer, $\mathcal{H}$ represents HuBert features, $\mathcal{P}^*$ represents phoneme ids aligned with $\mathcal{H}$, $\mathcal{B}^*$ represents Bert embeddings aligned with $\mathcal{H}$, $\mathcal{P}$ are unaligned phoneme ids and $\mathcal{B}$ are unaligned bert embeddings. Top Left: Model A, Top Right: Model B, Middle left: Model C, Middle right: Model E, Bottom: Model F. Model D (which is not shown) has the same figure as model B except that the inputs are unaligned versions.

TABLE 4.1: Number of emotion utterances per speaker for each of the four emotion classes in IEMOCAP

	1	2	3	4	5	6	7	8	9	10
ang	92	147	78	82	148	67	122	70	205	92
hap	66	69	77	66	55	70	34	47	31	80
sad	113	78	132	116	133	113	81	84	62	172
neu	163	171	221	213	190	135	182	227	76	130

TABLE 4.2: Number of positive and negative emotion utterances in CMU-MOSI in train and test split

	train	test
pos	676	401
neg	602	513

4.3.2 Model

4.3.2.1 Transformers

Transformer architecture, originally proposed (118) for the task of machine translation (sequence generation task), has successfully been used for classification tasks aswell. It is an encoder, decoder style architecture. In the encoder module, it forgoes the use of seq2seq models (generic way for encoding sequential inputs for modelling the sequential information in an input), instead it utilizes feed-forward layers with activation, or convolutional layers (conformer encoders (123)). It makes use of multi-head attention mechanisms for attending to the entire input streams, bypassing the problem of modelling long-term dependencies that general seq2seq models face. Furthermore, to add the order information in the input the model uses positional encodings (explained in next section). Figure 4.1 shows the encoder architecture of the model, proposed in the original paper, which we use in this work.

4.3.2.2 Positional Encoding

Positional encodings are suppose to represent the absolute or relative positions of the input streams. We use the originally proposed sinusoidal positional encoding, which allows the model to attend by the relative positions of the inputs. The sinusoidal PE are defined as the sequence of M vectors: $PE_M^D = [PE_D(0), PE_D(1), \cdots, PE_D(M-1)]$. where the indivdual vectors $PE_D(k) = [PE_{(k,0)}, PE_{(k,1)}, \cdots, PE_{(k,D-1)}]^\top$ are composed as

$$\text{PE}_{(k,2i)} = sin(k/10000^{2i/D})$$
$$\text{PE}_{(k,2i+1)} = cos(k/10000^{2i/D})$$

where D is the dimension of the encoded inputs and also the dimension of the PE.

4.4 Result

Table 4.3: Accuracy for each model

	A	B	C	D	E	F
IEMOCAP	64.7	**68**	62.1	62.2	67.8	66.7
CMU-MOSI	67.1	**68.9**	61.4	63.2	67.4	66.7

4.4.1 Separate Positional Encoding vs Shared

Comparing the CRA and CNA strategies: separate PE for each modality has shown to work much better in comparison to using only one shared PE. The difference can be seen while

comparing model B and model C. Separate PEs lead to a 5.9% improvement in IEMOCAP and 7.5% improvement in CMU-MOSI dataset. This difference shows that the positional encodings are important at capturing the local cadence changes, i.e change rates in the audio, text and phoneme identities separately.

4.4.2 Aligned vs unaligned inputs

Comparing the CRA and ULC strategies: aligned input modalities have shown to perform better than unaligned ones. This can be seen while comparing model B and D. Aligned streams lead to a 5.8% improvement in IEMOCAP and 5.7% in CMU-MOSI. Aligned inputs are better at capturing the changes over the input modalities and also the correlated changes in between the modalities. However aligning the inputs are expensive. For word and phoneme, forced alignment needs to be used, which we note that if the model is being used in computationally low resource settings, would not be feasible to perform.

4.4.3 Fusion vs none

We have found that fusion models, like model E and F, which use the aligned or unaligned versions of the modalities twice do not perform as well as best strategy CRA, i.e model B where aligned inputs are used. Model E uses both the aligned and unaligned inputs with separate PEs, however it can be speculated that with the addition of the unaligned inputs, the performance degrades in comparison to model B. Model F uses both aligned and unaligned inputs, however the aligned inputs have only one PE. It is exciting to see that the performance difference between model E and F is not that big in both the datasets.

4.4.4 Single modality vs multi modality

Having multi-modalities is important for the task, as shown in many earlier works as well. Comparing model A and other models, it can be observed that using only audio modality, model A, the performance does not improve drastically when other modalities are brought into the picture. We attribute this to the choice of HuBert features, since these features are rich representations of the audio, and not only audio but include linguist and phonetic information in the representations, so in a sense using only HuBert feature does not qualify for 'only audio modality'. When the model A is trained using directly on spectrograms or MFCC, the performance is quite low, and the performance difference between single and multi-modalities can be observed there (we however do not include those results).

4.4.5 Importance of aligned inputs

From the results, it can be observed that models which use aligned inputs but only a shared PE and models which use unaligned inputs but individual PEs perform almost similarly. This can be seen while comparing model C and D and also comparing E and F. In model C where inputs are aligned but the PE is shared performs almost similarly to the model D where unaligned inputs are used with individual PEs, 62.1% versus 62.2% in IEMOCAP and 61.4% versus 63.2% in CMU-MOSI respectively. Comparing models E and F, it can be observed that the difference is 1.1% in IEMOCAP and 0.7% in CMU-MOSI. This shows that the benefit that aligning the inputs together gives can be fully achieved by having individual PEs for each modality even when they are unaligned. The performance degradation is small enough for the cost of aligning the inputs.

4.5 Conclusion

Combining multi-modalities has been shown to be important for the task of emotion recognition. However, combining different modalities together effectively is an open problem. Specific to this task, there are numerous studies explaining how each of the modalities, e.g. speech signal, phoneme sequence, and word sequence change individually and also in conjunction with each other, i.e. have local cadence and inter-modality dependent cadence. Hence, it is important to model the individual cadence of each modality, and not assume a common global clock. We propose a simple framework for modeling this, by assigning a positional encoding specific to each modality, which allows the transformer model to attend to the input streams based on the positions of the elements in the streams. From our experiments, we show that local cadence is highly important for the task; the performance improves by at least 5.9% compared to when the modalities share only one PE.

Furthermore, we show that the performance improvement that aligning the input modalities with each other gives (with a shared positional encoding) can be equally achieved by using unaligned input modalities but each with their own individual positional encoding.

Going forward, we change the focus on representing strategies for emotion. The decoding capabilities of emotion rest on how the emotions are represented, and limiting representational abilities hinder the decoding capabilities. In the following chapters, we focus on how current emotion representation could be better utilized and further how can they be enhanced.

Chapter 5

Representing Emotions: Unifying Discrete and Continuous Views

In the past two chapters, we focused on how the encoding of emotions differs and how the phonetic sequence carries this information and then extended this analysis to form better computational decoding methodologies that utilize multiple modalities in addition to the phoneme sequence. In the coming two chapters, we will focus on how computational decoding could be improved by utilizing better representations of emotions. In the current chapter, we will combine the two widely used emotion representations, i.e. discrete and continuous representations.

5.1 Introduction

Emotion is a complex entity that can be represented in different ways. When defined as a motor response to stimuli, it can be viewed as belonging to a discrete set; whereas when considered as a subjective feeling, it can be expressed as a continuous vector in multiple dimensions (78). Speech Emotion Recognition (SER), the task of identifying the emotional state of the speaker using speech as input, can be used to predict both discrete (86) or continuous emotions (87). The most commonly used continuous representation for this task comprises three attributes – Valence (V), Arousal (A), and Dominance(D) (from Russell (79)). Among these three dimensions, Valence represents the pleasantness of the emotion, Arousal denotes the intensity of it, and Dominance represents the degree of control over a social situation. Similarly, a number of discrete emotions have been identified – happy, sad, angry, etc. Though the discrete and continuous views of emotion appear to be different, they have shared and complementary information. Psychological models like Plutchik's wheel of emotions (32) or Russell's model space (84) represent discrete emotions on and within a circle on a continuous "arousal-valence" plane, demonstrating that the values of these continuous variables are correlated to the discrete emotion.

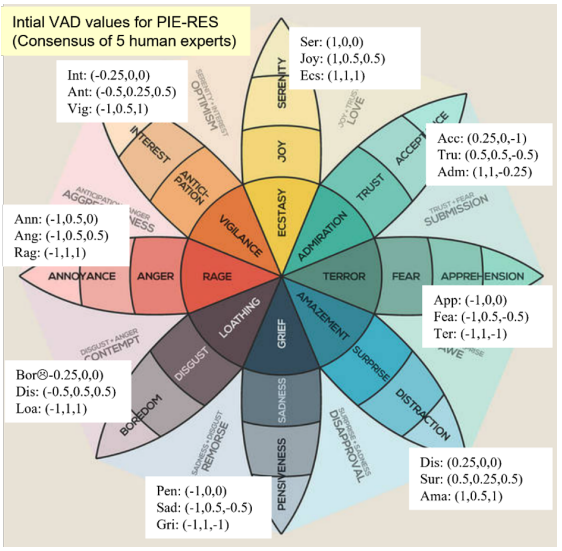


FIGURE 5.1: Plutchik wheel showing the discrete emotions, and some initial estimates of what V,A,D values could be for each emotion in the wheel.

Translating between these two forms of emotion representations has been of interest (80; 81; 82; 83), both with audio and textual input. Though there appear to be conflicting views (81; 83), many works have established different extents of correlations between the continuous and discrete emotions- like anger has a lower valence that is high in arousal (84; 85). Therefore in this chapter, we propose to examine such dependency relationships.

5.2 Motivation

Plutchik circumplex of emotions displays emotions as a flower. Each layer of the flower contains 8 emotions. The most inner layer is the basic emotions. There is some topological relationship of every discrete emotions in this wheel. For example, grief is right opposite ecstasy and loathing is opposite admiration, and so on. Any leaf in the flower either represents positive emotions - positive valence, or negative emotions - negative valence. As one goes out of the leaf, the intensity of the emotion decreases, so the arousal is lower. Therefore, given this circumplex, we can estimate some values for each V,A and D for the emotions. Figure 5.1 shows the Plutchik wheel and the initial estimates of what the V,A,D values could be for each emotion. These estimates are agreed upon by 5 experts in the lab. By plotting the initial estimates on the circumplex, we get the plots for valence, arousal and dominance as shown in the figure 5.2.

However these estimates of VAD are noisy and since they are only a consensus of few experts, we can improve upon these estimates. To improve upon them, we can imagine the Plutchik wheel as a circle, where each emotion has a radius and an angle from the center. We assign these two values as representative features for each emotion. To model these values, we form a simple linear function between the position of the discrete emotion and VAD. Position of the discrete emotion refers to the position of the emotion in the circumplex (hence denoted by its radius and angle in the circumplex). We form another linear layer which predicts VAD values from the

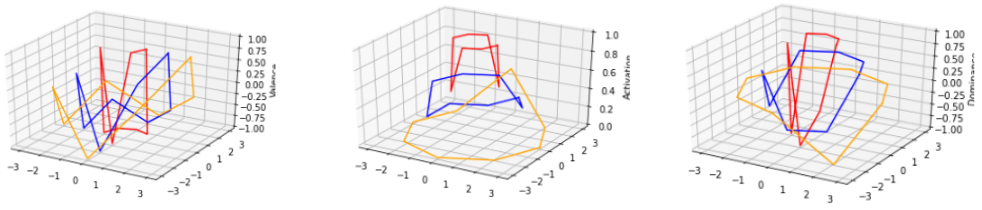


FIGURE 5.2: Plotting initial estimates of V,A,D on the circumplex using the radius and angle of the emotions from the Plutchik wheel. Left is valence, middle is arousal and right is dominance.

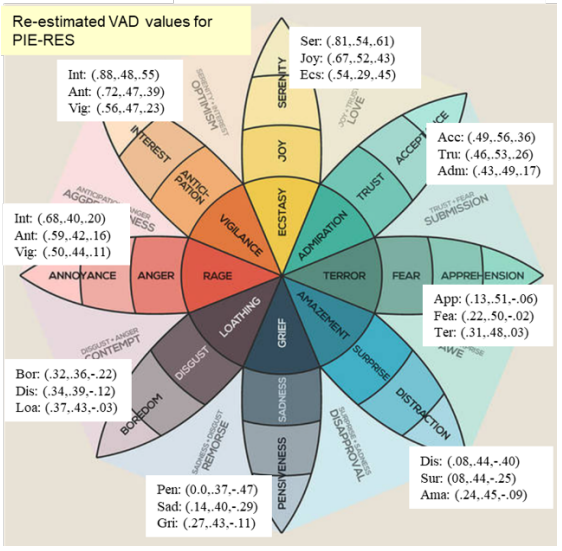


FIGURE 5.3: Plutchik wheel showing the discrete emotions, and learned estimates of what V,A,D values for each emotion in the wheel.

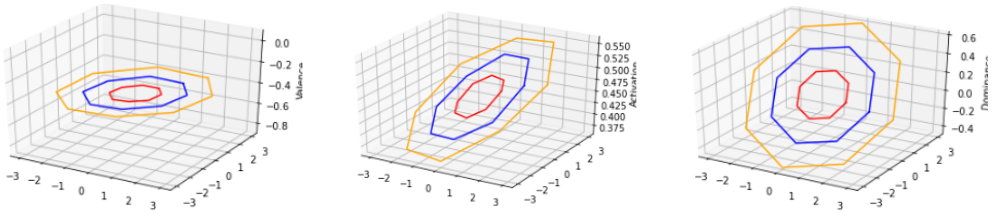


FIGURE 5.4: Plotting learned of V,A,D on the circumplex using the learned radius and angle of the emotions. Left is valence, middle is arousal and right is dominance.

position of the emotion. We repeat this prediction process until convergence of both VAD and the angle-radius. Figure 5.3 shows the final values of the VAD for each emotion. Furthermore, plotting the emotions again on the circumplex by using their updated angle and radius learned values, see figure 5.4. Focus on the activation and dominance plots and we observe that there is a better estimate of activation and dominance, covering a greater range in the space. A good observation is that this re-estimate encompasses the topology of the initial Plutchik wheel and introduces valence, activation and dominance into them. What this shows is that there is a better way of learning the relationship between the two emotion representations of discrete and continuous, but needs a more complex model than a linear layer.

5.3 Proposed Model

Holistic models for automatic SER must be able to capture both generic and specific notions of emotion contained in the continuous and discrete emotion labels. Prior work has investigated multi-task architectures for jointly predicting continuous and discrete emotion (124; 125; 87). However, such models do not adequately utilize the relationship between discrete and dimensional emotions.

We propose a hierarchical model structure that better utilizes the dependency of the discrete and continuous representations. Below, we first, explain the baseline architecture and models and then we explain the hierarchical model architectures.

5.3.1 Encoder Decoder Architecture

To perform SER on input speech, we use a speech encoder and pooling decoder. The encoder takes as input a sequence of N content-based speech features $X = [x_1, x_2, \cdots x_N]$, and produces as its output a sequence of hidden representations $H = [h_1, h_2, ...h_M]$. These hidden representations are fed to a decoder that predicts either continuous emotion $\hat{c}$, discrete emotion $\hat{y}$, or both.

The decoder comprises a temporal self-attentive pooling layer, followed by a Multi-layer Perceptron (MLP) that extracts task-specific embeddings, E_D for discrete prediction and E_C for continuous prediction. These embeddings are then passed through the final classification layer that maps to 3 dimensions for continuous prediction (corresponding to V,A,D), or 5 dimensions for discrete prediction (corresponding to the number of discrete emotion classes). The continuous and discrete baseline models use a similar neural network architecture, except for the number of output neurons.

Each of the models we describe below, baseline models that perform either continuous or discrete prediction, and multi-task or hierarchical multi-task models that predict both discrete and continuous emotion, share the same encoder structure but with slightly different decoder architectures.

5.3.2 Baseline Models

Baseline Discrete Model: To *independently* predict only the *discrete* attributes from input X, the model termed as **Baseline D** uses a decoder with self-attentive pooling followed by a Multi-Layer Perceptron which generates the embedding E_D. They are then fed into the classification layer that produces the discrete label $\hat{y}$ as its output. The model is shown in Fig. 5.6 (top left).

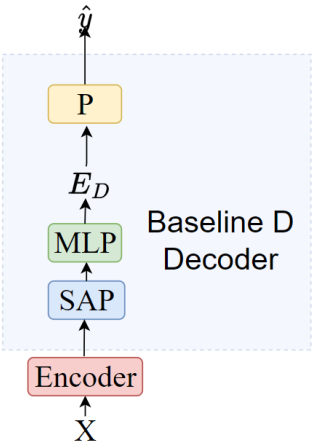


FIGURE 5.5: Baseline Discrete Model. The Baseline C model is similar (not shown) with intermediate E_C and final continuous prediction $\hat{c}$.

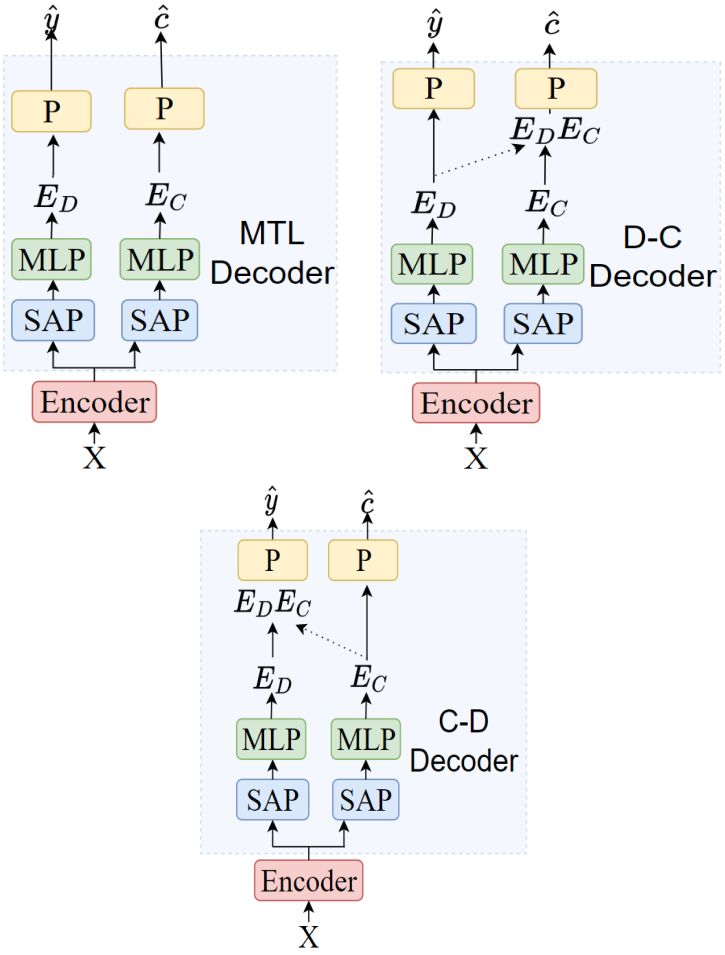


FIGURE 5.6: The top left shows the Multi-task model (MTL). The top right and bottom figures show the Hierarchical multi-task model D-C and C-D respectively. SAP stands for self-attention pooling. MLP stands for multi-layer perceptron. $\hat{y}$ is the discrete label prediction whereas $\hat{c}$ is the 3-dimensional continuous emotion prediction.

Given the true discrete label y, the model is optimized using the multi-class cross-entropy criterion. For a dataset with A utterances, the cross-entropy loss is computed as:

$$\mathcal{L}_{\text{disc}} = -\sum_{i=1}^{A}\sum_{c=1}^{K} y_{i,c}\log(\hat{y}_{i,c}) \tag{5.1}$$

Baseline Continuous Model: To predict the continuous attribute labels *independently* from the input speech, the encoder-decoder model, called **Baseline C** uses a continuous emotion pooling layer and MLP to predict the continuous emotion $\hat{c} = [\hat{c}_{Val}, \hat{c}_{Aro}, \hat{c}_{Dom}]$.

The variables $\hat{c}_{Val}$, $\hat{c}_{Aro}$, and $\hat{c}_{Dom}$ correspond to the valence, arousal and dominance predictions respectively.

This model is optimized using the Concordance Correlation Coefficient (CCC) Loss. Given the prediction $\hat{c}$ and the ground truth c, the CCC loss is defined as shown in Equation 5.3.

$$\text{CCC} = \frac{2s_{c\hat{c}}}{s_c^2 + s_{\hat{c}}^2 + (\bar{c} - \bar{\hat{c}})^2} \tag{5.2}$$

$$\mathcal{L}_{\text{ccc}} = 1 - \text{CCC} \tag{5.3}$$

where $s_{c\hat{c}}, s_c^2, s_{\hat{c}}^2, \bar{c}, \bar{\hat{c}}$ represent the covariance between the ground truth and prediction, variance of the ground truth, variance of the prediction, mean of the groundtruth and mean of the prediction respectively.

Multi-Task Model (MTL): Since continuous and discrete representations carry information about the same utterance, we are interested in understanding if *jointly* predicting continuous and discrete emotion attributes of the utterance improves model performance. To do this, we use a multi-task architecture, called **Multi-task**, to predict both the discrete and continuous emotion attributes simultaneously. We use a shared speech encoder that transforms the input speech into hidden representation H, capturing shared information between discrete and continuous emotion attributes.

This model predicts both discrete label $\hat{y}$ and continuous labels $\hat{c} = [\hat{c}_{Val}, \hat{c}_{Aro}, \hat{c}_{Dom}]$ for every utterance. The model uses a shared encoder. The decoder consists of two parallel branches that are used to learn task-specific pooling and MLP parameters. The discrete branch generates a discrete embedding E_D, which is then used to predict the discrete emotion $\hat{y}$. Similarly, the continuous branch predicts a continuous embedding E_C, which is used to predict the continuous emotion $\hat{c}$.

The model is optimized with the total loss as shown in Equation (5.4).

$$\mathcal{L}_{\text{total}} = \alpha(\mathcal{L}_{\text{ccc}}{}^{Val} + \mathcal{L}_{\text{ccc}}{}^{Dom} + \mathcal{L}_{\text{ccc}}{}^{Aro}) + \beta\mathcal{L}_{\text{disc}} \tag{5.4}$$

where the values of α and β are set to 1 based on hyperparameter search.

5.3.3 Hierarchical Multi-Task Models (HMTL)

Multi-task modeling described in the previous section seeks to utilize the shared information between discrete and continuous attributes to predict them jointly. However, it doesn't assume any direct dependency between them. Based on the hypothesis that knowledge of discrete emotion attributes would help improve continuous attribute predictions and vice-versa, we develop hierarchical multi-task models.

In this model, the continuous and discrete emotion prediction branches in the decoder are used to generate the respective embeddings E_C and E_D, as in the multi-task formulation. However, here, the predicted continuous emotion, i.e., $\hat{c}$ is not computed solely based on E_C, but also on E_D. In other words, the discrete emotion embedding is used as an auxiliary input to predict the continuous emotion. We term this **HMTL-DC** model. Similarly, the **HMTL-CD** model is where the continuous emotion embedding is used as an auxiliary input to predict the discrete emotion.

In **HMTL-DC** model, discrete emotion embedding E_D, is concatenated with the continuous embedding E_C. This concatenated embedding $[E_D E_C]$ is then used to predict the continuous emotion, $\hat{c}$. Similarly, in the **HMTL-CD** model E_C is concatenated with E_D and used to predict $\hat{y}$. The models are optimized with the total loss as shown in Equation (5.4).

5.4 Experimental Validation

5.4.1 Data and Evaluation Metrics

Our experiments are conducted on the MSPPodcast(126) and IEMOCAP (27). MSP Podcast is the largest human-labeled emotion dataset with 29,965 training examples and 10,013 test utterances.

MSPPodcast, comprising human annotated podcasts, has labels for three continuous dimensions - valence, arousal, and dominance, and for five discrete emotions - neutral, angry, happy, sad, and disgust. In this work, we report our results on the balanced test1 evaluation set with 30 male and 30 female speakers. We also use IEMOCAP, a 20-hour dataset, with annotations on acted emotion. It comprises 10 different speakers, 5 male, and 5 female. We use the same continuous dimensions and the same 5 emotions that are used from the first dataset.

Both of these datasets are publicly available and have contrasting labels, meaning one is acted and the other spontaneous (127). Discrete emotion recognition is evaluated using F1 or accuracy. For continuous prediction, we compute CCC for each of the attributes.

5.4.2 Model Hyperparameters

All our models are built using ESPNet2 (128)

We use `HUBERT-large` (121) embeddings as the input features following from past work (129). The encoder consists of a 4-layer conformer with 64 hidden units. We employ separate self-attentive pooling(130) layers for continuous and discrete emotion. The decoder consists of MLPs that map from the encoded output from 768, 64, 32, and then 3 for the continuous prediction and 5 for the discrete prediction. We use a ReLU activation to ensure that the predicted continuous attributes are strictly positive and use LeakyReLU elsewhere. Dropout of 0.2 is used in the decoder. The models are trained with the Adam and a peak learning rate of 1e-3 for 15,000 warmup steps.

TABLE 5.1: Emotion Recognition Results on IEMOCAP using 5-fold cross-validation: CCC, Valence, Arousal, and Dominance are reported for continuous emotions, and Unweighted accuracy is reported for discrete emotion prediction

Model Description	CCC	CCC-V	CCC-A	CCC-D	Acc
Baseline C	0.580	0.548	0.606	0.566	-
Baseline D	-	-	-	-	0.737
MTL	0.603	0.571	0.669	0.567	0.723
HMTL-DC	**0.667**	**0.660**	**0.717**	**0.625**	0.744
HMTL-CD	0.648	0.651	0.694	0.599	**0.749**
SoTA (131)	0.573	0.527	0.663	0.530	-
SoTA (132)	-	-	-	-	0.740

TABLE 5.2: Results on MSPPodcast: Concordance Correlation Coefficient(CCC) - overall, Valence, Arousal and Dominance is reported on the test1 evaluation set. For discrete emotions, unweighted F1 is reported.

Model Description	CCC	CCC-V	CCC-A	CCC-D	F1
Baseline C	0.593	**0.597**	0.646	0.538	-
Baseline D	-	-	-	-	0.368
MTL	0.587	0.591	0.637	0.533	0.393
HMTL-DC	**0.617**	0.588	**0.675**	**0.584**	0.404
HMTL-CD	0.605	0.554	0.661	0.569	**0.411**
SoTA (131)	0.619	0.485	0.733	0.640	-
SoTA (133)	-	-	-	-	0.340

5.5 Results

5.5.1 Same Dataset Setting

Table 5.1 shows our experimental results on IEMOCAP dataset, where all results are computed using standard 5-fold cross validation (134). Table 5.2 reports the results of experiments on MSPPodcast.

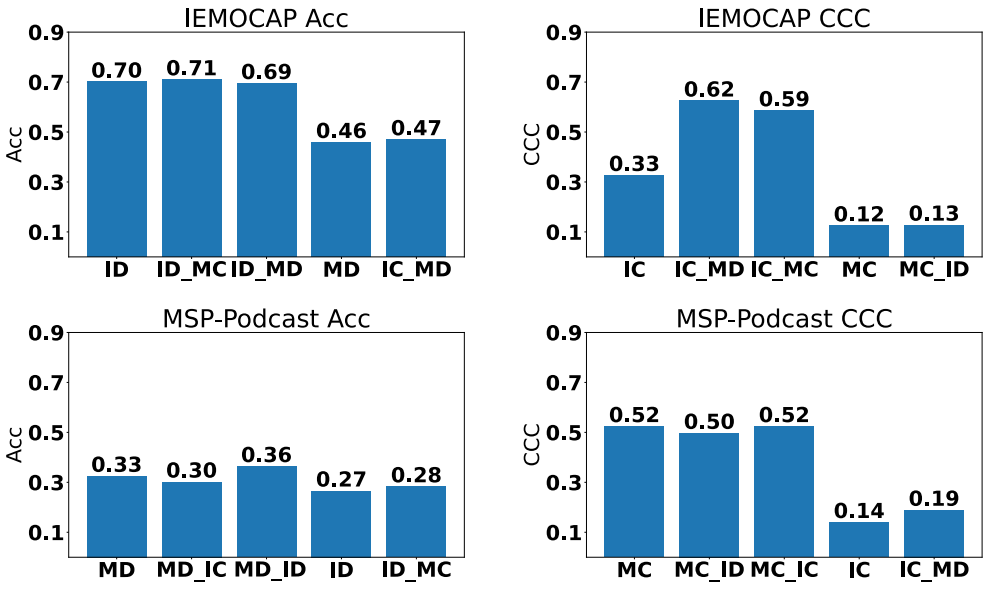
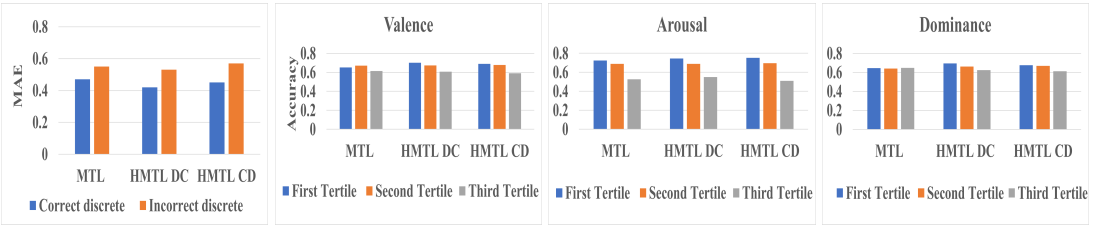


FIGURE 5.7: The plots compare the performance of the discrete and continuous predictions for IEMOCAP and MSPPodcast under five different training conditions. IC refers to when only IEMOCAP continuous emotions are in the training data. MC are when MSPPodcast continuous are in training data. ID refers to IEMOCAP discrete and MD for MSP Podcast discrete. IC_MD refers to when IEMOCAP continuous and MSPPodcast discrete are in training data. Similarly for others.



We make several observations from these results. Firstly, `HMTL-DC` has the best overall CCC for both datasets. On IEMOCAP, it has the best CCC for all three dimensions Valence, Arousal, and Dominance. Whereas in MSPPodcast, the CCC is best in all the dimensions except for Valence. Secondly, the `HMTL-CD` model has the best discrete prediction metric for both datasets.

These results support our hypothesis that knowledge of discrete and continuous emotion attributes can be used to improve performance on continuous and discrete emotion prediction respectively. To determine if `HMTL-DC` and `HMTL-CD` models perform significantly differently from the baseline models, we perform statistical significance tests. For a given utterance U_i, let the model 1 prediction be m_1^i and the model 2, prediction be m_2^i, let $d_i = m_1^i - m_2^i$. The null hypothesis is $H_0 : \mathbb{E}[d_i] = 0$ and the alternate hypothesis is $H_A : \mathbb{E}[d_i] \neq 0$. Under the assumption that the scores are normally distributed, we can perform a standard t-test. Our results show that at the significance of 5%, all the models are statistically different in both discrete and continuous predictions.

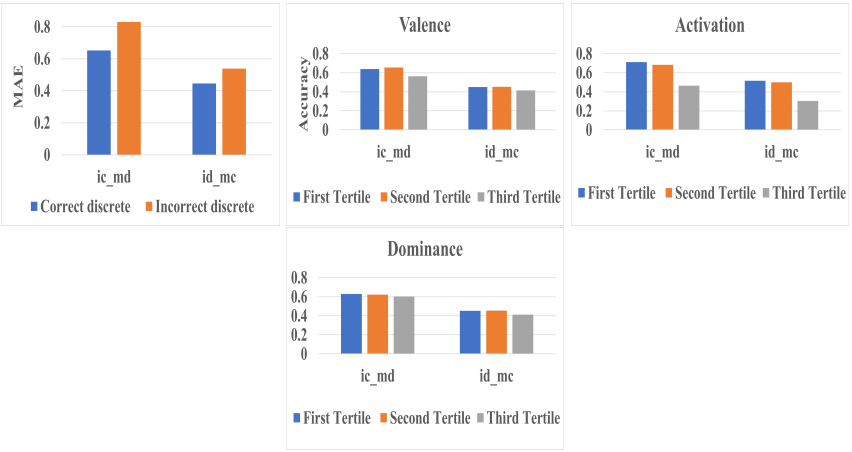


FIGURE 5.8: First row shows the same dataset setting analysis. 1st plot shows CCC change when the discrete predictions are correct vs incorrect. Followed by accuracy change for instances in three MAE bins for Valence, Arousal and Dominance respectively. The second row shows same analysis for cross dataset setting.

5.5.2 Cross Dataset Setting

In this section, we attempt to use discrete and continuous labels from different datasets and analyze gains when training models on multiple datasets with matched (e.g. discrete-discrete) and mismatched (e.g. continuous-discrete) labels. We perform all experiments using a subset of the MSPPodcast data chosen randomly such that the resulting size is the same as that of the IEMOCAP training data. This is done in order to be able to make fair comparisons on the transferability of representations. Figure 5.7 summarizes the results of our experiment on transferring labels across datasets. We perform this experiment in several different settings: (1) only IEMOCAP discrete (ID) labels are used for training, (2) only IEMOCAP continuous (IC) labels are used for training, (3) only MSPPodcast discrete (MD), (4) only MSPPodcast continuous (MC), and other combinations of MD, ID, MC, IC are used for training the models.

IEMOCAP Discrete Prediction: Figure 5.7 top left shows the results of IEMOCAP discrete from different models where the dataset and labels are varying. We observe that when training data contains IEMOCAP discrete and MSPPodcast continuous, the results are the best, i.e in mismatched label conditions.

IEMOCAP Continuous Prediction: Refer to Figure 5.7 top right. We observe the best performance is achieved when IEMOCAP continuous and MSP Podcast discrete are used in training i.e in mismatched label conditions.

MSPPodcast Discrete Prediction: We observe that when the labels are matched, using discrete labels from both IEMOCAP and MSP Podcast dataset, the performance is the highest i.e in matched label conditions.

MSPPodcast Continuous Prediction: We observe a similar trend in this case, where the match label condition helps the MSPPodcast continuous predictions.

In conclusion, we observe that using mismatched labels from the MSPPodcast data for training improves performance on discrete and continuous emotion prediction for IEMOCAP. However the same is not observed for MSP-Podcast where the matched conditions lead to better performance for both discrete and continuous predictions. We believe this might be because the inter-annotator agreement for continuous values in MSPPodcast is much lower than in IEMOCAP (135), which makes it challenging to obtain gains in MSPPodcast using transferred representations from IEMOCAP.

5.5.3 Analysis

To quantitatively examine the interior workings of the model, we examine how the performance changes for the `HMTL-DC` and `HMTL-CD` models in comparison to the MTL model, under different settings. We compare the accuracy of the discrete prediction for the instances where the MAE is lower vs when the AME is higher. Furthermore, we compare the MAE of the continuous predictions for instances where the discrete prediction is correct versus when it is incorrect.

Figure 5.8 shows the performance comparison in MSP Podcast. For the top row of the figure, we compare the models from the same dataset setting. For the first plot, we observe that for all three models, MTL Baseline, `HMTL-CD` and `HMTL-DC`, the MAE of the instances where the discrete predictions are correct is lower compared to when the predictions are incorrect. This shows that discrete predictions help with continuous predictions. The following three charts show the accuracy for the instances in three different tertiles of the MAEs for Valence, Arousal, and Dominance respectively. It can be observed that in the lower tertiles (where the MAEs are lower), the accuracy is higher, and in the third tertile (where the MAEs are higher), the accuracy is lower. This shows that continuous predictions are helping with discrete predictions.

Figure 5.8 lower row shows the same analysis performed in the cross-dataset setting. Here we compare models which use the mismatch label settings i.e. IC_MD and ID_MC. We observe a similar trend as in the same dataset setting. We also analyze the cross-dataset models and observe a similar trend.

Limitation: One limitation of this work is that the different datasets that are used in training hierarchical models need to be in sync with their labeling strategy of discrete and continuous emotions. Otherwise, this strategy won't lead to improvement.

5.6 Conclusion

Different datasets are labeled with either discrete or continuous labels, making it challenging to train large-scale emotion models utilizing all the datasets. In this paper, we introduce hierarchical multi-task learning models that use discrete labels to predict continuous emotion attributes and vice versa. With our method, we obtain absolute improvements of 1.2% Accuracy

and 4.3% F-1 for discrete prediction on IEMOCAP and MSPPodcast respectively. On continuous labels, we improve 0.09 CCC for IEMOCAP and 0.025 CCC in MSPPodcast. Furthermore,

Furthermore, we demonstrate that these continuous and discrete labels need not necessarily be manually annotated within the same corpus to improve recognition performance. Of the prevailing annotated datasets for emotion recognition, only a few such as MSPPodcast (136), and IEMOCAP(137) are annotated for both continuous and discrete attributes. Most, like MELD(138), are annotated for discrete attributes only. We demonstrate that even if the model is trained on continuous emotion labels from MSPPodcast and discrete labels from IEMOCAP, the proposed Hierarchical Multi-Task approach improves performance and generalizability. This implies that emotion recognition datasets can be trained jointly on multiple corpora with different labels. Under this mismatched label setting for IEMOCAP and MSPPodcast, we observe that mismatched labels from MSPPodcast help improve performance on IEMOCAP.

In the next chapter, we will expand the emotion representations beyond the current handful of discrete classes. We use the acoustic properties that are known to be correlated with various emotions and include them as descriptors of the emotions. This effectively increases the number of classes and enhances the model's ability to learn the relationship between acoustic properties and emotions.

Chapter 6

Representing Emotions: Natural Language Acoustic Prompts

In the previous chapter, we focused on how we can combine the use of current discrete and continuous emotion representations together. This allows us to achieve better emotion detection results and also allows models to be trained on multiple datasets that are annotated with either discrete or continuous or both. In this chapter, we will expand this work to present new representation strategies for emotion, which go beyond the use of discrete or continuous labels. We propose the use of natural language, specifically language that includes acoustic correlates of emotions. We show that this improves the emotion decoding performance, in two tasks; speech emotion recognition and speech emotion retrieval.

6.1 Introduction

Emotions are usually described using discrete labels like 'angry', or 'happy' following psychological models like the Plutchik wheel of emotion (139) or Ekman's model of emotion (140). Although these frameworks are extremely popular and provide ease of modeling, they do not fully capture the diversity in emotion expression. This makes using such discrete representations sub-optimal for downstream tasks.

Understanding the source of diversity in emotion expression is the key to formulating more accurate emotion representations. There are many sources of diversity in emotion, like the speaker, culture, and context, among other factors (141; 127). Labeling two instances of emotion with the same label of say 'anger', ignores the intricacies of the expression of anger. Therefore, we believe it is important to represent the fine-grained characteristics of emotion.

These fine-grained characteristics of emotions can be better captured by the flexibility that natural language provides. In general, such descriptions can describe the low-level information in the audio like the acoustic properties or they can describe the high-level information like

who is expressing the emotion and what the context is. Humans often use affective language to casually describe emotion in speech, for example, 'An angry man shouting loudly'. In this example 'loudness' has a direct acoustic correlate 'intensity' which can be used to form a description e.g. 'this is the sound of high-intensity anger'.

The choice of natural language description affects the high dimensional representation learned from the text, hence it is very important to choose the right description for the emotion. This leads to the question:

How do we describe an emotion using natural language and how can a model learn it?

In this work, we propose a method to describe the emotion in audio by using the low-level information in the audio. Previous research shows that there are numerous acoustic correlates of emotion (142; 143; 144). These acoustic correlates include measurements like the average pitch, intensity, speech rate, and articulation rate. We extract these correlates from each utterance and use them to form the description in an automatic and scalable way. We call descriptions generated in this manner *'acoustic prompts'*.

Given these acoustic prompts, we train models that associate them with corresponding audio by fine-tuning the Contrastive Language-Audio Pretraining (CLAP) model (89; 145). CLAP uses contrastive learning to associate the audio and their descriptions and yields state-of-the-art performance in learning audio concepts with natural language descriptions. We then evaluate this fine-tuned model on downstream tasks.

We evaluate on Emotion Audio Retrieval (EAR) and Speech Emotion Recognition (SER). SER is a well-known task defined as given a speech utterance, determine the emotion present in the utterance (86; 127). The task of EAR is not a commonly performed task. There are tangential works e.g. (146; 147) which examine retrieval of music audios, however, this task has not been explored for speech emotion. We believe that EAR is an important task to address since it can be useful in speech forensics, recommendation systems, search engines, social media, etc. Since emotions are also indicators of certain events, EAR methods can help in retrieving hate speech, and violence from audio. We show that the acoustic prompts improve the model's performance in EAR significantly; Precision@K is consistently better for various values of K. We also find that in SER, the model performance improves. Specifically, recognition performance improves 3.8% relative on Ravdess dataset. In a fine-tuning classification setup, we observe 3.7% improvement on Ravdess.

6.2 Proposed Model - CLAP

Fig. 6.1 shows the Contrastive Language-Audio Pretraining (CLAP) model - the backbone architecture used in this paper. The audio-text pairs are passed through an audio encoder and a text encoder respectively. Let $f_a(.)$ represent the audio encoder and $f_t(.)$ represent the text

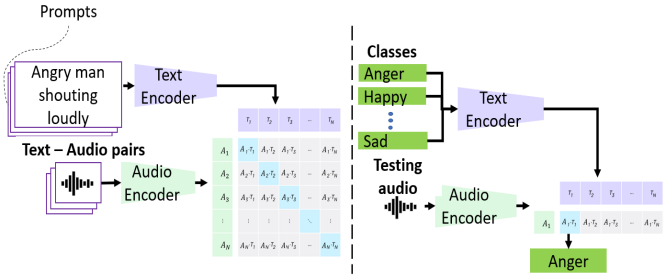


FIGURE 6.1: The left part of the image shows model training. Given a batch of N audio-text pairs, the model trains the audio and text encoders to learn their (dis)similarity using contrastive learning. On the right side is shown an evaluation scenario. Given an audio of unknown emotion, trained audio and text encoders are used to extract representations from the audio and the descriptions. The prediction is made based on the cosine similarity between the two representations.

encoder. For a batch of N:

$$\hat{X}_a = f_a(X_a); \hat{X}_t = f_t(X_t) \tag{6.1}$$

where $\hat{X}_a \in \mathbb{R}^{N \times V}$ are the audio representations of dimensionality V, and $\hat{X}_t \in \mathbb{R}^{N \times U}$ are the text representations of dimensionality U.

We brought audio and text representations into a joint multimodal space of dimension d by using a projection layer:

$$E_a = L_a(\hat{X}_a); E_t = L_t(\hat{X}_t) \tag{6.2}$$

where $E_a \in \mathbb{R}^{N \times d}$, $E_t \in \mathbb{R}^{N \times d}$, L_a and L_t are the linear projections for audio and text respectively.

Now that the audio and text embeddings (E_a, E_t) are comparable, we can measure similarity:

$$C = \tau * (E_t \cdot E_a^\top) \tag{6.3}$$

where τ is a temperature parameter to scale the range of logits. The similarity matrix $C \in \mathbb{R}^{N \times N}$ has N correct pairs in the diagonal and $N^2 - N$ incorrect pairs in the off-diagonal. The loss can be calculated as:

$$\mathcal{L} = 0.5 * (\ell_{text}(C) + \ell_{audio}(C)) \tag{6.4}$$

where $\ell_k = \frac{1}{N} \sum_{i=0}^{N} \log diag(softmax(C))$ along text and audio axis respectively. We used this symmetric cross-entropy loss ($\mathcal{L}$) over the similarity matrix to jointly train the audio and text encoders along with their linear projections.

We chose this architecture because it yields SoTA performance in learning audio concepts with natural language descriptions. We use log Mel spectrograms from the audios, sampled at 44K Hz, as input to the audio encoder - CNN14 (148), which is pre-trained on 2M audio clips from AudioSet. The text encoder is BERT uncased. The audio encodings are of 1024 dimensional from the HuggingFace library (149), whereas text encodings are 768 dimensional. Both encodings are then projected into a joint multimodal space of dimension 1024. Both audio and text encoders are frozen in our experiments, but the projection layers are learnable. We use PyTorch to

implement the model architecture. The model is trained with 0.0001 learning rate, batch size of 128, for 30 epochs using Adam optimizer.

6.3 Proposed Approach - CLAP

6.3.1 Datasets

We use 6 Emotion Datasets (ED) in this setup, see Table 6.2. The literature using these many datasets for emotion tasks are rare. The original CLAP model is trained with audio-text pairs sourced from three audio captioning datasets: ClothoV2 (150), AudioCaps (151), MACS (152), and one sound event dataset: FSD50K (153). Altogether they are referred to as 4D henceforth. All the datasets used are publicly available.

6.3.2 Prompt Generation

For all the emotion datasets being used, we only have the discrete class labels no associated descriptions. Therefore, we devise a scalable and automatic prompting method that is based on the acoustic properties of the speech audios. There are numerous acoustic correlates of emotion therefore, we hypothesize that including this information in the prompts would benefit downstream emotion tasks. We construct the prompts in the manner described below:

1. Class label Prompt: The simplest description for each audio can be the class label, i.e. audio with the discrete true label of 'anger' will be labeled as 'anger'. We use this as the baseline prompt to compare against the proposed prompts.

2. Pitch Prompt: Pitch is known to be affected by emotion, lower pitch is related to negative emotions like fear and high pitch is related to positive emotions like happiness or surprise (143). We bin pitch into four bins, since pitch is naturally sex-specific i.e. low-male pitch (< 132.5 Hz), high-male pitch (> 132.5 Hz, < 180 Hz), low-female pitch (> 180 Hz, < 210 Hz) and high-female pitch (> 210 Hz) However, we also experiment with binning into two classes, based on a cutoff of 170 Hz. The cutoffs are obtained from the average numbers for vocal pitch reported in the literature (154). The prompt is set as 'bin-class emotion-class', an example of which is 'low pitch anger' (without sex information) or 'low male pitch anger' (otherwise).

3. Intensity Prompt: Intensity is known to be affected by emotion, low intensity is linked with negative emotions like sadness or melancholy and high intensity is linked with joy or excitement (143). We bin the average intensity over the audio clip in two bins, low and high intensity at 60 dB (155). The cutoffs are based on average intensity numbers reported for human speech in literature. The same rule as pitch prompt is followed to form the intensity prompt, an example of which is 'high intensity anger'.

TABLE 6.1: Given audio of class label {emotion}, the prompts generated will be one among the following.

Property	Prompt
Class label	{emotion}
Pitch	high female pitch {emotion} low female pitch {emotion} high male pitch {emotion} low male pitch {emotion}
Intensity	high intensity {emotion} low intensity {emotion}
Speech rate	high speech rate {emotion} low speech rate {emotion}
Articulation rate	high articulation rate {emotion} low articulation rate {emotion}

4. Speech-rate Prompt: It has been observed that faster-spoken speech is linked with highly potent emotions such as anger and happiness whilst slower speech is linked with sadness, disgust, and boredom (142). Speech rate is calculated by extracting the number of syllables spoken divided by the total duration of the audio clip. We use 3.12 syllables/sec as the cutoff to bin the speech rate into two bins, low and high speech rate (156). An example of a speech-rate prompt is 'high speech rate anger'.

5. Articulation-rate Prompt: Similarly to speech rate, fast articulation rate is linked with emotions of interest, fear, or happiness; whereas slow articulation rate is indicative of sadness and disgust (142). The articulation rate is calculated as the total number of syllables divided by the total phonation time. We bin the audio into low and high articulation rate at the cutoff of 4 syllables/sec (156). An example of articulation-rate prompt is 'high articulation rate anger'. Even though speech and articulation rate are similar concepts, speech rate captures speaker-specific information in the form of the number of pauses and hesitation whereas articulation rate ignores such information.

6. Prompt Augmentation: To combine all 5 prompts, we pair an audio clip independently with each acoustic prompt. Thus, one audio clip will result in 5 pairs used for training our model. Note: we also tried making one prompt with all the acoustic properties combined together. However, this does not perform as well as when the prompts are paired separately with a given audio.

Table 6.1 shows all the acoustic prompts that are used in this work. We calculate the pitch and intensity using Librosa (157) and we calculate speech rate and articulation rate using Praat (158). Note: Other methods to select thresholds (used in prompt creation) like dataset-specific thresholds showed little effect on the final results, therefore we choose to use the literature-inspired thresholds.

Table 6.2: Details of the 6 emotion datasets used in this paper.

Dataset	Files	Class	Emotions
CMU-MOSEI(122)	23K	9	ang, exc, fear, sad frus, neu, sur, hap, dis
IEMOCAP(27)	10K	9	hap, fear, sad, sur, exc, ang, neu, disappoint, frus
MELD(159)	10K	7	neu, sur, fear, sad, joy, disgust, ang
CREMA-D(160)	7K	6	ang, dis, fear, hap, neu, sad
RAVDESS(161)	2.5K	8	neu, calm, hap, sad, ang, fear, disgust, sur
CMU-MOSI(122)	2.2K	3	neu, positive, negative

Table 6.3: Precision@K achieved under different training conditions and prompt settings. The rows show three different models. The first row is the baseline CLAP model. The second and third rows are models trained on 5 emotion datasets, not including the IEMOCAP dataset. The second row is when the prompts used for training are the emotion class labels (CL) of the audios and the third row is when the prompts are acoustic prompts. PA refers to Prompt-Augmentation The queries here are the acoustic prompts also shown in Table 6.1. The model trained with acoustic prompt augmentation (PA) is consistently better.

	Class Label Queries			Pitch Queries			Intensity Queries			Speech Rate Queries			Articulation Rate Queries		
	P@1	P@5	P@10	P@1	P@5	P@10	P1	P@5	P@10	P@1	P@5	P@10	P@1	P@5	P@10
4D	0.50	0.35	0.35	0.07	0.12	0.10	0.13	0.18	0.19	0.25	0.20	0.15	0.13	0.10	0.14
4D + [5 ED - *CL*]	0.50	0.40	0.35	0.00	0.04	0.05	0.25	0.13	0.15	0.13	0.18	0.19	0.13	0.13	0.13
4D + [5 ED - *PA*]	**0.75**	**0.45**	**0.38**	**0.20**	**0.13**	**0.15**	**0.25**	**0.20**	**0.20**	**0.38**	**0.25**	**0.23**	**0.38**	**0.23**	**0.21**

6.4 Experimental Validation and Results

6.4.1 Emotion Audio Retrieval

We evaluate our trained models for the task of emotion audio retrieval (EAR). With the increasing sizes of audio databases, being able to search such databases for specific types of audio is important. We compare (1) the baseline CLAP model, (2) the model where the prompts used for training are the emotion class labels of the audio, and (3) the model trained with our acoustic prompting method using prompt augmentation.

The first three columns in Table 6.3 show the results when the queries are among the four emotion classes, i.e. happy, sad, angry, and neutral and the collection consists of IEMOCAP dataset. Row 1 model is trained on only 4 audio captioning datasets. Rows 2 and 3 models are trained on 5 emotion datasets, not including IEMOCAP. For a given query, the model outputs top K audios whose audio embeddings have the highest cosine similarity to the text embedding of the query.

We observe that the model trained on acoustic prompts performs significantly better for all the precision@K metrics. This shows that training the model with acoustic prompts is resulting in better-learned emotion representations.

TABLE 6.4: Accuracy % on Ravdess when the model is trained under different settings. The second column shows when Ravdess is not in the training sets. The third column shows when the model is finetuned on Ravdess. The second row shows the CLAP Baseline trained on 4 audio captioning datasets (4D). Third row is when the model is trained using only 5 Emotion Datasets (5 ED). The following rows include 4D and 5ED in training and for the ED, the prompts during training are either the class labels (CL) or the acoustic prompt augmentation (PA) respectively.

Training dataset	Leave one out	Finetune
Random	12.50	12.50
4D	15.99	68.50
5 ED - *CL*	22.88	68.50
4D + [5ED - *CL*]	**38.46**	68.69
4D + [5ED - *PA*]	27.88	**72.46**
SoTA	-	81.82 (163)

Furthermore, we also access whether the trained model learns associations between the acoustic properties and the speech emotion. We test this in a similar framework as in the last experiment. The queries are made similar to the prompts as shown in Table 6.1.

The rest of the columns in Table 6.3 show the results of audio retrieval when queries are from the acoustic prompts. We calculate precision@K for each acoustic prompt shown on the columns. From the results, we observe that the model trained on the proposed acoustic prompting method performs best in all cases. The takeaway here is that our model is able to retrieve audio significantly better when trained using acoustic prompt augmentation. The precision@K numbers are comparable to numbers observed in audio retrieval tasks (162). The results suggest that we can introduce even more elaborate descriptions for each audio at training time and the model will learn associations and be able to retrieve audios with those descriptions.

6.4.2 Speech Emotion Recognition

To evaluate how the acoustic prompts would help in SER, we perform the following two experiments. The first is a zero-shot like setup where we leave one dataset out, which is used during the testing stage. The second is a fine-tuning setup where the model from the first setup is fine-tuned on the left-out dataset.

6.4.2.1 Leave one out

This setup evaluates how well a model trained on a pre-defined set of classes generalizes to a new dataset, which might have same or different sets of classes. Out of the 6 emotion datasets, we leave one out for testing and train the model on the other 5 emotion datasets. Therefore the training and testing datasets are completely different. In the case where Ravdess is the testing dataset, 'calm' class is not represented in any of the other training datasets and is a zero-shot classification result.

We train 5 different models shown in the rows of Table 6.4. There are two main takeaways from this experiment. Firstly adding Emotion datasets in the training stage helps the performance on the left-out emotion dataset. This can be observed in the second column where the performance improves from 15.99% to 22.88%.

Secondly using acoustic prompt augmentation (PA) is not helping in the fine-tuning setup. We believe this is because there is a distribution shift in the training and testing datasets, which effects the acoustics and hence the acoustic prompts. For example, 'high intensity anger' prompt might not be prevalent in the training datasets but is present in the testing dataset. This harms the transferability of the learned acoustic prompts to a completely new dataset. Note that the SoTA performance for this evaluation setup is not found in literature because the general evaluation setup is when the dataset is present in both training and testing sets.

6.4.2.2 Finetune

In this experiment, we fine-tune the model from the previous stage on the left-out dataset.

The results for SER are shown in the last column of Table 6.4. We observe that when using acoustic prompt augmentation, we get the best accuracy metric. We see improvement in performance by absolute 3.77%, from 68.69% to 72.46%.

6.4.3 Prompt Analysis

To evaluate which of the proposed acoustic prompts is better, we apply the trained model on SER with a smaller setup as in the last experiment, where the testing dataset is present in the training dataset. The model is trained 6 different times, where each time the description associated with emotion audios are varied. Among the 6, 1 uses the class label prompt and 4 uses the acoustic prompts as described in Section ??, and 1 uses the prompt augmentation - which combines all the acoustic prompts.

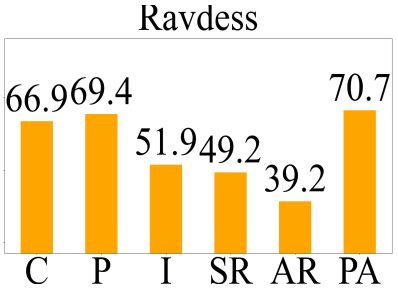


Figure 6.2: Accuracy achieved using different acoustic prompts on Ravdess. C=Class label, P=Pitch prompt, I=Intensity prompt, SR=Speech-Rate prompt, AR=Articulation-Rate prompt, PA=Prompt Augmentation.

We train the model on 4 audio captioning datasets and 1 emotion dataset. The left part of Figure 6.2 shows the performance achieved when the model is trained on the training set (including

4D and Ravdess) and tested on the testing set of Ravdess. We observe that among the 4 acoustic prompts, the pitch prompt gives the best performance. The second-best performance is achieved by the intensity prompt, followed by speech rate and then articulation rate. Secondly, we observe that overall acoustic prompt augmentation is giving the best performance in both datasets.

6.5 Limitation of CLAP Model

There are certain limitations to our work. Firstly, we use only four acoustic properties, however, there are other acoustic properties that are effected by emotion and should be explored. Secondly for each prompt, we create 2 or 4 bins per acoustic property, while these bins could be more fine-grained. Our future study will include work in alleviating the need for thresholding and relying on data-centric methods of binning the prompts. This work performs SER and EAR using the audios and their automatically generated descriptions. We use the acoustics of emotions to prompt the audios, in fact, there can be more complicated descriptions, invoking the semantics, environment, and context among other factors. We envision that as methods of describing emotions become more complicated, our ability to model emotions will become better. The acoustic properties we extract include pitch, intensity, speech rate, and articulation rate extracted from the audio. We find that among the acoustic prompts, pitch prompt is the best performing. Overall for EAR when we do acoustic prompt augmentation, we achieve consistently better Precision@K metric. For SER, we also achieve an improvement in performance in Ravdess by 3.8% in the finetuning setup.

6.6 Going beyond CLAP

One of the limitations of the CLAP model is that the skeleton of the prompts have to be explicitly defined, as shown in Table 6.1. However having defined the prompts, we are limiting the sort of natural language descriptors that the model can use to describe the emotions and hence further improve the classification results. This gives us the motivation for the following improvement on the model architecture of CLAP. We introduce a learnable part of the prompt called the 'prefix' that is in the latent space and learned during the training of the model. The following section provides further details of the model architecture, that we call SELM 'Speech Emotion Language Model'.

6.7 Proposed Model - SELM

Figure 6.3 shows the architecture of the SELM Model. The architecure of the model can be broken down into the following components:

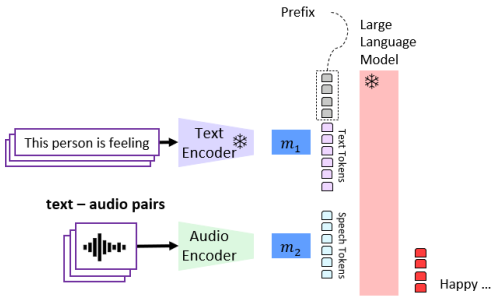


FIGURE 6.3: SELM: Speech Emotion Language Model. The model is fed with audio and text description prompts, which get independently encoded by audio projection audio mapper, and text embedder. The encoded audio and text is used to prompt a Language Model. In the figure, the input text prompt is "this person is feeling" and SELM outputs "emotion of happy". The audio projection and audio mapper are learned during training while the Language Model and Audio Encoder are frozen.

Audio encoder. The audio encoder extracts dense representations from audio. We use Wav2Vec2 as the choice of audio encoder due to its near SOTA performance for SER (164). The 4th layer of Wav2Vec2 contains the information relevant to the task of SER (164). Therefore, we extract the 4th layer of Wav2vec2 as our frozen audio feature extractor.

Text Embedder. The text embedder consists of a Language Model tokenizer followed by embedding lookup to convert language model tokens to continuous embeddings. We keep the tokenizer and embedding lookup same as the Language Model used, which in our case is GPT2 (165)

Audio Projection. The audio projection consists of two learnable linear layers with GeLU activation in between. This is similar to CLAP projection (89), and performs a non-linear transformation of the 4th layer of Wav2Vec2 hidden state. This transformed representation is passed to the Audio Mapper.

Audio Mapper. The audio mapper consists of a sequence mapper and a transformer layer. The sequence mapper is a linear layer followed by a reshape operation to convert a single embedding into a fixed set of continuous embeddings (t_1, t_2, ... t_k). The sequence is passed to a self-attention Transformer. The transformer learns to map the generated sequence from audio (s_1,... s_k) to the latent space of the Language Model (a_1,... a_k).

Text Mapper. The text mapper consists of a learnable Transformer layer to transform text embeddings into the latent space of the Language Model (t_1, t_2, ... t_k).

Language Model. The latent prefix for the Language Model is the concatenation of the audio prefix (a_1, a_2, ... a_k) and text prefix (t_1, t_2, ... t_k). The Language Model generates a text (*emotion of happy*) conditioned on this latent prefix. We use GPT2 (165) as the choice of Language Model and a prefix size of 10 for audio and text respectively.

6.8 Proposed Approach - SELM

Training: The model is trained using the next-token prediction objective. Let one example of training data consist of audio x_{a_i}, input text prompt x_{t_i} and ground truth response y_{t_i}. The audio encoder and audio mapper transform the audio x_{a_i} into a continuous sequence of embeddings $(a_{1_i}, a_{2_i}, ...a_{k_i})$, where k is the prefix size. Similarly, the text embedder and text mapper transforms the prompt x_{t_i} into a continuous sequence of embeddings $t_{1_i}, t_{2_i}, ...t_{k_i}$. Both sequences are concatenated to form the prefix z_i:

$$z_i = a_{1_i}, a_{2_i}, ...a_{k_i}, t_{1_i}, t_{2_i}, ...t_{k_i} \tag{6.5}$$

Then the Language Model conditioned on z_i produces a sequence of output tokens o_{t_i}. The model is trained to predict the next text tokens o_{t_i} ($t \in [0, l]$) conditioned on z_i in an autoregressive fashion. The loss function is Cross-Entropy:

$$\mathcal{L} = -\sum_{t=1}^{l} \log p_\theta(y_{t_i} | z_i, y_{0_i}, ..., y_{l_i}) \tag{6.6}$$

where θ denotes the model's trainable parameters which consist of Audio Projection, Audio Mapper, and Text Mapper. We keep the Wav2Vec2, Text embedder, and Language Model frozen.

Inference: The test audio and test input prompt are used to build prefix z_i to prompt the Language Model. The Language Model predicts the next token based on beam search decoding with a beam size of 3. The output of the model is a text and can be directly used in any downstream pipeline

For evaluating metrics, the free-form text answer from SELM has to be converted to 1 out of C classes for each dataset or user-specified classes. For example, SELM generates the emotion as frustration but the user wants classification into 4 classes of *Happy, Sad, Angry, Neutral*. Therefore, the generated text needs to be mapped into one of the C classes. For this, we encode the generated text and classes with CLAPS text encoder. Then use cosine similarity between text embedding of generated text and text embedding of classes to determine the most likely class.

6.9 Experimental Validation and Results

6.9.1 In Domain Setup

To check the in-distribution performance of SELM we use the in-domain setup. In this setup, the model is trained using the training subset of the Evaluation Datasets and evaluated on the testing subset of them. In explicit train-test subset are not available, we perform N-fold cross-validation.

In SELM, the audio encoder is Wav2Vec and is frozen i.e. not learned during training. We compare against literature baselines where the Wav2Vec2 is frozen followed by learning single or multiple linear layers. The results are shown in Table 6.5 for three datasets. We present SELM performance in the first row, and compare it to the methodology where Wav2Vec2 features are used followed by a simple neural network on top. We believe this is a fair comparison as the acoustic model architecture is the same in both models. SELM's better performance shows the benefit of using the learnable prefix layer and using the large language model.

	In-domain Dataset		
Model	RAVDESS ↑	CREMAD ↑	IEMOCAP ↑
Wav2Vec2 FT	56.53 (163)	46.02 (166)	69.90 (65)
SELM	75.70	88.16	73.09

Table 6.5: In-domain performance of SELM on three datasets. Similar to SELM, the benchmark numbers also use wav2vec2-base embeddings to extract acoustic features.

6.9.2 Out of Domain Setup

We simulate the Out-of-Distribution (OOD) setup by excluding some datasets from training. Therefore, the model is evaluated on three Evaluation Datasets namely RAVDESS, CREMA-D, and IEMOCAP, not used in training.

The existing literature models are also not trained on these three datasets, making this a valid OOD comparison. Table 6.6 presents the OOD experimental results for RAVDESS, CREAM-D and IEMOCAP. The first half of the table presents performance on all emotion classes in the dataset, while the lower half (denoted with *) presents numbers on a subset of classes. The subset of classes are chosen to be the four primary emotions: *happy, sad, angry, neutral*. The results show SELM outperforms existing models on all three datasets for both all-class and 4-class setup. Moreover, some of the datasets contain emotions that SELM has not seen during training, and hence showcasing the generalization ability of the model.

6.9.3 Few Shot Learning

Adapting to new distribution requires annotated examples for the target domain. To test the efficiency of this adaptation, we explore Few-Shot Learning (FSL) for SELM. For the FSL setup, we inherit the pretrained SELM model. Then the model is finetuned using N examples from target domain. The finetuning is restricted to specific parts of model to be parameter efficient and to prevent loss of generalization ability of SELM. We perform FSL in two settings, 4-shot and 8-shot, where 4 examples or 8 examples per class respectively are randomly picked from the training set to finetune specific parameters of SELM.

We compare the Few-Shot learning performance of SELM against literature benchmarks. The results are shown in Table 6.7. The table shows the performance of the model when it is finetuned on either 4 examples per class (4-shot) or 8 examples per class (8-shot). For the

	Out-of-Domain Dataset		
Model	RAVD ↑	CREMA ↑	IEMOC ↑
Random	12.50	16.70	-
CLAP (Audio) (89)	16.00	17.80	13.71
Pengi (167)	20.32	18.46	-
MMS large (168)	13.50	17.20	-
Whisper medium.en (168)	15.30	20.90	-
Whisper medium (168)	16.70	19.90	-
Whisper large-v2 (168)	15.10	20.20	-
AudioFlamingo (169)	20.90	26.50	-
SELM (Ours)	**24.51**	**28.30**	**21.42**
CLAP * (89)	29.48	31.05	33.72
LanSER PS* (170)	-	15.90	30.90
LanSER CM* (170)	-	23.50	34.30
CLAP (Ours)* (171)	38.46	35.22	-
SELM (Ours)*	**52.53**	**42.79**	**40.02**

TABLE 6.6: OOD Performance of different models across three datasets. The dataset is considered OOD when labeled or unlabelled audio is not used during training or unsupervised adaptation during testing. The * symbol indicates only four emotion classes (anger, happiness, sadness, and neutral) are used for evaluation. The metric used is unweighted.

all-class settings (presented in the earlier rows of the table), we compare SELM's performance against Audio-Flamingo (169). The AudioFlamingo reports 4-shot and 8-shot numbers, however the testing strategy and split is not reported. For the 4 class emotion classification, we compare againts LanSER (170), which reports the performance when the model is fine-tuned using 10% of the data. We note that 10% data from target domain is significantly more than 4-shot. For example, 10% of data from CREMA-D is equivalent to 744 audio files and hence the setups are not directly comparable. Due to the lack of Few-Shot Learning approaches in SER, we compare SELM against LanSER

The results shown in Table 6.7 lead to multiple conclusions. First, SELM performs better than prior work on CREMA-D and IEMOCAP. Second, 8-shot always leads to performance improvement over 4 shot. Similarly, 16-shot performs better than 8-shots, i.e. providing SELM more target domain data will lead to better performance. However, there SELM Few-Shot has some limitations. First, the AudioFlamingo performs better on RAVDESS in all-class settings than SELM. We believe that this is due to the specific training strategy of AudioFlamingo which enables use of In-Context Learning. Moreover, AudioFlamingo is trained on 6 million instances, which is 20 times higher than our settings. Second, in RAVDESS, the 4-shot leads to higher performance than 8-shot in the 4-class settings. This might be because RAVDESS has song and speech samples, while SELM has never encountered songs in training data. As songs have different audio distribution from speech samples, the parameter-efficient Few-Shot learning is not sufficient to improve performance and adapt to the song distribution in a few examples.

6.10 Conclusion

SELM is an improvement over the CLAP model that was proposed earlier. Comparing the zero shot performance of SELM to CLAP, we can observe that the performance improves from 38.46

		Dataset for Few-Shot Learning		
Model	Setup	RAVD ↑	CREMA ↑	IEMOC ↑
AudioFlam.	4-shot	-	30.47	-
AudioFlam.	8-shot	**35.20**	31.80	-
SELM (Ours)	4-shot	30.09	30.10	25.32
SELM (Ours)	8-shot	31.81	**32.27**	**27.01**
LanSER PS*	10%	-	35.50	42.00
LanSER CM*	10%	-	43.70	50.00
SELM (Ours)*	4-shot	**57.14**	43.93	50.17
SELM (Ours)*	8-shot	55.77	**46.02**	**50.17**

TABLE 6.7: Few-Shot Learning results of different models and methods on OOD dataset. The dataset is considered OOD when labeled or unlabelled audio is not used during training or unsupervised adaptation during testing. The * symbol indicates only four emotion classes (anger, happiness, sadness, and neutral) are used for evaluation. The metric used is unweighted accuracy.

to 52.53 in RAVDESS. Using few shot learning in SELM, this performance improves to 57.14 using 4-shot and 55.77 using 8-shot examples. This shows how improving the representation capabilities of emotion, by using natural language descriptions (in CLAP) or by making the model learn such descriptions (in SELM), we can improve the learning capabilities of our model, and improve emotion recognition performance.

Part II

Chapter 7

Prior Work

7.1 Theoretical Personality Models

The question of what constitutes 'Personality' has garnered a lot of interest from various fields. Researchers have tried to understand how to classify individuals into different personalities, how many unique personalities there are, what are the different dimensions that each individual should be classified over, and how should each dimension be assessed. There are numerous propositions for each. We mention some of the more accepted/used theories here:

- Five-factor Model: Researchers initially studied the lexical terms in the English language which can be used to describe all the traits for different human personalities. Using factor analysis, five dimensions were discovered (172; 173). These five basic dimensions included: Openness to Experience (O), Conscientiousness (C), Extraversion (E), Agreeableness (A), and Neuroticism (N). OCEAN for short. These five traits have been consistently found in subsequent studies. As mentioned in (174): "In many ways, it seems remarkable that such stability should be found in an area which to date has granted anything but consistent results." The same methodology was performed using other languages like Czech, Dutch, Filipino, German, Greek, Hebrew, Hungarian, Italian, Korean, Polish, Russian, Spanish, Turkish, and other languages, and the findings suggest the existence of the same five personality factors (175).
- NEO Personality Inventory-Revised (NEO PI-R): Further work on the Five-factor model elaborated each of the five factors into sub-traits called facets (176). For example, the six facets the researchers have identified for agreeableness are trust, straightforwardness, compliance, altruism, modesty, and tender-mindedness.
- Three Dimensions of Personality (PEN): Introduced by (177) who identified two super-factors: Extraversion (E) and Neuroticism (N) and Psychoticism (P). Each of these super-factors contained low-level traits.

- Myers-Briggs Type Indicator (MBTI): Introduced by (178) where individuals are classified on four different dimensions (1) Extroversion (E) versus Introversion (I); (2) Sensing (S) versus Intuition (N); (3) Thinking (T) and Feeling (F) and (4) Judging (J) versus Perception (P). Each individual is given a four-letter personality depending on where the individual falls in each of the four dimensions. For example, INFP would mean Introverted-Intuition-Feeling-Perception individual. Although MBTI is widely acceptable among individuals, it is criticized for its oversimplification and for limiting the unlimited space of personality into 16 discrete types.

- ARC types: Introduced by (179), the authors cluster the Big-Five personality traits from the Five-factor model into three personality traits. (1) Resilient, (2) Overcontrolled, and (3) Undercontrolled. These three personality traits are different combinations from the five personality traits identified in the Five-Factor model. For example, people with overcontrolled personalities have high neuroticism and low extraversion. Undercontrolled personalities are low on agreeableness and conscientiousness but low on neuroticism and openness.

- Four Personality types: A much recently proposed is a four-factor model (180), which clusters individuals into four clusters, where individuals around identified clusters show existence of four robust personality types. These four-personality types are also some combination of the personliaty identified in the Five-Factor model. These four include (1) Role Models: Socially desirable, low on neuroticism, and high on all other four traits. (2) self-centred - not as socially desirable, low scores on openness, agreeableness, conscientiousness, (3) Reserved - not so socially desirable, Low scores on neuroticism and openness (4) Average - least robust, average score on all traits.

7.2 Computational Personality Models

[TODO] Over the years, there have been numerous works that have tried to detect personality traits from various modalities directly. Personality detection from text has been vastly studied, especially because there are more cleanly compiled datasets available for this task. However there are very few speech datasets available for personality detection.

Works studying personality from speech have studied various features that are useful for the task. One of the earliest works (15; 18) find correlations between different speech features and personality traits, i.e. they find that 'significant positive association between personality trait of dominance and the voice characteristics of loudness, resonance, and lower pitch.' and find that there exists positive correlations between submissiveness and high speech rate, etc. Studies like (181) further report similar correlations between voice qualities and personality traits e.g. it states that 'result might indicate that a more aggressive trait is associated to a tenser or creakier voice'. Another study (182) specifically shows how extroversion can be assessed

from voice. More recent explorations have studied features like prosody (183; 28; 184), MFCC (185), non-verbal features like speaking time length, number of pauses (186).

For modeling the speech personality task, studies have explored various models like GMM (187), SVM (188). Among recent works, (189) proposes to use frame level low level speech features from opensmile toolkit, and train a neural auto encoder model for binary classification of the OCEAN traits into high and low classes.

When we started exploring this research topic, I realized the scarcity of data in the field. My first intention was to contribute a dataset for task that is more reliable in terms of its annotation, and is much larger in scale. However, i realized soon that this is an extremely difficult task to delve into. In the next chapter, I describe the work done on the data collection front and the setbacks faced. Furthermore, instead of relying on the annotated data for personality, and building more data intensive models, I move towards more knowledge based exploration of the task. In the coming chapters, I explain how we do that.

Chapter 8

Data Collection - VoxCeleb for Personality

8.1 Introduction

One of the biggest problems with studying personality from speech is the lack of datasets in the field. Dataset collection is not an easy task especially in as subjective a task as personality. Secondly getting reliable annotation is a bigger challenge.

Some of the available datasets, including speech modality, are listed in Table 8.1. Among these, the widely used dataset is SSP-Net, which is a collection of French TV broadcasts. The reason it is the most widely used is that each audio file in the dataset is annotated by 11 different raters, among which some are professionals and some are laymen. The dataset with the greatest number of hours is UDIVA, consisting of 90.5 hours of speech.

Our aim was to release the dataset with the highest number of audio for this task, divided into two sets, one consisting of shorter audio and the other longer ones. The data consists of 20000 audio clips with a total of 52.2 hours. With this dataset, the unique aspect is that we will be getting annotations for shorter and longer audios, where the smaller audios range from an

Dataset	# Clips	# Spks	# Hrs	Annt. Type
SSPNet (28)	640	322	1.7	External Raters
ELEA (190)	22	85	5	External Raters
FirstImpression (191)	10000	3000	41.7	MTurk
YouTube Lens (186)	2269	469	150	MTurk
MHHRI (192)	746	18	4.25	Self Reported
UDIVA (193)	940	147	90.5	Self Reported
VoxCeleb-Pers (Ours)	20000	958	52.2	MTurk

Table 8.1: Datasets present for Speech Personality. From left to right: the total number of clips, number of speakers, total number of hours, personality annotation type (how personality was annotated).

President	N	E	O	A	C
Lyndon Johnson	95	99.4	7	0.07	72
Richard Nixon	97	7	14	0.02	98
Gerald Ford	15	91	8	53	69
Ronald Reagan	4	98	10	26	9
George H. W. Bush	58	55	18	29	23
James Carter	76	58	77	56	98
John F. Kennedy	27	99.6	82	11	5
Bill Clinton	58	99.9	82	24	5
George W. Bush	49	99.6	0.03	4	9

Table 8.2: Personality OCEAN Trait Scores for the Presidents. These scores are percentiles with respect to the U.S. general population.

Presidents	Carter	Clinton	Ford	HW Bush	Johnson	Kennedy	Nixon	Reagan	W Bush	Total
# Hours	8.38	17.80	7.93	10.05	27.80	14.37	9.65	21.77	14.70	132.5

Table 8.3: Duration of Audio (Hrs) for each President.

average of 5 seconds, and the longer audios an average of 1 minute. Personality perception as a function of audio length has not been studied before and informs us how differently or more accurately people make judgments about others when hearing shorter or longer audio.

One of the lacking aspects of the current datasets is the way that they are annotated for personality labels. The labels are either self annotations or crowd-sourced through Amazon Turk. However, neither self-reporting nor crowd-sources are experts in the understanding of personality and cannot therefore accurately rate personality traits. Therefore, we aimed to work on the creation of a speech personality dataset where the annotations are expert labelers, and therefore the labels are more trustworthy.

United States Presidents' lives have been studied extensively over the years. In the book *'Personality, character, and leadership in the White House: Psychologists assess the presidents'* (194), many expert psychologists study 21 United States presidents, studying each president throughout their lives, from birth to their time in office to afterward. They have rated the president on the Big-5 OCEAN traits. Table 8.2 shows the percentile scores for the OCEAN traits for some of the presidents. There are 9 Presidents, for whom we can find their voices online. The earliest is John F. Kennedy and the latest is George Walker Bush (the later presidents are not rated in the book and therefore are not included in the study). We collect audio files for each of these presidents from the Miller Centre (195). Table 8.3 shows the number of hours we have per president. We will use this data to analyze our models.

8.2 Voxceleb Dataset

Voxceleb dataset (196) is one of the largest datasets, used to study many different tasks. It is a collection of celebrity interviews, consisting of audio-visual modality. We use Amazon Turk to

#	Question
1	This person is reserved
2	This person is generally trusting
3	This person tends to be lazy
4	This person is relaxed, handles stress well
5	This person has few artistic interests
6	This person is outgoing, sociable
7	This person tends to find fault with others
8	This person does a thorough job
9	This person gets nervous easily
10	This person has an active imagination

TABLE 8.4: BFI-10 Questionnaire

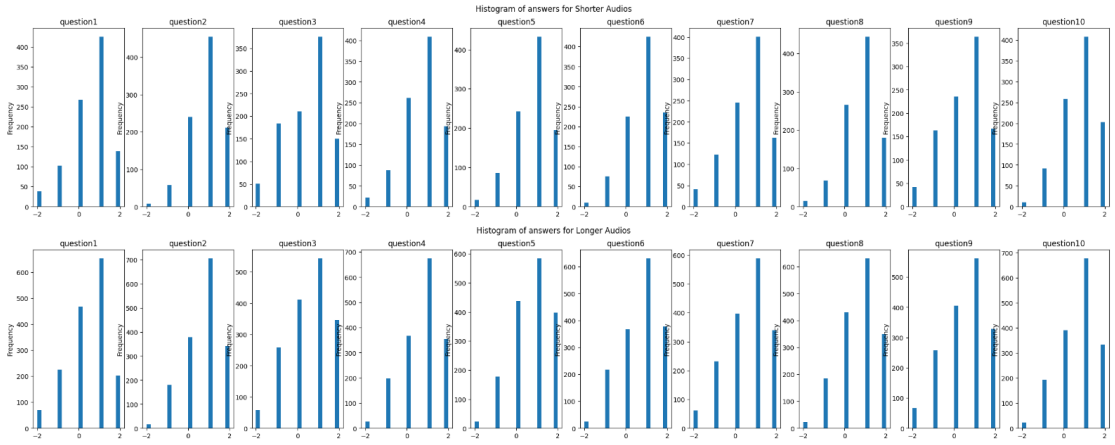


FIGURE 8.1: Histogram depicting the frequency of the chosen answer by raters for the short audios (above) and long audios (below).

collect personality annotations for this dataset. We use the BFI-10 questionnaire (197), shown in Figure 8.4. The questionnaire is a compressed version of the original BFI-44, a set of 44 questions (198). These 10 questionnaires have been shown to cover the same subset of phenomena that the original set covers. The lower number of questions makes it easier to collect responses from the raters. Each question can be answered on a scale of 1-5 where the rater selects from Strong agrees, to Strong disagrees. The rater listens to an audio and then proceeds to answer the 10 questions. We collect some demographic information for each rater, including their gender, age, and education level. As previously mentioned, there are numerous ways in which personality traits can be annotated, but the most widely accepted is the Big 5 OCEAN traits. Therefore, we also annotate the speakers on these traits. These traits can be extracted from the BFI-10 questionnaire easily (28).

8.3 Results and Conclusion

Here I present the number of audio files for which we collected the annotations for. We collected annotations for 324 short audio files and 90 long audio files. Each audio file is rated by 3 raters

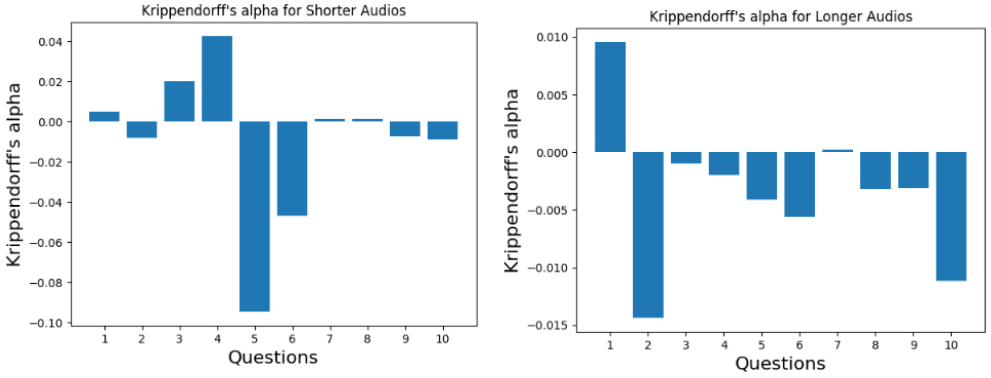


FIGURE 8.2: Krippendorff's alpha for the shorter audios (top) and for longer audios (bottom) for each of the 10 questions. We can observe that the Krippendorff's alpha is poor for all the questions for both settings.

on the 10 BFI questions. Figure 8.1 shows the histogram of the chosen answer by the raters. We can observe that majority of the raters prefer to choose "I Agree" option for each of the 10 questions. The second frequent answer is "Neutral". The trend of the answers tend to more towards the "I Agree" and "Strongly Agree". Very few raters select the "Disagree option".

Figure 8.2 shows the Krippendorff's alpha for the short and longer version of the audios. We can observe from the bar chart that the Krippendorff's alpha is poor for both scenarios for all the questions. Above 0.8 Krippendorff's alpha is acceptable and considered good agreement, however what we observe in the ratings collected shows poor agreement.

In conclusion, from the above results we can observe that the BFI-10 questionnaire has very little agreement among the raters. This agreement leads to poorer agreement among the OCEAN traits as well. Our understanding is that this low agreement is either due to the platform used in this data collection effort and the bad faith of the raters, or that it is genuinely a hard task. This agreement is in the same ballpark as reported for other datasets. The field of OCEAN trait understanding could benefit from a better questionnaire based assessment from personality that is more refined than the BFI-questionnaire, especially one that is geared towards the understanding personality from speech.

Therefore, instead of going down this route, we focus more on knowledge based analysis of personality. We delve deeper into prior literature which link low level speech features to personality. This leads to handcrafted features, more explainable model building and better understanding of personality from speech.

Chapter 9

Knowledge Based Features for Personality Evaluation

There has been a lot of work in building computation models for personality as explored in the previous section. However building computational automatic models required enough data and also good quality labels, both of which are hugely lacking in this area.

In our own effort of making automatic models for personality detection, we faced with these issues whereby data is scarce, especially speech based data, and secondly, the quality of the labels is really low, with very low inter-annotator agreement.

To tackle the above issues, we take a more knowledge-based approach to the problem then a data-intensive approach. There is bunch of literature which study how voice quality features are reflective of certain personality traits, and in order to reach to these voice quality features, we delve deeper into the low level features. We form this chain of formulas, which first go from low level features, to voice quality features, finally to personality ratings. Below we g o in depth of how this is done.

9.1 Introduction

The fact that voice carries information about the speaker has been observed and noted for centuries. In daily life, humans make myriad judgments about people based on their voice, especially the quality of their voice. Voice quality has indeed long been recognized as a fundamental aspect of human communication, influencing and supporting practice in various fields such as music, speech therapy, interpersonal communication, and speaker profiling. For example, (199) highlights the importance of vocal quality in conveying emotion and style. (200) emphasize the impact of voice quality on speech clarity and the therapeutic strategies to rehabilitate voice disorders. In interpersonal communication, the nuances of voice quality can influence perceptions of trustworthiness, attractiveness, and authority, as explored by (201).

Studies like (202) discuss how vocal characteristics can be leveraged for forensic purposes, providing critical evidence in legal contexts.

Despite its significance, the assessment of voice quality has predominantly been subjective, described with terms like “rough,” “breathy,” “twangy,” etc. This subjectivity, while capturing the nuanced perceptions of human listeners, presents challenges for consistent and accurate analysis, especially in applications requiring precise differentiation of voice characteristics. Such examples include speaker profiling for security and law enforcement, where the automated deduction of a multitude of speaker characteristics from voice is desired, but their relationships to voice have only previously been documented in terms of subjective assessments of the corresponding voice qualities.

Similarly human personality is a subjective characteristics, however there have been numerous studies to highlight the different objective ways in which humans can be characterized and hence rated on different personality bases. The Big-5 OCEAN traits is one such methodology, where every individual is rated on each of the 5 dimensions, including Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. Personality understanding, however more advanced than voice quality assessment, proves to be challenging task for rating individuals using this methodology. Raters when ranking an individual on OCEAN traits result in extremely low inter annotator agreement, proving the challenging nature of the task, and making it more so important to have a more standardized way of ranking individuals on various personality traits.

The need for objective measurement arises from the limitations of subjective evaluation, which can vary widely between listeners and lack the precision necessary for applications such as medical diagnostics, forensic analysis, and automated speech processing. Objective measures can provide a standardized way to evaluate voice quality and personality, enabling consistent assessments across different contexts and applications.

The goal of this paper is to offer a path toward achieving this goal - to propose objective assessment of voice qualities and personality traits. By analyzing the physical properties of voice signals, many prior studies have established correlations between these low-level signal features and the perceived qualities of the human voice and between voice qualities and perceived personality OCEAN traits. Such studies have identified and noted many specific signal characteristics – such as frequency, amplitude, pitch, and other spectral characterizations – that relate to the subjective qualities traditionally used to describe voice. Secondly, there are numerous studies which correlate e.g. voice qualities like the hoarseness of voice to personality traits like Extraversion (182). Leveraging these correlations, it is possible to develop formulae or algorithms that quantitatively assess voice quality based on measurable signal properties and personality based on quantitative voice qualities.

Towards this goal of quantitative expression and evaluation of voice quality, we propose a clear 3-step procedure: 1) We choose 24 voice quality features that are widely used in different fields, especially in clinical literature. For each of these, we collate the results of prior scientific studies

that have documented the statistical relationships of these voice qualities to members of a set of 25 low-level signal characterizations. For OCEAN personality traits, we document for each trait, prior literature which have found correlations between the 5 personality traits and 24 voice quality features. 2) We translate these observations into a set of formulae with linear terms and weights. 3) We experimentally evaluate the consistency and accuracy of these formulae by devising human assessment and data-driven experiments. We show that the formulae are consistent and strongly follow the expected subjective trends of voice quality and personality traits assessment by trained humans.

9.2 Related Work

Given the usage and significance of voice quality in various scenarios especially in diagnosing health conditions, speaker profiling etc, over the years multiple methodologies have been developed for measuring voice quality. These efforts include visual analysis, perceptual evaluation, aerodynamic measures, acoustic analysis, and self-evaluation by the individual (203; 204). Below we give examples of some of these efforts.

Auditory perceptual assessment is the most widely used methodology for measuring voice quality. Laver's Voice Profile Analysis (205) proposed a phonetic classification system, using differences in the laryngeal and supra-laryngeal settings of the voice production systems for describing differences in voice qualities. Other methods include evaluation protocols like GRBAS (206) scale. It evaluates an individual on 5 voice qualities including roughness, breathiness, asthenia, strain and grade of the voice. Another example is RBH scale (207) which evaluates 3 voice qualities including roughness, breathiness, and hoarseness. CAPE-V (208) evaluates several voice qualities including loudness, diplophonia, vocal fry, falsetto, asthenia, aphonia, pitch instability, tremor, wet/gurgly, roughness, breathiness and strain. (209) proposes a metric based on the phase spectrum, demonstrating its effectiveness in detecting voice disorders by capturing irregularities in the glottal source. They test their methodology on private medical dataset and verify their ranking of voice quality against doctor's rankings. Some other methodologies analyse power spectra of a given speech signal over time, which reflect differences in voice qualities.

Other proposed models evaluate sustained vowels by an individual. Such models include Acoustic Voice Quality Index (i.e., AVQI) (210) and Cepstral Spectral Index of Dysphonia (211), widely used in clinical settings to identify vocal abnormalities.

While existing methodologies contribute to voice quality assessment, they often suffer from limitations. Some require expert analysis, while others impose specific requirements on the individual, such as sustaining vowels. Additionally, these methods typically focus on a subset of voice qualities, neglecting a comprehensive evaluation. This necessitates the development of readily deployable, user-friendly approaches that capture the full spectrum of voice qualities without imposing stringent pre-conditions.

Secondly, personality OCEAN trait prediction is a very useful task for more natural human-computer interaction, for aiding in the assessment and monitoring of mental health and psychological problems in humans, etc. There are existing datasets as well for which speech samples are labeled for the OCEAN traits of the speakers. However such datasets are small in quantity and more over are of poor quality. The OCEAN labels are either coming from self-ratings from the speaker (which are less reliable since speakers have little knowledge about the OCEAN traits), or are labeled using crowd-sourcing. Crowd-sourcing labels bring numerous challenges with them, including poor inter-annotator agreement (IAA). Whereby the most widely used datasets have IAA of 0.12-0.2 (28), which is poor.

Regardless of the poor quality of the openly available datasets for personality, many prior works have tried to automatically deduce personality from speech (183; 212; 189) by performing a binary classification task, which is predicting for each of the 5 OCEAN traits, if an individual scores high or low on that trait.

Instead of going towards a more data-driven approach for personality analysis, we shift our focus towards a more knowledge-driven approach. We develop a framework based on more knowledge based features, relying on findings from prior literature which have found correlations between voice quality features and personality traits, and report results for personality prediction task.

9.3 Proposed Formulation - Voice qualities, OCEAN traits, signal characteristics and rationale for selection

In this paper we focus on 24 voice qualities (VQF). These are listed below. The detailed definition and description of each of these qualities can be found here (213).

1. Coveredness (Cov)
2. Aphonicity (Aph)
3. Biphonicity (Biph)
4. Breathiness (Brea)
5. Creakiness (Crea)
6. Diplophonicity (Dip)
7. Flutter (Flu)
8. Glottalization (Glo)
9. Hoarseness (Hoa)
10. Roughness (Rou)
11. Nasality (Nas)
12. Jitter (Jit)
13. Pressed (Pre)
14. Pulsed (Pul)
15. Resonant (Res)
16. Shimmer (Shim)
17. Strained (Stra)
18. Strohbassness (Stro)
19. Tremor (Tre)
20. Twanginess (Twa)
21. Ventricular (Ven)
22. Wobble (Wob)
23. Yawniness (Yaw)
24. Loudness (Lou)

We focus on the following 25 low level speech features (LLF), which are listed in the table below, and whose detailed definitions can be found in (213).

1. Loudness
2. alphaRatio
3. hammerbergIndex
4. slope0-500
5. slope500-1500
6. spectralFlux
7. mfcc1
8. mfcc2
9. mfcc3
10. mfcc4
11. F0semitoneFrom27
12. jitterLocal
13. shimmerlocaldB
14. HNRdBACF
15. logRelF0-H1-H2
16. logRelF0-H1-A3
17. F1frequency
18. F1bandwidth
19. F1amplitudeLogRelF0
20. F2frequency
21. F2bandwidth
22. F2amplitudeLogRelF0
23. F3frequency
24. F3bandwidth
25. F3amplitudeLogRelF0

9.3.1 Correlating LLF to VQF, and VQF to OCEAN

Correlations for voice qualities are established based on extensive analysis of prior literature. Table 9.1 shows the correlations between 24 voice quality and 25 low level speech features. To give an example of how a column is filled up in Table 9.1, lets look at the voice quality of 'Breathiness'. (301; 302; 303; 275) demonstrates that breathiness has a negative correlation with loudness, alphaRatio (the ratio of the summed energy from 50-1000 Hz and 1-5 kHz), hammarbergIndex (the ratio of the strongest energy peak in the 0-2 kHz region to the strongest peak in the 2-5 kHz region), a positive correlation with spectral flux (a measure of the change in the spectral content of a sound over time), has a positive correlation with F1, F2 and F3 bandwidth and (304) shows that breathiness has a negative correlation with F2 and F3 frequency.

The other columns in Table 9.1 have been similarly completed based on a collated analysis of findings in prior literature. For example, a subset of findings we refer to in this work are: nasality (243; 305), tremor (248; 306; 307), creakiness (308; 309; 286), hoarseness (310), pressed (311; 215), loudness (312), and many others.

The collection of all the formulae can be found here: `https://github.com/ydhira/vqf_ocean_formulas_pipeline`

To derive the formulae, we use the correlation information that we have collected in Table 9.1, as follows: A given voice quality i (vq_i) can be written as a function of each of the low-level features j (llf_j) as shown below:

$$vq_i = 1/Z * \sum_j \frac{c_{i,j} * w_{i,j}(v_{\text{llf}_j} - \mu_{llf_j})}{\sigma_{\text{llf}_j}} \tag{9.1}$$

TABLE 9.1: Shows the correlation of 24 voice quality features (columns) with the 25 low level features (rows). For reference the color palette can be found in Table 9.3.

Feat↓ VQuality→	Cov	Aph	Biph	Brea	Crea	Dip	Flu	Glo	Hoa	Rou	Nas	Jit
Loudness			-	(214)	(215)	-	(216)	-	(217)	-	-	(218)
alphaRatio		(213)		(219)	(220)	(221)	-				(222)	
hammarbergIndex		(213)	-	(219)	(223)	(221)					(222)	-
slope0-500		(213)		(224)	(225)	(223)	-					
slope500-1500		(213)		(224)	(224)	-	-	(226)	-			
spectralFlux				(227)		(221)	(216)	(226)	(217)	(217)		
mfcc1			-		(225)			(228)	-			-
mfcc2			-		(225)	-		(228)	-			-
mfcc3			-		(225)	-		(228)	-			-
mfcc4			-		(225)			(228)	-			-
F0semitoneFrom27.5Hz	(213)	(229)	(230)	(214)	(215)	(223)		(226)	(215)	(231)	(232) -	
jitterLocal				(214)	(233)	(234)	-	(226)	(215)	(231)	-	(235)
shimmerLocaldB			-	(214)	(233)	(234)	-	(226)			-	
HNRdBACF			-	(214)	(215)	(234)	(236)	(228)	(237)	(215)	(222)	(238)
logRelF0-H1-H2			-	(214)	(239)			(228)	(240)			-
logRelF0-H1-A3			-	(214)	(239)				(240)			-
F1frequency			-	(227)	(241)	(221)	-		(217)	(217)	(242)	-
F1bandwidth	-		-	(227)	(241)		-		(217)	(217)	(221)	
F1amplitudeLogRelF0			-	(227)			-		(217)		(242)	
F2frequency	(213)	-	-	(241)	(241)	(221)	-		(217)	(217)	(243)	
F2bandwidth		-	-	(227)	(244)	(221)	-		(217)	(217)	(232)	
F2amplitudeLogRelF0			-				-		(217)		(242)	-
F3frequency	(213)	(213)	-	(245)	(244)	(221)			(217)	(217)	(246)	-
F3bandwidth	-	-		(245)	(244)	(221)			(217)	(217)	(232)	-
F3amplitudeLogRelF0	-	-					-		(217)		(242)	-

Feat↓ VQuality→	Pre	Pul	Res	Shim	Stra	Stro	Tre	Twa	Ven	Wob	Yaw	Lou
Loudness	(227)		(213)	(218)	(247)	-	(248)		(249)	-	-	
alphaRatio	-		Neu	-	(250)	-	(251)	(252)		-	-	-
hammarbergIndex	-		Neu	-		-	(251)	(252)		-	(253)	-
slope0-500			(254)	-	(247)	(255)		(252)	(256)		(253)	(257)
slope500-1500			(254)	-	(247)	(255)		(252)	(256)		-	(257)
spectralFlux			-	-	(247)	(255)	(251)					(257)
mfcc1			(254)	-	-		(258)					(259)
mfcc2			(254)	-	-		(258)			-		(259)
mfcc3		-	(254)	-	-		(258)					(259)
mfcc4		-	(254)	-	-		(258)			-		(259)
F0semitoneFrom27.5Hz	(227)		-	-	(231)	(255)	(260)		(256)		(261)	(257)
jitterLocal	-		-		(231)	(215)	(260)		(249)	-		
shimmerLocaldB	-		-	(235)	-	(215)	(260)		(249)	-	(261)	
HNRdBACF	(254)		(254)		(262)	(215)	(260)		(256)	-	(261)	(259)
logRelF0-H1-H2	(213)	-						(253)	-			- (259)
logRelF0-H1-A3		-							-			- (259)
F1frequency	(254)		(254)					(254)	(263)		(253)	(259)
F1bandwidth			(264)				(265)		(249)			(257)
F1amplitudeLogRelF0			(213)								(253)	
F2frequency	(254)		(254)					(253)	(263)		(253)	(259)
F2bandwidth			(264)						(249)			(257)
F2amplitudeLogRelF0		-	(213)				-				(253)	(257)
F3frequency	(254)	-	(254)				-	(254)	(263)		(253)	(259)
F3bandwidth		-	(264)				(265)		(249)			(257)
F3amplitudeLogRelF0		-	(213)									-

where Z is the normalizing factor which equals to the number of llf involved in the equation. This ensures that the absolute value of the voice quality is between 0 and 1. $c_{i,j}$ is 1 when the

8pt8pt

TABLE 9.2: Shows the correlation of the 24 voice quality features (columns) with the 5 OCEAN traits (rows). For reference the color palette can be found in Table 9.3.

OCEAN↓ VQuality→	Cov	Aph	Biph	Brea	Crea	Dip	Flu	Glo	Hoa	Rou	Nas	Jit
Openness (O)	(266)	(267)		(268)	-	-	-	-	(269)	(270)	(271)	(184)
Conscientiousness (C)		(267)		(272)	-	-	-	-	(184)	(233)	(233)	(233)
Extraversion (E)	(273)	(274)	-	(275)	-	-	-	-	(182)	(184)	(276)	(182)
Agreeableness (A)	(277)	(267)	-	(233)	-	-	-	-	(278)	(279)	(280)	(281)
Neuroticism (N)	(277)	(282)	-	(283)	-	-	(284)	-	(266)	(266)	(233)	(281)

OCEAN↓ VQuality→	Pre	Pul	Res	Shim	Stra	Stro	Tre	Twa	Ven	Wob	Yaw	Lou
Openness (O)	(285)	(286)	(287)	-	(19)	-	-	(288)	-	-	(289)	(290)
Conscientiousness (C)	(291)	(292)	(19)	(281)	(293)	-	(267)	(294)	-	(295)	(289)	(19)
Extraversion (E)	(296)	(286)	(182)	-	(274)	-	-	(297)	-	-	(182)	(182)
Agreeableness (A)	(285)	(298)	(287)	-	(280)	-	-	(297)	-	-	(289)	(12)
Neuroticism (N)	(269)	(298)	(287)	(281)	(298)	-	(299)	(288)	-	(295)	(300)	(12)

TABLE 9.3: Color Palette for the Correlations Table

Correlation	Color	Weight
Strong Negative (SN)		1
Negative (N)		0.75
Weak Negative (WN)		0.25
Neutral (N)	-	0
Weak Positive (WP)		0.25
Positive (P)		0.75
Strong Positive (SP)		1
Inconclusive (IC)		0

relation between vq_i and llf_j is a positive correlation, it is -1 if the correlation is negative, else it is 0. The $w_{i,j}$ is the weight assigned to the correlation. μ_{llf_j} and σ_{llf_j} are the statistics of the llf_j calculated as explained later in section **??**.

Table 9.2 shows the correlations between the 24 voice qualities and the 5 OCEAN traits. To illustrate an example of one of the OCEAN traits, lets look at Openness. Openness is negatively correlated with voice quality of coveredness, aphonicity, nasality, jitter, pulsed, strained and yawniness. Openness is positively correlated with biphonicity, breathiness, pressed, resonant, twanginess and loudness. There are some voice qualities, for which we do not find prior literature and hence and are not cited.

9.4 Experimental Validation and Results

The above formulae are extracted from prior literature, however validation of the formulae is highly important for their correctness and usefulness. Figure 9.1 presents the overview of the research questions that we ask in this paper.

(RQ1) Do the voice qualities, extracted from the formulae, agree with human perception of the voice qualities, i.e., do the formulae ranking of two utterances for a voice quality agree with human ranking?

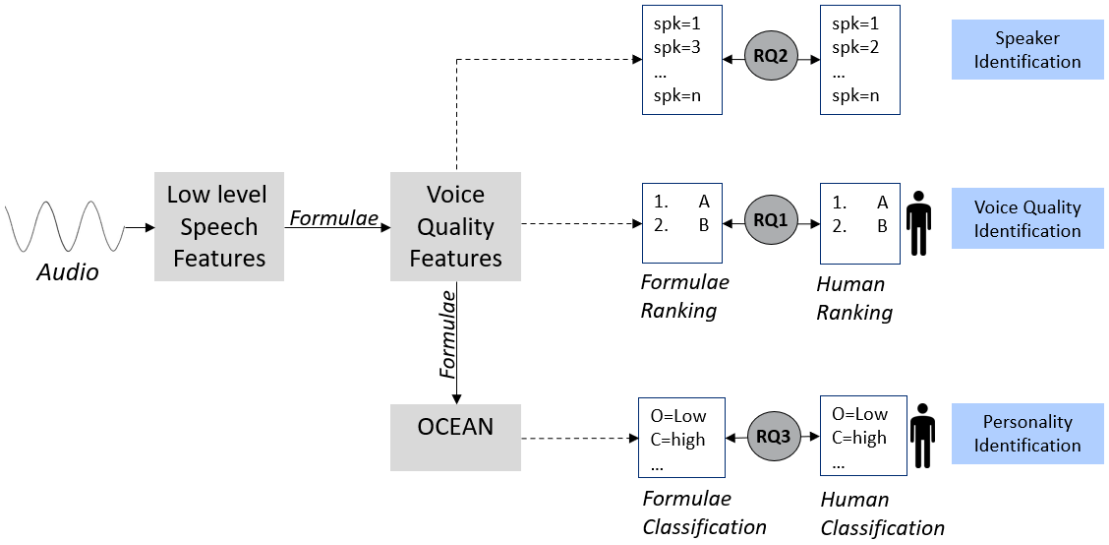


FIGURE 9.1: Overview of the research questions

(RQ2) Are the voice quality features useful for downstream tasks like speaker identification?

(RQ3) Do the OCEAN personality formulae agree with the OCEAN ranking assigned by raters?

In order to address the above RQ's, we perform the following tasks and analysis:

- (RQ1) *Ranking of voice qualities*: The voice quality features highlight the correct voice qualities in a given speech sample. For this purpose, we make the assumption that a given speaker would have one most-prominent voice quality, e.g. the nasal voice quality would have a higher score for the speech sample as compare to the other voice qualities. We rank a pair of audio samples A and B based on this scoring. We then ask raters to rank the pair A and B as per their perceived voice quality. We calculate how many times do the human raters agree with the automatic scoring by the formulae. Section 9.4.1 provides the details of how this task is performed.

- (RQ2) *Speaker identification using voice quality formulae*: We assess the usefulness of the voice quality formulae by analyzing their performance on a close set speaker identification task. We use the Librispeech clean dataset and analyze which voice qualities have the most distinctive power for speaker identification. Section 9.4.2 provides details about this task.

- (RQ3) *Personality Classification using OCEAN formulae*: We employ the OCEAN trait formulae for downstream personality analysis. We use the SSPNet speaker trait corpus, which is rated for the OCEAN traits by human raters. We then compare how the OCEAN formulae ranking compares with the human rankings. Section 9.4.3 provides details for this task.

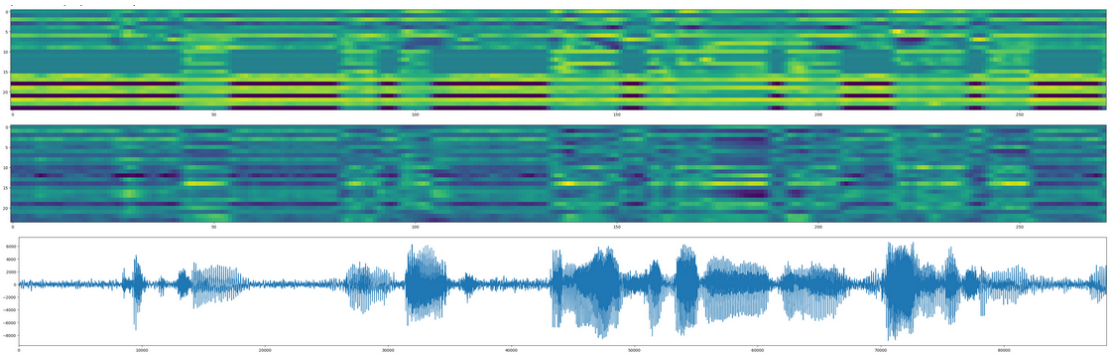


FIGURE 9.2: Top: The Low Level Feature for a speech signal. Mid: Voice Quality Feature for the same speech signal. Bottom: The speech signal.

9.4.1 Ranking of voice qualities

9.4.1.1 Methodology

To extract the low level signal characteristics, we use the Temporal Acoustic Parameter (TAP) model (313). The model is trained on speech data and is robust in predicting the low level features. For this purpose we use the VoxCeleb dataset (314). Once the low level features are extracted, we extract the statistics (mean and variance) of the multiple low level speech features. These statistics are then used for the formulae as explained in the previous sections. We then extract the voice quality features, a 24 dimensional vector per utterance.

For each voice quality, we rank the audios in the VoxCeleb dataset for that voice quality. We select the audios which rank the highest for that voice quality against the ones which rank lowest for that voice quality and form pairs in such a way. While forming pairs we make sure that sex of the speaker is the same so that it does not influence the listener's judgement of the voice quality. For example for voice quality of nasality, we form pairs among the top 5 samples ranking highest for nasality randomly with the 5 samples which rank lowest for nasality. These 5 pairs are then given to raters, who are provided with the following 4 choices: 1. Speech sample A is more nasal than B, 2. Speech sample B is more nasal than A, 3. They are both equally nasal, 4. They both are not nasal. The listener are also provided with a representative speech sample for that voice quality, i.e a speech sample which has been determined to be highly nasal by two experts in the lab. These representative speech sample primes the listener as to what that particular voice quality sounds like.

9.4.1.2 Results

Each voice quality questionnaire has 5 questions and is answered by 20 people. A majority vote is taken as the human rating for that voice quality. For each of the 5 questions, we then note how many pairs agree with the formulae ranking.

Table 9.4: Human judgment alignment with the formulaic predictions when accounting for listener sex

Voice Quality	Correct		
	Listener=Female	Listener=Male	Overall
Creakiness (f=1, m=4)	5	4	5
Twanginess (f=4, m=1)	4	5	5
Hoarseness (f=0,m=5)	5	5	5
Strohbassness (f=0,m=5)	4	4	4
Nasality (f=3, m=2)	3	3	4
Coveredness (f=2, m=3)	4	5	4
Glottalization (f=2, m=3)	0	3	1
Breathiness (f=4, m=1)	1	1	1
Pulsed (f=0,m=5)	0	3	0

Table 9.5: Examining the percentage accuracy across varying conditions of speaker and listener sex.

	Listener	
Speaker	Male	Female
Male	31.5	33.4
Female	12.9	11.8

Table 9.4 shows the overall correct pairs out of 5. It can be observed that voice quality creakiness, twanginess, hoarseness are perceived quite well and all 5 pairs agree with the formulae ranking. For the voice quality strohbassness, nasality, coveredness 4 out of 5 agree with formulae ranking. Voice qualities of glottalization, breathiness, and pulsed have a low score of 1 out 5 and 0 out of 5 respectively, and hence human ranking does not agree with the formulae ranking.

We also notice that for male speakers, more female listeners get the ranking correct and vice versa, i.e for female speakers, more male listeners get the ranking correct. The exact percentage is shown in table 9.5.

A secondary analyses was run by giving the listener a 5 second of audio and the entire 10 sec of audio; where the 5 second of audio is either the first half, the second half and a random 5 second chunk of the audio. We then run the human ranking test again for each. We run this test for the voice quality of nasality, breathy and creaky. The results are shown in table **??**. We can observe that the voice quality is not consistently expressed in a speech sample and raters are not consistent in the first or the second half of the audios. A random chunk of sample leads to better results for creaky voice quality. We can observe that longer speech sample does not necessarily lead to a higher correctness, infact longer samples are actually worse than the most correct shorter 5 second segment.

Table 9.6: Human judgment alignment with the formulaic predictions when accounting for listener sex when audio length is ≈ 5 sec (short) vs ≈ 10 sec (long).

Voice Quality	Correct		
	Listener=Female	Listener=Male	Overall
Nasality (first half)	2	2	4
Nasality (second half)	0	2	2
Nasality (random chunk)	3	3	4
Nasality (long)	2	2	1
Breathy (first half)	1	0	0
Breathy (second half)	1	2	1
Breathy (random chunk)	1	1	1
Breathy (long)	0	0	0
Creaky (first half)	4	4	4
Creaky (second half)	4	4	4
Creaky (random chunk)	5	4	5
Creaky (long)	5	5	5

9.4.2 Speaker Identification using voice quality formulae

9.4.2.1 Methodology

To assess the usefulness of the voice quality formulae, we hypothesize that voice quality features capture speaker specific characteristics. For example is the speaker voice is predominantly nasal, then the nasal voice quality would capture this. In addition, the pattern of nasality would also be specific to each speaker, therefore proving useful for the speaker identification task. For this purpose, we conduct a close set speaker identification task. We use the voice quality extraction in a windowed fashion, where we end up with features as shown in figure 9.2. For this task, we use the Librispeech dataset (315) and just the clean subset of the dataset.

9.4.2.2 Dataset

The Librispeech clean subset, which is a collection of speakers reading audio books. There are a total of 22831 audio files, 251 speakers and the average audio recording is 4 seconds long. For each speaker, we use 80% of the audio files in the training set and the rest 20% in the test set.

9.4.2.3 Results

Left part of figure 9.3 shows how the voice quality features perform on the test set of Librispeech clean. Around 6th epoch the model is able to achieve above 90% F1 score.

To analyze further as to which voice quality features are most useful for speaker identification task, we train the model 24 times, each time training on a different voice quality feature. Right part of figure 9.3 shows the bar chart demonstrating F1 for each voice quality. We can

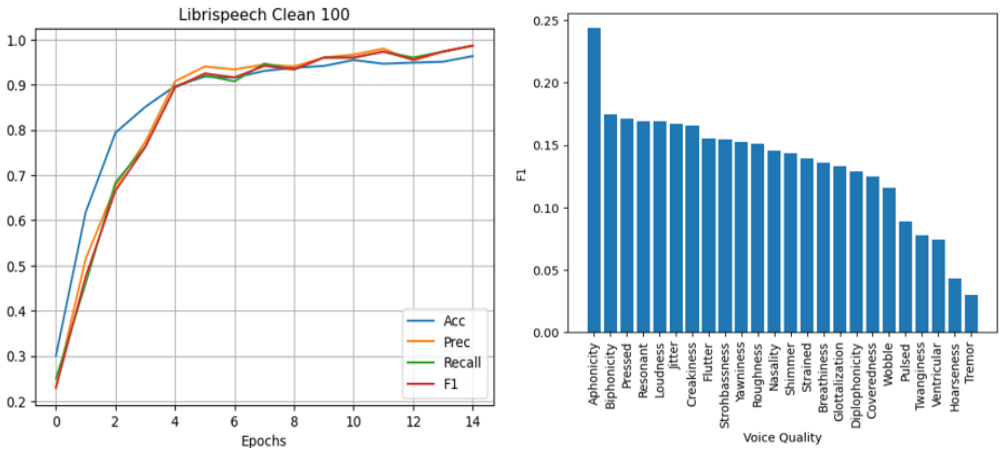


FIGURE 9.3: [Left] Testing metrics on close set speaker Identification on the Librispeech Clean subset. [Right] F1 metric when a single voice quality feature is used for model training for the task of close-set speaker identification on Librispeech clean subset.

observe that Aphonicity voice quality feature is the most distinctive and results in the highest performance, leading to 24.35% F1 metric, followed by Biphonicity which results in 17.49% F1. The reason that aphonicity and biphonicty voice qualities features are performing well is because they fluctuate the most as a time series and this trend captures speaker specific patterns

9.4.3 Personality classification using OCEAN formulae

9.4.3.1 Methodology

In order to validate the formulae presented in the table 9.2, we assess their usefulness in downstream personality prediction task. We perform the task as a binary classification task for predicting low or high OCEAN trait.

9.4.3.2 Dataset

We use the SSPNet Speaker Personality Corpus dataset (28) for personality analysis. SSPNet is a collection of 640 utterances of french news broadcasts and each audio file has been labeled by 11 raters for the OCEAN traits. The audio files are on average 10 seconds long. The number of speakers in the dataset are 322. We use this dataset to evaluate the performance of the proposed voice quality and OCEAN formulae on the task.

9.4.3.3 Results

We extract the OCEAN traits for the SSPNet dataset using our proposed OCEAN formulae, and perform the binary classification task. We also train a neural network model on the proposed voice quality formulae (extracted similarly as done in the speaker ID task) and perform the

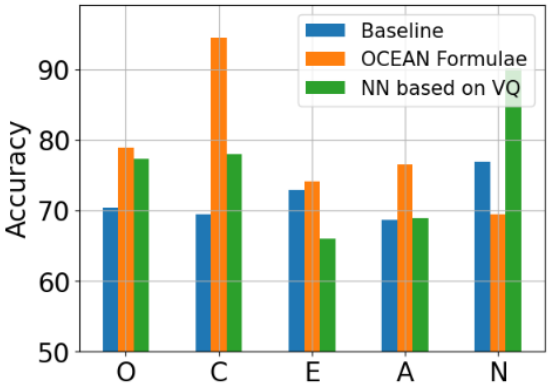


FIGURE 9.4: Accuracy for personality OCEAN traits on the SSP-Net dataset.

OCEAN classification task. We also compare with a baseline model (189), which proposes to use the frame level low level speech features from opensmile toolkit. They train a neural auto encoder model. For the second neural network model, the architecture is as follows: it is a Convolutional Neural Network model followed by a linear layer, ReLU activation and batch-norm layer, followed by a average pooling to convert the output of the convolutions into a 64 dimensional vector which is then passed through a linear layer and then softmax to predict the output into two classes.

Figure 9.4 shows the performance of the proposed voice quality and OCEAN formulae on the OCEAN trait prediction task.

We observe that OCEAN formulae performance is comparable to the baseline and the voice quality formulae's performance. The OCEAN formulae outperforms in Openness, Conscientiousness, Agreeableness. Whereas voice quality formulae outperforms greatly on Neuroticism. In Extraversion, however the baseline model performs the best compared to others. On average, on all the traits, the OCEAN formulae achieve the best performance, 78.74%, followed by 76.08% and then by the baseline model 70.42%.

9.5 Conclusion

One limitation of the current work is the assumption that the relationship between voice quality and low level feature and secondly between OCEAN and voice quality is linear. The relationship could in fact be more complicated, i.e. non-linear. Future work in this direction includes exploring this more complex relationship with the help of neural network.

In conclusion, in this work we proposed formulation for extracting 24 different voice qualities from low level speech signal properties and henceforth extracting 5 OCEAN traits using the 24 voice qualities. This is the first work, to the best of our knowledge, which gives objective methodology for extracting voice qualities and OCEAN traits instead of the traditional way of subjectively describing them. We test the usefulness and correctness of the objective values of the voice qualities and OCEAN traits in multiple tasks. Firstly, we assess how well do human

ranking agree with the voice quality formulae ranking on pairs of speech samples. We launch this experiment for various voice qualities, and find that on average 3.1 / 5 pairs humans agree with the formulae ranking. Secondly, we assess how well do the voice quality formulae perform on the speaker identification task and find that on Librispeech clean subset, on a close set speaker identification task, voice quality formulae can lead to 94.9% accuracy. Thirdly, we perform personality binary classification task for each of the OCEAN traits on the SSPNet personality corpus. We find that on average the OCEAN formulae achieves the best performance of 78.74% compared to the 76.08% achieved by the voice quality formulae and 70.42% by the baseline method.

Going forward, we change the focus on representing strategies for personality. Widely personality has been represented on the Big 5 traits, the OCEAN. This representation has been reached at by analyzing the different ways in which humans describe other or themselves. However there is little to no work done in order to understand the bases using newer techniques. Understanding the personality bases impacts the way data is collected, is more in line with the personality perception by humans and in effect impact the data or knowledge based personality modeling. In the following chapter, we focus on how current personality representations are reached at and if the newer techniques agree with traditional understanding.

Chapter 10

Representing Personalities - Revisiting Personality Bases

Personality has been represented in numerous ways. As stated in Chapter 4.1, these are various theoretical models, with either 3, 4, 5 etc number of dimensions (bases) on which personality is defined. Identifying the correct number of dimensions is important as it impacts all downstream tasks performed on such data. The work done in previous chapter also demonstrates how important the number of bases on which personality is defined is.

10.1 Introduction

Understanding personality is important for numerous fields, like behavioral science (316), sociology (317), human-computer interaction (318; 319; 320). The question of what constitutes 'Personality' has garnered a lot of interest from various fields. Researchers have tried to understand how to classify individuals into different personalities (321), how many unique personalities there are (a good comprehensive summary of which can be found here (322)), what are the different dimensions that each individual should be classified over, and how should each dimension be assessed (322). There are numerous propositions for each. One of the most accepted theories is the Big-5 personality trait theory (172; 173), also called the the OCEAN traits. OCEAN stands for Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. However there are many other theories as well like the NEO Personality Inventory-Revised (NEO PI-R) (176), Three Dimensions of Personality (PEN) (177), Myers-Briggs Type Indicator (MBTI) (178) etc, four personality types (180), etc.

Understanding personality is useful for developing technologies that are used for human-computer interaction. At the least, human interactions with machines can become more natural if computers are able to understanding personality as humans do. Such technologies can also aid in the assessment and monitoring of mental health and psychological problems in humans (323; 324), helping all involved – healthcare providers and affected people – in positive ways.

They can also help predict the long-term susceptibility of individuals to specific types of work, social situations, and other factors (325).

One way to understanding personality is to identify unique bases across which an individual's personality could be described. For example, some works state 5 bases; i.e the OCEAN traits mentioned earlier. Others state 4 bases (180), and some 3 (177). Infact in earlier works, there was so much a large range of opinions (326), where Cattell (1973) (327) held the opinion that the number of bases were as high as 16, while Eysenck (1976) (328) and Wiggins (329) believed that the number is as low as 2 or 3, the focus being on extraversion and neuroticism being the most important bases. The question arises as to why are the opinions so diverse, how have these works landed on these number of bases and is there a way to evaluate how close we are to the true number of bases? We go into detail on how the 5 widely accepted bases have been identified in Section **??**.

The methodology of identification of these unique bases can be roughly described as follows: (1) A list of all possible English words are shortlisted which can be used for describing human's personality traits, (2) Individuals are either asked to perform self-ratings, where they describe themselves using a subset of these words, or they describe others using a subset of words, (3) These ratings are then analyzed using factor extraction methods like principle-component, varimax and oblique algorithm to reduce the number of dimensions that are most used in discriminating individuals. In this framework, there are a sources of uncertainty. Firstly, the involvement of human ratings brings variability of opinions, biases like individual biases, cultural biases, language differences, etc. Secondly, there is little understanding of the differences that arise from self ratings versus ratings other individuals.

In this paper, we propose new methodology to verify the already established personality bases, understand how many 'number' of bases are optimal. We analyse that standalone list of words from previously proposed literature and we use various Large Language Models (LLMs) to extract their word embeddings and cluster them. We make observations on the optimal number of clustering of these words. Secondly, we hypothesize that words in context would provide more information about how they used to describe people/things. We use the Project Gutenberg (330) and use this dataset to extract sentences in which a collection of personality words are used. We examine the sentence embeddings from these sentences and make inferences on these.

We propose a metric called cluster entropy to measure the goodness of the cluster sizes. We find that the cluster entropy is lowest when the cluster size is 2, followed by 5. We also identify the representative words for different cluster sizes and observe that when the cluster size equals 2, each LLM model selects words from one of the OCEAN bases. When the cluster size equals 5, there is a word from each of the 5 OCEAN bases.

10.2 Background work

One of the widely accepted ways of representing personality is using the OCEAN personality bases. OCEAN stands for Openness, Conscientiousness, Extraversion, Agreeableness and Neuroticism. But where do these 5 bases come from? In this section we explore the origin of these 5 bases.

Originally, Allport and Odbert in 1936 (331) did a landmark lexical study of listing all the different words from the English dictionary that can be used for describing humans. They made a collection of 18000 such words (331). Cattell reduced the extensive 18 K list into 4500 trait terms and later reduced it to 35 terms by eliminating different words like synonyms. This allowed him to perform data analytic techniques on the terms like factor analysis (172; 173). He identified 12 personality factors which he also included in the 16 personality factor questionnaire proposed in order to assess an individual's personality (332).

Norman in 1967 (333) also worked on the original 18K words from Allport and Odbert and reduced these list of 18K words to just 2800 words in various phases (333). Firstly from the initial master pool of 18K words, they manually curtailed words. For example, they removed words that are not used in contemporary discourse, rather pertain to obscure literary, historical or mythological referents; and they removed words which are broader terms denoting various anatomical or physiognomic characteristics. Further categorizing on these words lead to a shorter word list of 2800 terms only. These traits were judged to reflect stable, "biophysical" traits, which were further divided into 14 lists of 200 terms each. Later Norman (334) reduced this list into 75 smaller semantic categories. He did this based on his understanding of the similarities in words' meaning. In making his classifications, Norman began by sorting the terms into a few broad categories, and then he subsequently developed more fine-grained classifications within each of the initial categories.

Goldberg in 1990 (335; 24) worked on the list of 2800 terms from Norman and used a subset of 1710 words to be included in a self-report inventory of trait-descriptive adjectives. His aim was to uncover the underlying dimensions in the terms. For this purpose, he employed a number of students to describe themselves using the terms. Under instructions to work on this task for no more than an hour at a time, 187 college students (70 men and 117 women) described themselves on each of the 1,710 terms, using an 8-step rating scale ranging from extremely inaccurate to extremely accurate as a self-descriptor. Factor extraction methodologies are then performed on these ratings. like principal-components, principal-factors, alpha-factoring, image factoring, and maximum-likelihood procedures. In a a similar manner, the 75 Norman categories were analyzed using a wide variety of such procedures. Goldberg identifies the five-factor space in English.

Such procedure have repeatedly in various studies led to establishing the five factors. For example, Tupes and Christal (174) found five relatively strong factors. This five factors have been replicated by Norman (1963) (334) , Borgatte (336) , and Digman and Takemoto-Chock

(337) and Goldberg (24). Since these initial landmark studies in the 90's, these bases have been widely accepted and there has been little exploration of the bases using newer methodologies. A high level timeline of work on discovering personality bases in shown in Figure 10.1

The Big-Five personality bases are widely accepted, however there are other competing hypothesis as well. The studies presented in Section **??** falls under the psycho-lexical hypothesis. The lexical hypothesis assumes that our usage of language and words to describe humans would reveal the underlying bases of the personality. However opponents to this hypothesis suggested that the complexity in personality could not be captured by just words alone, and entire sentences have to be considered. There are other hypothesis like the Five Factor Model (Costa and McCrae), which propose a different methodology to understand personality. Costa and McCrae (338) who suggested the NEO personality model, which uses longer sentences for rating individuals. In connection to this is the BFI-10 questionnaire which assesses people on these extensive NEO traits.

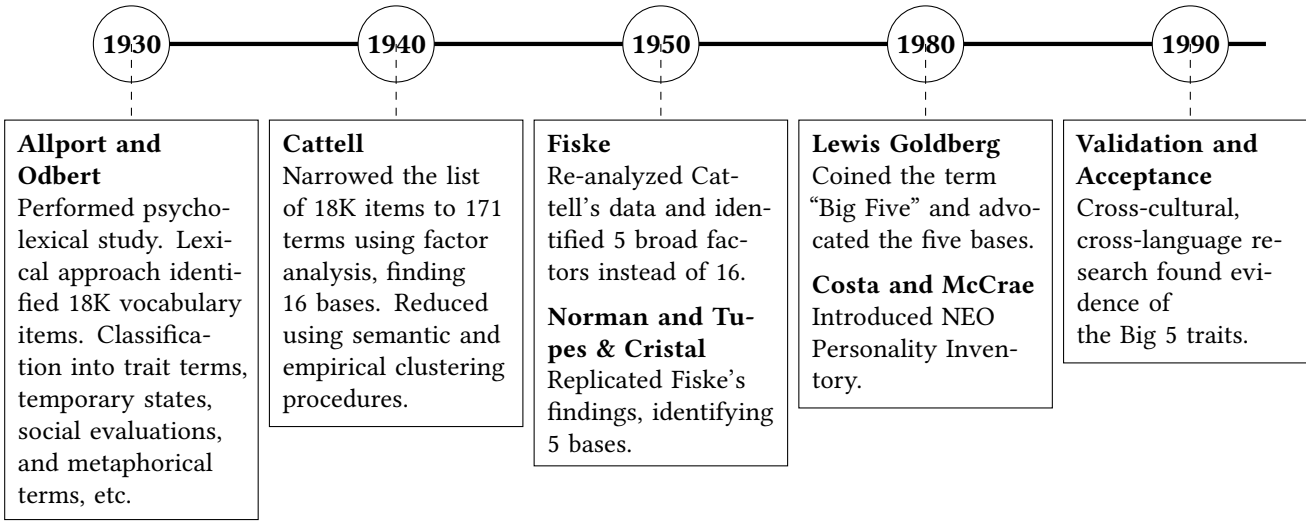


FIGURE 10.1: Timeline of the Development of the Big Five Personality Bases

10.3 Proposed Approach

10.3.1 Dataset

In our to perform our experiment, we require a collection of text where individuals have been described in diverse manners. For this purpose we use the Project Gutenberg (PG) dataset, founded in 1971. It is a library of over $70,000$ free eBooks, and more than $3 \times 10e9$ word-tokens (339).

TABLE 10.1: Some example sentences. The top part of the table presents sentences for the word 'talkativeness', where the lower half shows examples for the word 'silence'.

1. During lunch both girls were struck by his unusual *talkativeness*.
2. You must forgive my *talkativeness*; I am hot upon this subject and forget that others may grow weary of it.
3. The *talkativeness* of these Europeans [...]
4. He spoke also of the *talkativeness* of his Katie.
5. Often there is restlessness, *talkativeness*, indifference, carelessness and disturbances of volition.
1. There was, moreover, great *silence* in the chamber.
2. [...] she, thinking herself snubbed by his *silence* after her avowal, grew hot and uncomfortable.
3. Our God shall come, and shall not keep *silence*;
4. After a long and gloomy *silence* he spoke again.
5. You needn't be afraid; I daren't say more; my *silence* is prepaid.
6. Francis walked on for some moments in *silence*.

10.3.2 Analysis

In order to perform our analysis, we use the 75 words reduced by Norman (334). For each word, we perform two different experiments, where we extract individual word representations from different Large Language Models (LLMs). Secondly, we extract sentences from the PG dataset and similarly extract sentence representation from the LLMs. Below we go into detail about each category of experimentation.

Individual Word Analysis:
We use the 75 words reduced by Norman (334) (note that these 75 words are coming from the original 18K word list from Allport and Odbert, as described in Section **??**) [1]. To extract the embeddings of the words, we use three different Large Language Models: BERT (112), Llama (340), OPT (341) and T5 (342). BERT is an encoder-only model, Llama and OPT are a decoder-only models and T5 is a encoder-decoder model. We cluster these embeddings using KMeans using different number of clusters.

Sentence Analysis:
Instead of doing the analysis for individual word embeddings, it is important to consider the context in which they occur. For this purpose, we use the Project Gutenberg collection of books. The motivation is that books would contain many instances of words used in describing people, things, or circumstances. These description would be more useful in understanding the usage of such words and how similarly/dis-similarly these words are used as.

We again use the 75 word list as above from Norman as above. For each of the word in this list, we identify $10,000$ sentences in PG where these words occur. We extract word embeddings

[1] The list of all the words can be found in the GitHub repository here: https://github.com/ydhira/personality_bases/blob/main/norman_75.txt

using the three LLMs; Bert, Llama and OPT. Note that for a word that occurs in two sentences, the word embedding extracted from these two sentences would be different since the embedding is influenced by the context (what words occur before and after).

We cluster the word embeddings into different number of clusters and calculate the metric that we call: 'cluster-entropy' (CE). CE measures the entropy of all the word embeddings for a single word belonging to the same cluster. In detail, lets suppose for word w, from sentence k, whose embedding is denoted as e_{wk}, belongs to cluster c_i where i goes from $1 \cdots N$ where N is the total number of clusters. Belonging to a cluster is determined by the minimum euclidean distance of e_{wk} to the cluster center. For all the word embeddings of word w represented as e_w, we calculate the clusters that it belongs to, and from this calculate the probability of clusters, simply by counting the times that c_i appears in that cluster. The entropy of this probability vector represents the cluster-entropy for word w. An average over all the words, represents overall cluster entropy. It could be written as:

$$\text{CE} = \mathbb{E}_w[\text{H}(p_{wi})]$$

where p_{wi} is the probability of word embedding of word w belonging to cluster i

Using NN for reducing cluster entropy:
Instead of using KMeans for clustering, we experiment with using Neural networks for the task of clustering. We have multiple objective functions to optimize for: 1. Entropy of probability of an instance belonging to a cluster should be low 2. Entropy of the probability of all the sentence embeddings for the same word belonging to the same cluster should be low 3. Entropy of the probability of all the instances belonging to each cluster

Formally, to define the loss function, let the batchsize be N, the output logit be l_i where i goes from $1 \cdot N$, and let K be the total number of words.

$$L_1(\Phi) = \frac{1}{N}\sum_{i=1}^{N} \text{H}(\text{softmax}(l_i))$$

$$L_2(\Phi) = \frac{1}{K}\sum_{k=1}^{K} \text{softmax}(l_i \in w_k)$$

$$L_3(\Phi) = \text{H}(\frac{1}{N}\sum_{i=1}^{N} \text{softmax}(l_i))$$

$$L(\Phi) = w_1(L_1(\Phi) + w_2 L_2(\Phi) - w_3 L_3(\Phi)) \tag{10.1}$$

$$\hat{\Phi} = \arg\min_{\Phi} w_1(L_1(\Phi) + w_2 L_2(\Phi) - w_3 L_3(\Phi)) \tag{10.2}$$

where w_1, w_2, and w_3 are weights for combining the losses.

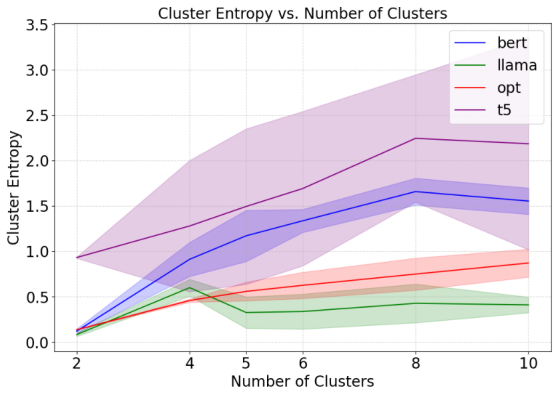


FIGURE 10.2: Cluster Entropy across the different configs for each LLM.

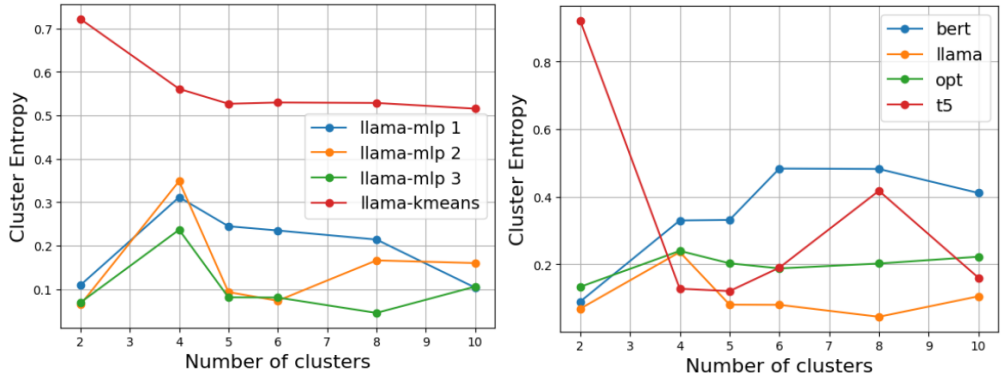


FIGURE 10.3: [Left] Cluster entropy using different number of clusters using the Llama LLM. [Right] Cluster entropy comparing the different LLM Models and Kmeans model.

We compare the cluster entropy when clustered using KMeans versus Neural Networks. We discuss the results in the following section.

Cluster Representative Word:
For each cluster, we extract one representative word for that cluster. We do this by picking the word which appears the most in a cluster, normalized by the word frequency. We compare the representative word for when KMeans is used for clustering versus when Neural Networks are used for clustering.

10.4 Experimental Validation and Results

10.4.1 Cluster Entropy

Figure 10.3 shows the cluster entropy achieved when the number of clusters are set to 2,4,5,6,8,10 and using embeddings extracted from the llama model. We experiment with three different weight settings as shown in equation (1), which are labeled as mlp 1, mlp 2, and mlp 3 in figure 10.3. We can observe the neural networks are able to perform better clustering than KMeans is able to, by observing that the cluster entropy is lower for mlp models compared to the kmeans

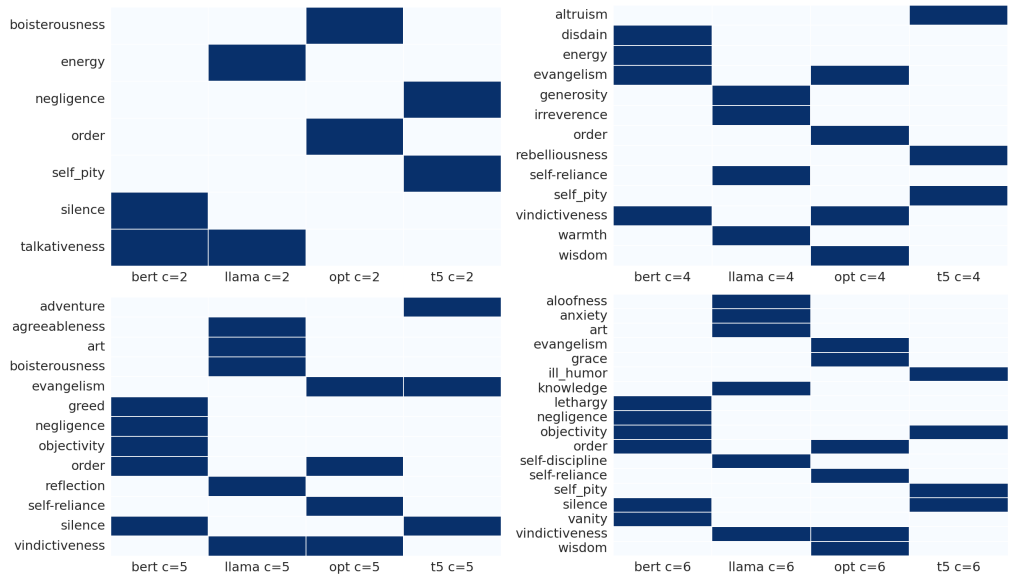


FIGURE 10.4: Representative words from the different models using different number of clusters (c). E.g. the number of cluster is shown as c=2, when the number of clusters=2.

model. We can also observe that the trend of the cluster entropy is similar across the mlp models, where the lowest cluster entropy is when the number of clusters are 2, 5, 6 and 8.

The lower part of Figure 10.3 shows the cluster entropy for the four different LLM used. We can observe that llama embeddings lead to better cluster entropy compared to the other models, we can also observe that the consistently the cluster entropy is lowest or similar to the previous cluster number is 5, otherwise the cluster entropy is not consistent at a given cluster number across the LLMs. This points at the stability of when the number of clusters are set at 5.

10.4.2 Cluster Representatives

The identified cluster representative words are shown in figure 10.4. We show the identified words for different number of clusters across the different LLMs. We can observe the level of agreement that different LLMs have for each value of c. We will go into depths of two scenarios: when the number of clusters is 2 and when it is 5.

Two Clusters

When the number of clusters is set at two, the two cluster center identified by Bert are 'silence' and 'talkativeness'. These dimension has also been observed previously has the dimension with the most discriminative power in the factor analysis methodologies, reported in (174). Another interpretation of this finding is that these two words point towards the introversion and extroversion category established in the Big-5 trait hypothesis. At a higher level, one big observation is that each model picks out two words in the Big 5 OCEAN traits (24). For example, Bert picks out silence-talkativeness from the Extraversion bases. Llama picks outs Talkativeness-Energy also from Extraversion bases. OPT picks Order-Boisterousness from the

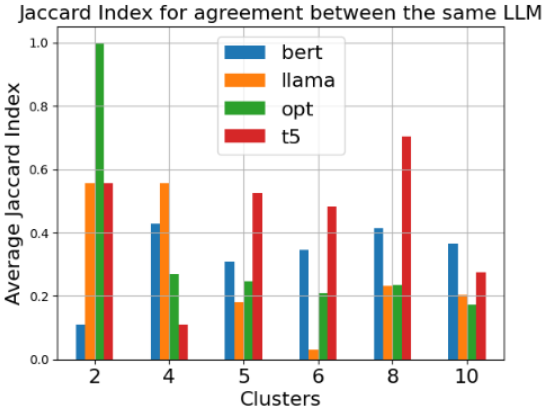


FIGURE 10.5: Bar chart shows the Jaccard Index between representatives words within each LLM as a function of the number of clusters used.

Conscientiousness scale. T5 picks self pity-Negligence from the Neuroticism scale. This is an interesting observation since individually these bases seem to be the most informative when it comes to dividing personality into two clusters.

Five Clusters

When the number of clusters is set at five, it can be observed that each representative words comes from one of the OCEAN bases as listed in (24), especially in Bert and Llama. For Bert, silence is from Extraversion bases, Order is from Conscientiousness bases, Objectivity is Openness, Negligence is from Neuroticism and Greed is from Agreeableness. For Llama, Vindictiveness is from Neuroticism, Reflection is from Extraversion, Boisterousness is from Conscientiousness, Art is from Openness and Agreeableness is from Agreeableness. For OPT and T5, there are less than 5 words highlighted because the same word is picked as a representative word for more than one cluster, therefore there is an overlap. For OPT, vindictiveness is Neuroticism, self-reliance is Openness, order is Conscientiousness, evangelism is Extraversion. For model T5, self-reliance is Openness, evangelism and adventure is Extraversion.

Agreement Within the Same LLMs

In order to assess how often do different configurations of the LLMs results in the same representative words, we calculate Jaccard Index of the representative words. Jaccard Index (JI), measures the similarity between two finite sets. It is defined as the size of the intersection divided by the size of the union of the sample sets. Figure 10.5 shows the bar chart of average jaccard index over different configurations for each LLMs over the different number of clusters. We can observe that when the cluster=2, OPT model results in JI equals to 1. LLama results in higher JI when clusters=4, T5 results in high JI when clusters=5,6, and 8, and Bert results in high JI when clusters=10.

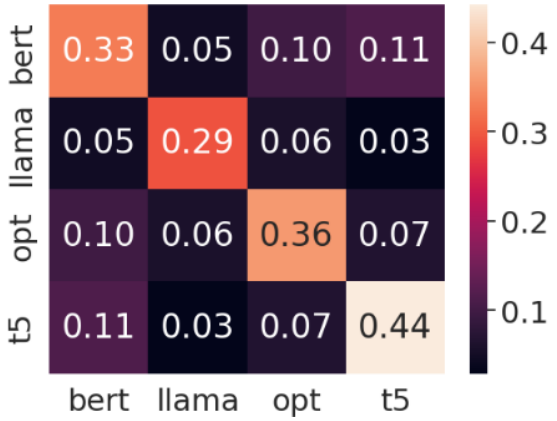


FIGURE 10.6: Heatmap shows that average Jaccard Index between the different LLMs.

Agreement Between Different LLMs

To assess the level of agreement in between the different LLMs, we calculate Jaccard Index over all the different configurations and over all the clusters sizes. Figure 10.6 shows the heatmap of the average JI in between the different LLMs. We can observe that on average T5 has the highest JI, followed by OPT, Bert and then Llama. Bert and T5 tend to agree more, followed by Bert and OPT and then OPT and T5.

10.5 Conclusion

In this work, our aim is to uncover the main bases across which personality could be best represented in. Traditionally personality is widely represented on the OCEAN traits, which had been established in the 1900's and the field has not been revisited recently. In this work, we aim to utilize newer methodologies to re-analyze the lexical bases on top of which the OCEAN traits are based on. We utilize language models, including Bert, Llama, OPT and T5 to extract context representations of words identified in earlier works for personality analyses. Using these higher dimension representations, we first identify the metric called cluster-entropy based on the purity of a cluster-size which best clusters the lexical bases. When the cluster size is 4, considering all the 4 LLMs, we observe that the variance in cluster entropy is the smallest, followed closely by cluster size of 5. We also identify representative word for each cluster, and observe that interestingly the representative word comes from each of the 5 OCEAN bases when cluster size is 5. In conclusion, according to the cluster entropy metric, bases 4 an 5 are the most optimum number where most models.

Part III

Chapter 11

Conclusion and Suggested Future Directions

11.1 Thesis Conclusion

This thesis delves into the topic of understanding psychological traits from human speech, focusing on emotion and personality. We divide the pipeline of understanding psychological traits into three components. The first part is the 'encoding process' - where the psychological trait encodes into various signals including the speech signal, which is then communicated. The second part is the 'decoding process' which involves how the features encoded in the signals (i.e the speech signal) can be decoded from it and optimally used to understand the trait. Thirdly the 'representation process' - which tries to entangle the problem of representing the psychological traits.

The three constituents of the understanding pipeline are dealt individually for both emotion and personality. Firstly for encoding emotions, we delve deeper into how the encoding process in humans differ when expressing natural versus acted emotion. We perform this study by focusing on the phonetic bases of the vocal expression. Phonetic bases of emotion comprise 'what' phonemes are used and the 'manner' they are delivered in to express the emotion. We find statistically significant differences in the expression of phonemes under acted and natural expression. We find that fricatives, stops and nasal are more important in natural speech than in acted speech for classifying for emotion. We also find that vowels are more important in acted speech than in natural speech.

Secondly for decoding emotions from speech we delve deeper into how the specific features in speech carry information about emotions. There are an array of speech features, including acoustic, prosodic features that have been studied for the task, however one very important feature, the 'cadence' of speech has not been. That is because it is difficult to quantify cadence and secondly to utilize it in a methodical way in determining emotions. We present a way in which cadence of speech signal can be used methodologically and in addition cadence of

other modalities like the word sequence and phoneme sequence, can be used for understanding emotion. Our experimentation analysis reveal that cadence is a very beneficial feature to utilize in understanding emotion and boosts performance significantly leading to a 5.9% improvement in IEMOCAP and 7.5% improvement in CMU-MOSI. We also find that the cadence of each modality is different from each other and need to be modeled separately. Most importantly, our findings suggest that if the cadence is modeled well, it eliminates the need to align modalities together.

Thirdly, we focus on ways in which emotion representation can be improved for their understanding. Emotions have traditionally been modeled as discrete (angry, happy, sad), or in three dimensional continuous space representing valence, activation and dominance (VAD). In our first study, we focus on how these two views of emotions could be combined together to utilize their complementary information for better emotion understanding. We propose a hierarchical multitask neural network architecture, where the discrete and continuous emotion representations are predicted in a multi task framework. We find that they do have complementary information and using both views brings benefits in the detection task leading to a 1.2% accuracy improvement in discrete classification and 0.09 CCC improvement for VAD prediction in IEMOCAP. Secondly, we move beyond these traditional views of emotions and use the natural way that humans describe emotions in real life. Emotions can be described in greater detail using the flexibility that natural-language provides. We note that there are multiple acoustic correlates of emotion and use these acoustic features to expand the natural language descriptions of emotion. We use this for training a Contrastive-Language-Audio model and show that indeed using these longer acoustic descriptions, we improve the emotion classification performance on top of using simple emotion labels, leading to 5.6% relative improvement in Ravdess dataset. Furthermore, instead of using explicit acoustic prompts, we allow the model to learn such prompts automatically while training, which is called 'prefix'. Using this strategy, we improve on the emotion classification task under various settings: 'zero shot', and 'few shot' and in-domain, for example this strategy leads to 36.58% relative improvement in Ravdess zero shot setting.

In the second half of this thesis, we delve into understanding personality from speech. Firstly, we focus on what features encoded in the speech could be best used to understanding personality. Due to the lack of labeled speech data for the task and also the low reliability (low inter annotator agreement) of the labeled data, we take a more knowledge base approach. In order to study the 'encoding' of personality in speech, we dive deeper into literature studying effects on voice per personality trait, finding that personalities have a specific effect on the voice, for example high extroverted personalities have more loudness and more jittery voices. This motivated us to hand-craft feature sets for studying personality. We validate our hand crafted features sets collecting human labels on them and on multiple down stream tasks like speaker identification and personality analysis. We find that our handcrafted features lead to better results on certain personality traits like openness, consciousnesses, extraversion and agreeableness.

Secondly we delve a bit deeper into how personality representation can be improved. Traditionally, personality is represented using the Big-5 OCEAN traits. However these five bases have been established a long time ago, with the help of human annotations and low amount of data. With availability of more data, and Large Language Models (LLM), we revisit discovering the bases that are emerge from the understanding that LLMs have. We find that 2 bases are the most informative when grouping people, followed by 4 and then 5. However when focusing on the 5 bases, we find that the model anchors on the identified OCEAN bases, which validates the bases already established.

11.2 Suggested Future Directions

The work done in this thesis has implications in research and development for emotion and personality understanding from speech. I list a few ideas here, in the hopes that who ever is reading this thesis gets inspired to take these ideas forward.

11.2.1 Incorporating emotion and personality models in the physical world

The progress that Artificial Intelligence has made in the recent years has been phenomenal. This progress when incorporated into products that people can interact with is the ultimate test of the potential and progress of AI. One example of this is the advancement in Large Language Models, and the phenomenal adaptation of ChatGPT by users all around the world. I believe that psychological traits like emotion and personality are an integral part of seamless communication between humans and machines. Human computer interaction would need to incorporate the 'encoding', 'decoding' and 'understanding' of the psychological traits, and such models would need to be effective, scalable, reliable and should have low latency.

11.2.2 Psychological trait understanding in the wild

Building on top of the previous point, one of the big requirements to incorporate emotion and personality models in physical objects is the need for it to work over multiple domains. It needs to work irrespective of the culture, age, gender, language etc differences between the users. This also remains one of the most challenging problems in the field. With the advent of foundation models in various fields, and their exhibit of developing emergent properties, I believe that such models would work well for this task as well, and provide a solution for domain adaptation in this very subjective task.

11.2.3 Revisiting the categorization of the psychological traits

We made an attempt in this direction by firstly combining multiple emotion representations, or coming up with natural language descriptions for the emotions, or re evaluating the bases for

personality. However I believe that there is still potential in this direction. Having a single set of defined classes for emotion or personality limits any model's ability to learn generalizable representations, and there has to be a way to incorporate subjectivity in the way the classes are established. This area is relatively new with our work being the first using LLMs or contrastive models for learning new ways of representing emotions and personality, and I believe it is an important one worth exploring.

Bibliography

[1] Yeonggwang Park, Feng Wang, Manuel Díaz-Cádiz, Jennifer M Vojtech, Matti D Groll, and Cara E Stepp, "Vocal fold kinematics and relative fundamental frequency as a function of obstruent type and speaker age," *The Journal of the Acoustical Society of America*, vol. 149, no. 4, pp. 2189–2199, 2021.

[2] Evelyne Van Houtte, Kristiane Van Lierde, and Sofie Claeys, "Pathophysiology and treatment of muscle tension dysphonia: a review of the current knowledge," *Journal of Voice*, vol. 25, no. 2, pp. 202–207, 2011.

[3] Ingo R Titze, "Physiologic and acoustic differences between male and female voices," *The Journal of the Acoustical Society of America*, vol. 85, no. 4, pp. 1699–1707, 1989.

[4] Jack Jiang, Emily Lin, and David G Hanson, "Vocal fold physiology," *Otolaryngologic Clinics of North America*, vol. 33, no. 4, pp. 699–718, 2000.

[5] Sungbok Lee, Serdar Yildirim, Abe Kazemzadeh, and Shrikanth Narayanan, "An articulatory study of emotional speech production," in *Ninth European Conference on Speech Communication and Technology*, 2005.

[6] Donna Erickson, Caroline Menezes, and Akinori Fujino, "Some articulatory measurements of real sadness," in *Eighth International Conference on Spoken Language Processing*, 2004.

[7] Guofeng Ren, Xueying Zhang, and Shufei Duan, "Articulatory-acoustic analyses of mandarin words in emotional context speech for smart campus," *IEEE Access*, vol. 6, pp. 48418–48427, 2018.

[8] Jangwon Kim, Asterios Toutios, Sungbok Lee, and Shrikanth S Narayanan, "Vocal tract shaping of emotional speech," *Computer speech & language*, vol. 64, pp. 101100, 2020.

[9] Vivien C Tartter, "Happy talk: Perceptual and acoustic effects of smiling on speech," *Perception & psychophysics*, vol. 27, pp. 24–27, 1980.

[10] Eva Lasarcyk and Jürgen Trouvain, "Spread lips+ raised larynx+ higher f0= smiled speech?-an articulatory synthesis approach," *Proceedings of ISSP*, pp. 43–48, 2008.

[11] Jeff Pittam, *Voice in social interaction*, vol. 5, Sage, 1994.

[12] Charles D Aronovitch, "The voice of personality: Stereotyped judgments and their relation to voice quality and sex of speaker," *The Journal of social psychology*, vol. 99, no. 2, pp. 207–220, 1976.

[13] Klaus R Scherer, "Vocal affect expression: a review and a model for future research.," *Psychological bulletin*, vol. 99, no. 2, pp. 143, 1986.

[14] Klaus R Scherer, Harvey London, and Jared J Wolf, "The voice of confidence: Paralinguistic cues and audience evaluation," *Journal of Research in Personality*, vol. 7, no. 1, pp. 31–44, 1973.

[15] Edith B Mallory and Virginia R Miller, "A possible basis for the association of voice characteristics and personality traits," *Communications Monographs*, vol. 25, no. 4, pp. 255–260, 1958.

[16] David Andrew Puts, Carolyn R Hodges, Rodrigo A Cárdenas, and Steven JC Gaulin, "Men's voices as dominance signals: vocal fundamental and formant frequencies influence dominance attributions among men," *Evolution and Human Behavior*, vol. 28, no. 5, pp. 340–344, 2007.

[17] Sarah E Wolff and David A Puts, "Vocal masculinity is a robust dominance signal in men," *Behavioral Ecology and Sociobiology*, vol. 64, pp. 1673–1683, 2010.

[18] Carolyn R Hodges-Simeon, Steven JC Gaulin, and David A Puts, "Different vocal parameters predict perceptions of dominance and attractiveness," *Human Nature*, vol. 21, pp. 406–427, 2010.

[19] Tim Polzehl, "Personality in speech," *Assessment and automatic classification*, 2015.

[20] Matthias R Mehl, Samuel D Gosling, and James W Pennebaker, "Personality in its natural habitat: manifestations and implicit folk theories of personality in daily life.," *Journal of personality and social psychology*, vol. 90, no. 5, pp. 862, 2006.

[21] Nelson Roy, Diane M Bless, and Dennis Heisey, "Personality and voice disorders: a multitrait-multidisorder analysis," *Journal of Voice*, vol. 14, no. 4, pp. 521–548, 2000.

[22] Yuanyuan Xin, Jianhui Wu, Zhuxi Yao, Qing Guan, André Aleman, and Yuejia Luo, "The relationship between personality and the response to acute psychological stress," *Scientific reports*, vol. 7, no. 1, pp. 16906, 2017.

[23] Daniel P Skarlicki, Robert Folger, and Paul Tesluk, "Personality as a moderator in the relationship between fairness and retaliation," *Academy of management journal*, vol. 42, no. 1, pp. 100–108, 1999.

[24] Lewis R Goldberg, "The development of markers for the big-five factor structure.," *Psychological assessment*, vol. 4, no. 1, pp. 26, 1992.

[25] Patrick H Raymark, Mark J Schmit, and Robert M Guion, "Identifying potentially useful personality constructs for employee selection," *Personnel Psychology*, vol. 50, no. 3, pp. 723–736, 1997.

[26] Laurence Devillers, Laurence Vidrascu, and Lori Lamel, "Challenges in real-life emotion annotation and machine learning based detection," *Neural Networks*, vol. 18, no. 4, pp. 407–422, 2005.

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

[28] Gelareh Mohammadi and Alessandro Vinciarelli, "Automatic personality perception: Prediction of trait attribution based on prosodic features," *IEEE Transactions on Affective Computing*, vol. 3, no. 3, pp. 273–284, 2012.

[29] Charles Darwin and Phillip Prodger, *The expression of the emotions in man and animals*, Oxford University Press, USA, 1998.

[30] Jessica L Tracy and Daniel Randles, "Four models of basic emotions: A review of ekman and cordaro, izard, levenson, and panksepp and watt," *Emotion review*, vol. 3, no. 4, pp. 397–405, 2011.

[31] Silvan Tomkins and BP Karon, "Effect, imagery, consciousness," 1962.

[32] Robert Plutchik, "A general psychoevolutionary theory of emotion," in *Theories of emotion*, pp. 3–33. Elsevier, 1980.

[33] Paul Ekman, "Facial expressions of emotion: an old controversy and new findings," *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, vol. 335, no. 1273, pp. 63–69, 1992.

[34] Harold Schlosberg, "Three dimensions of emotion.," *Psychological review*, vol. 61, no. 2, pp. 81, 1954.

[35] Prajakta P Dahake, Kailash Shaw, and P Malathi, "Speaker dependent speech emotion recognition using mfcc and support vector machine," in *2016 International Conference on Automatic Control and Dynamic Optimization Techniques (ICACDOT)*. IEEE, 2016, pp. 1080–1084.

[36] Daniel Neiberg, Kjell Elenius, and Kornel Laskowski, "Emotion recognition in spontaneous speech using gmms," in *Ninth international conference on spoken language processing*, 2006.

[37] Fatemeh Daneshfar, Seyed Jahanshah Kabudian, and Abbas Neekabadi, "Speech emotion recognition using hybrid spectral-prosodic features of speech signal/glottal waveform,

metaheuristic-based dimensionality reduction, and gaussian elliptical basis function network classifier," *Applied Acoustics*, vol. 166, pp. 107360, 2020.

[38] Lijiang Chen, Xia Mao, Yuli Xue, and Lee Lung Cheng, "Speech emotion recognition: Features and classification models," *Digital signal processing*, vol. 22, no. 6, pp. 1154–1160, 2012.

[39] Yuanbo Gao, Baobin Li, Ning Wang, and Tingshao Zhu, "Speech emotion recognition using local and global features," in *Brain Informatics: International Conference, BI 2017, Beijing, China, November 16-18, 2017, Proceedings*. Springer, 2017, pp. 3–13.

[40] Leila Kerkeni, Youssef Serrestou, Mohamed Mbarki, Kosai Raoof, Mohamed Ali Mahjoub, and Catherine Cleder, "Automatic speech emotion recognition using machine learning," 2019.

[41] Ashwini Rajasekhar and Malaya Kumar Hota, "A study of speech, speaker and emotion recognition using mel frequency cepstrum coefficients and support vector machines," in *2018 International Conference on Communication and Signal Processing (ICCSP)*. IEEE, 2018, pp. 0114–0118.

[42] Chien Shing Ooi, Kah Phooi Seng, Li-Minn Ang, and Li Wern Chew, "A new approach of audio emotion recognition," *Expert systems with applications*, vol. 41, no. 13, pp. 5858–5869, 2014.

[43] Chengfu Yang, Luping Ji, and Guisong Liu, "Study to speech emotion recognition based on twinssvm," in *2009 Fifth international conference on natural computation*. IEEE, 2009, vol. 2, pp. 312–316.

[44] Fangfang Zhu-Zhou, Roberto Gil-Pita, Joaquín García-Gómez, and Manuel Rosa-Zurera, "Robust multi-scenario speech-based emotion recognition system," *Sensors*, vol. 22, no. 6, pp. 2343, 2022.

[45] Björn Schuller, Ronald Müller, Manfred Lang, and Gerhard Rigoll, "Speaker independent emotion recognition by early fusion of acoustic and linguistic features within ensembles," 2005.

[46] A Milton, S Sharmy Roy, and S Tamil Selvi, "Svm scheme for speech emotion recognition using mfcc feature," *International Journal of Computer Applications*, vol. 69, no. 9, 2013.

[47] Soroosh Mariooryad and Carlos Busso, "Compensating for speaker or lexical variabilities in speech for emotion recognition," *Speech Communication*, vol. 57, pp. 1–12, 2014.

[48] Kunxia Wang, Guoxin Su, Li Liu, and Shu Wang, "Wavelet packet analysis for speaker-independent emotion recognition," *Neurocomputing*, vol. 398, pp. 257–264, 2020.

[49] Elif Bozkurt, Engin Erzin, Cigdem Eroglu Erdem, and A Tanju Erdem, "Formant position based weighted spectral features for emotion recognition," *Speech Communication*, vol. 53, no. 9-10, pp. 1186–1197, 2011.

[50] Qingli Zhang, Ning An, Kunxia Wang, Fuji Ren, and Lian Li, "Speech emotion recognition using combination of features," in *2013 Fourth International Conference on Intelligent Control and Information Processing (ICICIP)*. IEEE, 2013, pp. 523–528.

[51] Xia Mao, Lijiang Chen, and Liqin Fu, "Multi-level speech emotion recognition based on hmm and ann," in *2009 WRI World congress on computer science and information engineering*. IEEE, 2009, vol. 7, pp. 225–229.

[52] Sungrack Yun and Chang D Yoo, "Speech emotion recognition via a max-margin framework incorporating a loss function based on the watson and tellegen's emotion model," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2009, pp. 4169–4172.

[53] Farah Chenchah and Zied Lachiri, "A bio-inspired emotion recognition system under real-life conditions," *Applied Acoustics*, vol. 115, pp. 6–14, 2017.

[54] Laura Caponetti, Cosimo Alessandro Buscicchio, and Giovanna Castellano, "Biologically inspired emotion recognition from speech," *EURASIP journal on Advances in Signal Processing*, vol. 2011, pp. 1–10, 2011.

[55] Aseef Iqbal and Kakon Barua, "A real-time emotion recognition from speech using gradient boosting," in *2019 international conference on electrical, computer and communication engineering (ECCE)*. IEEE, 2019, pp. 1–5.

[56] T Anrarjon, Kwon Mustaqeem, and S Kwon, "Deep-net: A lightweight cnn-based speech emotion recognition system using deep system using deep," *Sensors*, vol. 20, pp. 5212, 2020.

[57] José Manuel Fuentes, Joaquin Taverner, Jaime Andres Rincon, and Vicente Botti, "Towards a classifier to recognize emotions using voice to improve recommendations," in *Highlights in Practical Applications of Agents, Multi-Agent Systems, and Trust-worthiness. The PAAMS Collection: International Workshops of PAAMS 2020, L'Aquila, Italy, October 7–9, 2020, Proceedings 18*. Springer, 2020, pp. 218–225.

[58] Jianfeng Zhao, Xia Mao, and Lijiang Chen, "Speech emotion recognition using deep 1d & 2d cnn lstm networks," *Biomedical signal processing and control*, vol. 47, pp. 312–323, 2019.

[59] Hao Meng, Tianhao Yan, Fei Yuan, and Hongwei Wei, "Speech emotion recognition from 3d log-mel spectrograms with deep learning network," *IEEE access*, vol. 7, pp. 125868–125881, 2019.

[60] Zhengwei Huang, Ming Dong, Qirong Mao, and Yongzhao Zhan, "Speech emotion recognition using cnn," in *Proceedings of the 22nd ACM international conference on Multimedia*, 2014, pp. 801–804.

[61] Wootaek Lim, Daeyoung Jang, and Taejin Lee, "Speech emotion recognition using convolutional and recurrent neural networks," in *2016 Asia-Pacific signal and information processing association annual summit and conference (APSIPA)*. IEEE, 2016, pp. 1–4.

[62] Jinkyu Lee and Ivan Tashev, "High-level feature representation using recurrent neural network for speech emotion recognition," in *Interspeech 2015*, 2015.

[63] Felicia Andayani, Lau Bee Theng, Mark Teekit Tsun, and Caslon Chua, "Hybrid lstm-transformer model for emotion recognition from speech audio files," *IEEE Access*, vol. 10, pp. 36018–36027, 2022.

[64] Felicia Andayani, Lau Bee Theng, Mark TeeKit Tsun, and Caslon Chua, "Recognition of emotion in speech-related audio files with lstm-transformer," in *2022 5th International Conference on Computing and Informatics (ICCI)*. IEEE, 2022, pp. 087–091.

[65] Li-Wei Chen and Alexander Rudnicky, "Exploring wav2vec 2.0 fine tuning for improved speech emotion recognition," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[66] Gaetan Ramet, Philip N Garner, Michael Baeriswyl, and Alexandros Lazaridis, "Context-aware attention mechanism for speech emotion recognition," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 126–131.

[67] Rashid Jahangir, Ying Wah Teh, Faiqa Hanif, and Ghulam Mujtaba, "Deep learning approaches for speech emotion recognition: State of the art and research challenges," *Multimedia Tools and Applications*, pp. 1–68, 2021.

[68] Taiba Majid Wani, Teddy Surya Gunawan, Syed Asif Ahmad Qadri, Mira Kartiwi, and Eliathamby Ambikairajah, "A comprehensive review of speech emotion recognition systems," *IEEE access*, vol. 9, pp. 47795–47814, 2021.

[69] Elisabeth Scheiner and Julia Fischer, "Emotion expression: The evolutionary heritage in the human voice," in *Interdisciplinary anthropology*, pp. 105–129. Springer, 2011.

[70] Nicolas Audibert, Véronique Aubergé, and Albert Rilliard, "How we are not equally competent for discriminating acted from spontaneous expressive speech," in *Proceedings of speech prosody*. Citeseer, 2008, pp. 693–696.

[71] Khiet Phuong Truong, "How does real affect affect affect recognition in speech?," 2009.

[72] Nicolas Audibert, Véronique Aubergé, and Albert Rilliard, "Prosodic correlates of acted vs. spontaneous discrimination of expressive speech: a pilot study," in *Speech Prosody 2010-Fifth International Conference*, 2010.

[73] Rebecca Jürgens, Kurt Hammerschmidt, and Julia Fischer, "Authentic and play-acted vocal emotion expressions reveal acoustic differences," *Frontiers in psychology*, vol. 2, pp. 180, 2011.

[74] Jian Huang, Jianhua Tao, Bin Liu, Zheng Lian, and Mingyue Niu, "Multimodal transformer fusion for continuous emotion recognition," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 3507–3511.

[75] Shamane Siriwardhana, Tharindu Kaluarachchi, Mark Billinghurst, and Suranga Nanayakkara, "Multimodal emotion recognition with transformer-based self supervised feature fusion," *IEEE Access*, vol. 8, pp. 176274–176285, 2020.

[76] Thaddeus L Bolton, "Rhythm," *The american journal of psychology*, vol. 6, no. 2, pp. 145–238, 1894.

[77] E Glenn Schellenberg, Ania M Krysciak, and R Jane Campbell, "Perceiving emotion in melody: Interactive effects of pitch and rhythm," *Music Perception*, vol. 18, no. 2, pp. 155–171, 2000.

[78] J.C. Borod, *The Neuropsychology of Emotion*, Affective Science. Oxford University Press, USA, 2000.

[79] James A Russell, "Culture and the categorization of emotions.," *Psychological bulletin*, vol. 110, no. 3, pp. 426, 1991.

[80] Henrik Kessler, Alexander Festini, Harald C Traue, Suzanne Filipic, Michael Weber, and Holger Hoffmann, "Simplex–simulation of personal emotion experience," *Affective Computing*, pp. 255–270, 2008.

[81] Holger Hoffmann, Andreas Scheck, Timo Schuster, Steffen Walter, Kerstin Limbrecht, Harald C Traue, and Henrik Kessler, "Mapping discrete emotions into the dimensional space: An empirical approach," in *2012 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, 2012, pp. 3316–3320.

[82] Sven Buechel and Udo Hahn, "A flexible mapping scheme for discrete and dimensional emotion representations: Evidence from textual stimuli," in *CogSci 2017—Proceedings of the 39th Annual Meeting of the Cognitive Science Society*, 2017, pp. 180–185.

[83] Marián Trnka, Sakhia Darjaa, Marian Ritomskỳ, Róbert Sabo, Milan Rusko, Meilin Schaper, and Tim H Stelkens-Kobsch, "Mapping discrete emotions in the dimensional space: an acoustic approach," *Electronics*, vol. 10, no. 23, pp. 2950, 2021.

[84] J.A. Russell, "A circumplex model of affect," *Journal of personality and social psychology*, vol. 39, no. 6, pp. 1161–1178, 1980.

[85] Eddie Harmon-Jones, Cindy Harmon-Jones, and Elizabeth Summerell, "On the importance of both dimensional and discrete models of emotion," *Behavioral sciences*, vol. 7, no. 4, pp. 66, 2017.

[86] Hira Dhamyal, Bhiksha Raj, and Rita Singh, "Positional encoding for capturing modality specific cadence for emotion detection," *Proc. Interspeech 2022*, pp. 166–170, 2022.

[87] Rui Xia and Yang Liu, "A multi-task learning framework for emotion recognition using 2d continuous space," *IEEE Transactions on affective computing*, vol. 8, no. 1, pp. 3–14, 2015.

[88] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al., "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*. PMLR, 2021, pp. 8748–8763.

[89] Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang, "Clap: Learning audio concepts from natural language supervision," *arXiv preprint arXiv:2206.04769*, 2022.

[90] Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel, "Audioclip: Extending clip to image, text and audio," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 976–980.

[91] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 776–780.

[92] Justin Salamon, Christopher Jacoby, and Juan Pablo Bello, "A dataset and taxonomy for urban sound research," in *Proceedings of the 22nd ACM international conference on Multimedia*, 2014, pp. 1041–1044.

[93] Karol J Piczak, "Esc: Dataset for environmental sound classification," in *Proceedings of the 23rd ACM international conference on Multimedia*, 2015, pp. 1015–1018.

[94] Sinuo Deng, Ge Shi, Lifang Wu, Lehao Xing, Wenjin Hu, Heng Zhang, and Ye Xiang, "Simemotion: A simple knowledgeable prompt tuning method for image emotion classification," in *International Conference on Database Systems for Advanced Applications*. Springer, 2022, pp. 222–229.

[95] Rebecca Jürgens, Annika Grass, Matthis Drolet, and Julia Fischer, "Effect of acting experience on emotion expression and recognition in voice: Non-actors provide better stimuli than expected," *Journal of nonverbal behavior*, vol. 39, no. 3, pp. 195–214, 2015.

[96] Janneke Wilting, Emiel Krahmer, and Marc Swerts, "Real vs. acted emotional speech," in *Ninth International Conference on Spoken Language Processing*, 2006.

[97] Klaus R Scherer, "Vocal markers of emotion: Comparing induction and acting elicitation," *Computer Speech & Language*, vol. 27, no. 1, pp. 40–58, 2013.

[98] Oliver Sacks, "The president's speech," *Language, Communication and Education*, p. 23, 1985.

[99] Tom Johnstone and Klaus R Scherer, "The effects of emotions on voice quality," in *Proceedings of the XIVth international congress of phonetic sciences*. Citeseer, 1999, pp. 2029–2032.

[100] Klaus R Scherer and James S Oshinsky, "Cue utilization in emotion attribution from auditory stimuli," *Motivation and emotion*, vol. 1, no. 4, pp. 331–346, 1977.

[101] Carl E Williams and Kenneth N Stevens, "Emotions and speech: Some acoustical correlates," *The Journal of the Acoustical Society of America*, vol. 52, no. 4B, pp. 1238–1250, 1972.

[102] Blake Myers-Schulz, Maia Pujara, Richard C Wolf, and Michael Koenigs, "Inherent emotional quality of human speech sounds," *Cognition & emotion*, vol. 27, no. 6, pp. 1105–1113, 2013.

[103] Klaus R Scherer, Tom Johnstone, and Gundrun Klasmeyer, "Vocal expression of emotion," *Handbook of affective sciences*, pp. 433–456, 2003.

[104] Sona Patel, Klaus R Scherer, Eva Björkner, and Johan Sundberg, "Mapping emotions into acoustic space: The role of voice production," *Biological psychology*, vol. 87, no. 1, pp. 93–98, 2011.

[105] Cynthia Whissell, "Phonosymbolism and the emotional nature of sounds: evidence of the preferential use of particular phonemes in texts of differing emotional tone," *Perceptual and Motor Skills*, vol. 89, no. 1, pp. 19–48, 1999.

[106] Shahan Ali Memon, Hira Dhamyal, Oren Wright, Daniel Justice, Vijaykumar Palat, William Boler, Bhiksha Raj, and Rita Singh, "Detecting gender differences in perception of emotion in crowdsourced data," *arXiv preprint arXiv:1910.11386*, 2019.

[107] "Podcast Directory," https://www.npr.org/podcasts/.

[108] "Top collections at the archive," https://archive.org/.

[109] Amazon Mechanical Turk, "Amazon mechanical turk," *Retrieved August*, vol. 17, pp. 2012, 2012.

[110] Frank Wilcoxon, "Individual comparisons by ranking methods," in *Breakthroughs in statistics*, pp. 196–202. Springer, 1992.

[111] "Google Speech To Test," https://cloud.google.com/speech-to-text/.

[112] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[113] Paul Lamere, Philip Kwok, Evandro Gouvea, Bhiksha Raj, Rita Singh, William Walker, Manfred Warmuth, and Peter Wolf, "The cmu sphinx-4 speech recognition system," in

IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP 2003), Hong Kong, 2003, vol. 1, pp. 2–5.

[114] Navonil Majumder, Devamanyu Hazarika, Alexander Gelbukh, Erik Cambria, and Soujanya Poria, "Multimodal sentiment analysis using hierarchical fusion with context modeling," *Knowledge-based systems*, vol. 161, pp. 124–133, 2018.

[115] Miriam Kienast and Walter F Sendlmeier, "Acoustical analysis of spectral and temporal changes in emotional speech," in *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*, 2000.

[116] Thomas Goldbeck, Frank Tolkmitt, and Klaus R Scherer, "Experimental studies on vocal affect communication.," 1988.

[117] Janneke Iven, "The effect of emotional valence on word choice," 2017.

[118] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[119] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency, "Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos," *arXiv preprint arXiv:1606.06259*, 2016.

[120] Jinhwan Park, Chanwoo Kim, and Wonyong Sung, "Convolution-based attention model with positional encoding for streaming speech recognition on embedded devices," in *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2021, pp. 30–37.

[121] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[122] Amir Zadeh and Paul Pu, "Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers)*, 2018.

[123] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al., "Conformer: Convolution-augmented transformer for speech recognition," *arXiv preprint arXiv:2005.08100*, 2020.

[124] Geng Tu, Jintao Wen, Hao Liu, Sentao Chen, Lin Zheng, and Dazhi Jiang, "Exploration meets exploitation: Multitask learning for emotion recognition based on discrete and dimensional models," *Knowledge-Based Systems*, vol. 235, pp. 107598, 2022.

[125] Ehab A. Albadawy and Yelin Kim, "Joint discrete and continuous emotion prediction using ensemble and end-to-end approaches," in *Proceedings of the 18th ACM International*

Conference on Multimodal Interaction, New York, NY, USA, 2018, ICMI 2018, pp. 366–375, ACM.

[126] R. Lotfian and C. Busso, "Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings," *IEEE Transactions on Affective Computing*, vol. 10, no. 4, pp. 471–483, October-December 2019.

[127] Hira Dhamyal, Shahan Ali Memon, Bhiksha Raj, and Rita Singh, "The phonetic bases of vocal expressed emotion: Natural versus acted," *Proc. Interspeech 2020*, pp. 3451–3455, 2020.

[128] Shinji Watanabe, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Jiro Nishitoba, Yuya Unno, Nelson Enrique Yalta Soplin, Jahn Heymann, Matthew Wiesner, Nanxin Chen, Adithya Renduchintala, and Tsubasa Ochiai, "ESPnet: End-to-end speech processing toolkit," in *Proceedings of Interspeech*, 2018, pp. 2207–2211.

[129] Roshan Sharma, Tyler Vuong, Mark Lindsey, Hira Dhamyal, Rita Singh, and Bhiksha Raj, "Self-supervision and learnable strfs for age, emotion, and country prediction," *arXiv preprint arXiv:2206.12568*, 2022.

[130] Pooyan Safari, Miquel India, and Javier Hernando, "Self-attention encoding and pooling for speaker recognition," *Proc. Interspeech 2020*, pp. 941–945, 2020.

[131] Sundararajan Srinivasan, Zhaocheng Huang, and Katrin Kirchhoff, "Representation learning through cross-modal conditional teacher-student training for speech emotion recognition," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6442–6446.

[132] Md Asif Jalal, Rosanna Milner, and Thomas Hain, "Empirical interpretation of speech emotion perception with attention based model for speech emotion recognition," in *Proceedings of Interspeech 2020*. International Speech Communication Association (ISCA), 2020, pp. 4113–4117.

[133] Yang Li, Constantinos Papayiannis, Viktor Rozgic, Elizabeth Shriberg, and Chao Wang, "Confidence estimation for speech emotion recognition based on the relationship between emotion categories and primitives," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7352–7356.

[134] Edmilson Morais, Ron Hoory, Weizhong Zhu, Itai Gat, Matheus Damasceno, and Hagai Aronowitz, "Speech emotion recognition using self-supervised features," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 6922–6926.

[135] Meysam Shamsi and Marie Tahon, "Training speech emotion classifier without categorical annotations," *arXiv preprint arXiv:2210.07642*, 2022.

[136] Reza Lotfian and Carlos Busso, "Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings," *IEEE Transactions on Affective Computing*, vol. 10, no. 4, pp. 471–483, 2017.

[137] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan, "Iemocap: interactive emotional dyadic motion capture database.," *Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[138] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea, "MELD: A multimodal multi-party dataset for emotion recognition in conversations," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, July 2019, pp. 527–536, Association for Computational Linguistics.

[139] Robert Plutchik, *The emotions*, University Press of America, 1991.

[140] Paul Ekman, *Are there basic emotions?*, American Psychological Association, 1992.

[141] Shahan Ali Memon, Hira Dhamyal, Oren Wright, Daniel Justice, Vijaykumar Palat, William Boler, Bhiksha Raj, and Rita Singh, "Detecting gender differences in perception of emotion in crowdsourced data," *arXiv preprint arXiv:1910.11386*, 2019.

[142] Robert W Frick, "Communicating emotion: The role of prosodic features.," *Psychological bulletin*, vol. 97, no. 3, pp. 412, 1985.

[143] Klaus R Scherer, "Acoustic concomitants of emotional dimensions: Judging affect from synthesized tone sequences.," 1972.

[144] Aneta Pavlenko, *Emotions and multilingualism.*, Cambridge University Press, 2005.

[145] Soham Deshmukh, Benjamin Elizalde, and Huaming Wang, "Audio retrieval with wavtext5k and clap training," *arXiv preprint arXiv:2209.14275*, 2022.

[146] Weixing Wang, Qianqian Li, Jingwen Xie, Ningfeng Hu, Ziao Wang, and Ning Zhang, "Research on emotional semantic retrieval of attention mechanism oriented to audio-visual synesthesia," *Neurocomputing*, vol. 519, pp. 194–204, 2023.

[147] Ha Thi Phuong Thao, Gemma Roig, and Dorien Herremans, "Emomv: Affective music-video correspondence learning datasets for classification and retrieval," *Information Fusion*, vol. 91, pp. 64–79, 2023.

[148] Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D Plumbley, "Panns: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2880–2894, 2020.

[149] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al., "Huggingface's transformers: State-of-the-art natural language processing," *arXiv preprint arXiv:1910.03771*, 2019.

[150] Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen, "Clotho: an audio captioning dataset," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.

[151] Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim, "AudioCaps: Generating Captions for Audios in The Wild," in *NAACL-HLT*, 2019.

[152] Irene Martín-Morató and Annamaria Mesaros, "What is the ground truth? reliability of multi-annotator data for audio tagging," in *2021 29th European Signal Processing Conference (EUSIPCO)*. IEEE, 2021, pp. 76–80.

[153] Eduardo Fonseca, Xavier Favory, Jordi Pons, Frederic Font, and Xavier Serra, "Fsd50k: An open dataset of human-labeled sound events," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2022.

[154] Hartmut Traunmüller and Anders Eriksson, "The frequency range of the voice fundamental in the speech of male and female adults," *Unpublished manuscript*, vol. 11, 1995.

[155] Hettie Roebuck, Kun Guo, and Patrick Bourke, "Attending at a low intensity increases impulsivity in an auditory sustained attention to response task," *Perception*, vol. 44, no. 12, pp. 1371–1382, 2015.

[156] F Martínez-Sánchez, JJG Meilán, J Carro, C Gómez Íñiguez, L Millian-Morell, IM Pujante Valverde, T López-Alburquerque, and DE López, "Speech rate in parkinson's disease: A controlled study," *Neurología (English Edition)*, vol. 31, no. 7, pp. 466–472, 2016.

[157] Brian McFee, Colin Raffel, Dawen Liang, Daniel P Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto, "librosa: Audio and music signal analysis in python," in *Proceedings of the 14th python in science conference*, 2015, vol. 8, pp. 18–25.

[158] "Praat.," https://www.fon.hum.uva.nl/praat, [Online].

[159] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea, "Meld: A multimodal multi-party dataset for emotion recognition in conversations," *arXiv preprint arXiv:1810.02508*, 2018.

[160] Houwei Cao, David G Cooper, Michael K Keutmann, Ruben C Gur, Ani Nenkova, and Ragini Verma, "Crema-d: Crowd-sourced emotional multimodal actors dataset," *IEEE transactions on affective computing*, vol. 5, no. 4, pp. 377–390, 2014.

[161] Steven R Livingstone and Frank A Russo, "The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english," *PloS one*, vol. 13, no. 5, pp. e0196391, 2018.

[162] Bongjun Kim and Bryan Pardo, "Improving content-based audio retrieval by vocal imitation feedback," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 4100–4104.

[163] Cristina Luna-Jiménez, Ricardo Kleinlein, David Griol, Zoraida Callejas, Juan M Montero, and Fernando Fernández-Martínez, "A proposal for multimodal emotion recognition using aural transformers and action units on ravdess dataset," *Applied Sciences*, vol. 12, no. 1, pp. 327, 2021.

[164] Alexandra Saliba, Yuanchao Li, Ramon Sanabria, and Catherine Lai, "Layer-wise analysis of self-supervised acoustic word embeddings: A study on speech emotion recognition," *arXiv preprint arXiv:2402.02617*, 2024.

[165] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al., "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, pp. 9, 2019.

[166] Orchid Chetia Phukan, Arun Balaji Buduru, and Rajesh Sharma, "A comparative study of pre-trained speech and audio embeddings for speech emotion recognition," *arXiv preprint arXiv:2304.11472*, 2023.

[167] Soham Deshmukh, Benjamin Elizalde, Rita Singh, and Huaming Wang, "Pengi: An audio language model for audio tasks," in *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds. 2023, vol. 36, pp. 18090–18108, Curran Associates, Inc.

[168] Rao Ma, Adian Liusie, Mark Gales, and Kate Knill, "Investigating the emergent audio classification ability of asr foundation models," in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 2024, pp. 4746–4760.

[169] Zhifeng Kong, Arushi Goel, Rohan Badlani, Wei Ping, Rafael Valle, and Bryan Catanzaro, "Audio flamingo: A novel audio language model with few-shot learning and dialogue abilities," 2024.

[170] Taesik Gong, Josh Belanich, Krishna Somandepalli, Arsha Nagrani, Brian Eoff, and Brendan Jou, "Lanser: Language-model supported speech emotion recognition," *arXiv preprint arXiv:2309.03978*, 2023.

[171] Hira Dhamyal, Benjamin Elizalde, Soham Deshmukh, Huaming Wang, Bhiksha Raj, and Rita Singh, "Describing emotions with acoustic property prompts for speech emotion recognition," *arXiv preprint arXiv:2211.07737*, 2022.

[172] Raymond B Cattell, "The description of personality: Basic traits resolved into clusters.," *The journal of abnormal and social psychology*, vol. 38, no. 4, pp. 476, 1943.

[173] Raymond B Cattell, "The description of personality: Principles and findings in a factor analysis," *The American journal of psychology*, vol. 58, no. 1, pp. 69–90, 1945.

[174] Ernest C Tupes and Raymond E Christal, "Recurrent personality factors based on trait ratings," *Journal of personality*, vol. 60, no. 2, pp. 225–251, 1992.

[175] Oliver P John, Laura P Naumann, and Christopher J Soto, "Paradigm shift to the integrative big five trait taxonomy: History, measurement, and conceptual issues.," 2008.

[176] Paul T Costa Jr and Robert R McCrae, "Revised neo personality inventory (neo-pi-r) and neo five-factor (neo-ffi) inventory professional manual," *Odessa, Fl: PAR*, 1992.

[177] Hans J Eysenck, *Dimensions of personality*, vol. 5, Transaction Publishers, 1947.

[178] Isabel Briggs-Meyers, A Hammer, Mary McCauley, and Naomi Quenk, *MBTI Manual: A Guide to the Development and Use of the Meyers-Briggs Type Indicator*, CPP Incorporated, 2003.

[179] Richard W Robins, Oliver P John, Avshalom Caspi, Terrie E Moffitt, and Magda Stouthamer-Loeber, "Resilient, overcontrolled, and undercontrolled boys: three replicable personality types.," *Journal of Personality and Social psychology*, vol. 70, no. 1, pp. 157, 1996.

[180] Martin Gerlach, Beatrice Farb, William Revelle, and Luis A Nunes Amaral, "A robust data-driven approach identifies four personality types across four large data sets," *Nature human behaviour*, vol. 2, no. 10, pp. 735–742, 2018.

[181] Andrea Guidi, Claudio Gentili, Enzo Pasquale Scilingo, and Nicola Vanello, "Analysis of speech features and personality traits," *Biomedical signal processing and control*, vol. 51, pp. 1–7, 2019.

[182] Klaus R Scherer, "Personality inference from voice quality: The loud voice of extroversion," *European Journal of Social Psychology*, vol. 8, no. 4, pp. 467–487, 1978.

[183] and others, "The interspeech 2012 speaker trait challenge," in *INTERSPEECH 2012, Portland, OR, USA*, 2012.

[184] Gelareh Mohammadi, Antonio Origlia, Maurizio Filippone, and Alessandro Vinciarelli, "From speech to personality: Mapping voice quality and intonation into personality differences," in *Proceedings of the 20th ACM international conference on Multimedia*, 2012, pp. 789–792.

[185] Tim Polzehl, Sebastian Möller, and Florian Metze, "Automatically assessing personality from speech," in *2010 IEEE fourth international conference on semantic computing*. IEEE, 2010, pp. 134–140.

[186] Joan-Isaac Biel and Daniel Gatica-Perez, "The youtube lens: Crowdsourced personality impressions and audiovisual analysis of vlogs," *IEEE Transactions on Multimedia*, vol. 15, no. 1, pp. 41–55, 2012.

[187] Michelle Hewlett Sanchez, Aaron Lawson, Dimitra Vergyri, and Harry Bratt, "Multi-system fusion of extended context prosodic and cepstral features for paralinguistic speaker trait classification.," in *INTERSPEECH*, 2012, pp. 514–517.

[188] Clément Chastagnol and Laurence Devillers, "Personality traits detection using a parallelized modified sffs algorithm," *computing*, vol. 15, pp. 16, 2012.

[189] Effat Jalaeian Zaferani, Mohammad Teshnehlab, and Mansour Vali, "Automatic personality traits perception using asymmetric auto-encoder," *IEEE access*, vol. 9, pp. 68595–68608, 2021.

[190] Dairazalia Sanchez-Cortes, Oya Aran, Dinesh Babu Jayagopi, Marianne Schmid Mast, and Daniel Gatica-Perez, "Emergent leaders through looking and speaking: from audio-visual data to multimodal recognition," *Journal on Multimodal User Interfaces*, vol. 7, pp. 39–53, 2013.

[191] Hugo Jair Escalante, Isabelle Guyon, Sergio Escalera, Julio Jacques, Meysam Madadi, Xavier Baró, Stephane Ayache, Evelyne Viegas, Yağmur Güçlütürk, Umut Güçlü, et al., "Design of an explainable machine learning challenge for video interviews," in *2017 International joint conference on neural networks (IJCNN)*. IEEE, 2017, pp. 3688–3695.

[192] Oya Celiktutan, Efstratios Skordos, and Hatice Gunes, "Multimodal human-human-robot interactions (mhhri) dataset for studying personality and engagement," *IEEE Transactions on Affective Computing*, vol. 10, no. 4, pp. 484–497, 2017.

[193] Cristina Palmero, Javier Selva, Sorina Smeureanu, Julio Junior, CS Jacques, Albert Clapés, Alexa Moseguí, Zejian Zhang, David Gallardo, Georgina Guilera, et al., "Context-aware personality inference in dyadic scenarios: Introducing the udiva dataset," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 1–12.

[194] Steven J Rubenzer and Thomas R Faschingbauer, *Personality, character, and leadership in the White House: Psychologists assess the presidents*, Potomac Books, Inc., 2004.

[195] "Miller Center of Public Affairs, University of Virginia. "Presidential Speeches: Downloadable Data."," Accessed on March 17, 2022.

[196] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman, "Voxceleb: a large-scale speaker identification dataset," *arXiv preprint arXiv:1706.08612*, 2017.

[197] Beatrice Rammstedt and Oliver P John, "Measuring personality in one minute or less: A 10-item short version of the big five inventory in english and german," *Journal of research in Personality*, vol. 41, no. 1, pp. 203–212, 2007.

[198] Oliver P John, Sanjay Srivastava, et al., "The big-five trait taxonomy: History, measurement, and theoretical perspectives," 1999.

[199] Johan Sundberg, "The acoustics of the singing voice," *Scientific American*, vol. 236, no. 3, pp. 82–91, 1977.

[200] DR Boone, "The voice and voice therapy," *Allyn and Bacon google schola*, vol. 2, pp. 830–843, 2005.

[201] Casey A Klofstad, Rindy C Anderson, and Stephen Nowicki, "Perceptions of competence, strength, and age influence voters to select leaders with lower-pitched voices," *PloS one*, vol. 10, no. 8, pp. e0133779, 2015.

[202] Francis Nolan and Harry Hollien, "The phonetic bases of speaker recognition by francis nolan," 1985.

[203] Philippe H Dejonckere, Patrick Bradley, Pais Clemente, Guy Cornut, Lise Crevier-Buchman, Gerhard Friedrich, Paul Van De Heyning, Marc Remacle, and Virginie Woisard, "A basic protocol for functional assessment of voice pathology, especially for investigating the efficacy of (phonosurgical) treatments and evaluating new assessment techniques: guideline elaborated by the committee on phoniatrics of the european laryngological society (els)," *European Archives of Oto-rhino-laryngology*, vol. 258, pp. 77–82, 2001.

[204] Prakash Boominathan, John Samuel, Ravikumar Arunachalam, Roopa Nagarajan, and Shenbagavalli Mahalingam, "Multi parametric voice assessment: Sri ramachandra university protocol," *Indian Journal of Otolaryngology and Head & Neck Surgery*, vol. 66, pp. 246–251, 2014.

[205] JA Maidment, "The phonetic description of voice quality.," *Journal of the International Phonetic Association*, vol. 11, no. 2, pp. 78–84, 1981.

[206] Minoru Hirano and Karen R McCormick, "Clinical examination of voice by minoru hirano," 1986.

[207] Tadeus Nawka, Lutz Christian Anders, and J Wendler, "Die auditive beurteilung heiserer stimmen nach dem rbh-system," *Sprache Stimme Gehör*, vol. 18, no. 3, pp. 130–133, 1994.

[208] Gail B Kempster, Bruce R Gerratt, Katherine Verdolini Abbott, Julie Barkmeier-Kraemer, and Robert E Hillman, "Consensus auditory-perceptual evaluation of voice: development of a standardized clinical protocol," 2009.

[209] Maria Koutsogiannaki, Olympia Simantiraki, Gilles Degottex, and Yannis Stylianou, "The importance of phase on voice quality assessment," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[210] Youri Maryn, Paul Corthals, Paul Van Cauwenberge, Nelson Roy, and Marc De Bodt, "Toward improved ecological validity in the acoustic measurement of overall voice quality:

combining continuous speech and sustained vowels," *Journal of voice*, vol. 24, no. 5, pp. 540–555, 2010.

[211] Shaheen N Awan, Nelson Roy, and Christopher Dromey, "Estimating dysphonia severity in continuous speech: application of a multi-parameter spectral/cepstral model," *Clinical linguistics & phonetics*, vol. 23, no. 11, pp. 825–841, 2009.

[212] Alessandro Vinciarelli and Gelareh Mohammadi, "A survey of personality computing," *IEEE Transactions on Affective Computing*, vol. 5, no. 3, pp. 273–291, 2014.

[213] Rita Singh, *PROFILING HUMANS FROM THEIR VOICE.*, Springer, 2019.

[214] Rahul Shrivastav and Christine M Sapienza, "Objective measures of breathy voice quality obtained using an auditory model," *The Journal of the Acoustical Society of America*, vol. 114, no. 4, pp. 2217–2224, 2003.

[215] Laurent Eskenazi, Donald G Childers, and Douglas M Hicks, "Acoustic correlates of vocal quality," *Journal of Speech, Language, and Hearing Research*, vol. 33, no. 2, pp. 298–306, 1990.

[216] Michela J Mir, Nicole E Herndon, Aparna Wagle Shukla, Karen Wheeler-Hegland, and Nikolaus R McFarland, "The vocal flutter of multiple system atrophy: A parkinsonian-type phenomenon?," *Movement Disorders Clinical Practice*, vol. 11, no. 4, pp. 403–410, 2024.

[217] Paulo AL Pontes, Vanessa P Vieira, Maria IR Gonçalves, and Antônio AL Pontes, "Characteristics of hoarse, rough and normal voices: acoustic spectrographic comparative analysis," *Rev Bras Otorrinolaringol*, vol. 68, no. 2, pp. 182–188, 2002.

[218] Meike Brockmann, Claudio Storck, Paul N Carding, and Michael J Drinnan, "Voice loudness and gender effects on jitter and shimmer in healthy adults," 2008.

[219] Richard J Klich, "Relationships of vowel characteristics to listener ratings of breathiness," *Journal of Speech, Language, and Hearing Research*, vol. 25, no. 4, pp. 574–580, 1982.

[220] Brian B Monson, Eric J Hunter, Andrew J Lotto, and Brad H Story, "The perceptual significance of high-frequency energy in the human voice," *Frontiers in psychology*, vol. 5, pp. 587, 2014.

[221] Bruce R Gerratt, Kristin Precoda, David G Hanson, and Gerald S Berke, "Source characteristics of diplophonia," *The Journal of the Acoustical Society of America*, vol. 83, no. S1, pp. S66–S66, 1988.

[222] Catherine Madill, Duong Duy Nguyen, Kristie Yick-Ning Cham, Daniel Novakovic, and Patricia McCabe, "The impact of nasalance on cepstral peak prominence and harmonics-to-noise ratio," *The Laryngoscope*, vol. 129, no. 8, pp. E299–E304, 2019.

[223] Diane Schreibweiss-Merin and Lee M. Terrio, "Acoustic analysis of diplophonia: A case study," *Perceptual and Motor Skills*, vol. 63, no. 2, pp. 755–765, 1986.

[224] Donald G Childers and Chih K Lee, "Vocal quality factors: Analysis, synthesis, and perception," *the Journal of the Acoustical Society of America*, vol. 90, no. 5, pp. 2394–2410, 1991.

[225] Richard Wright, Courtney Mansfield, and Laura Panfili, "Voice quality types and uses in north american english," *Anglophonia. French Journal of English Linguistics*, , no. 27, 2019.

[226] Michael Ashby and Joanna Przedlacka, "Measuring incompleteness: Acoustic correlates of glottal articulations," *Journal of the International Phonetic Association*, vol. 44, no. 3, pp. 283–296, 2014.

[227] Dennis H Klatt and Laura C Klatt, "Analysis, synthesis, and perception of voice quality variations among female and male talkers," *the Journal of the Acoustical Society of America*, vol. 87, no. 2, pp. 820–857, 1990.

[228] Gang Chen, Jody Kreiman, Yen-Liang Shue, and Abeer Alwan, "Acoustic correlates of glottal gaps," in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[229] Marshall Strome, Jeannine Stein, Ramon Esclamado, Douglas Hicks, Robert R Lorenz, William Braun, Randall Yetman, Isaac Eliachar, and James Mayes, "Laryngeal transplantation and 40-month follow-up," *New England Journal of Medicine*, vol. 344, no. 22, pp. 1676–1679, 2001.

[230] Monika Tigges, Patrick Mergell, Hanspeter Herzel, Thomas Wittenberg, and Ulrich Eysholdt, "Observation and modelling of glottal biphonation," *Acta Acustica united with Acustica*, vol. 83, no. 4, pp. 707–714, 1997.

[231] Ali Arabi, Maryam Tarameshlu, Roozbeh Behroozmand, and Leila Ghelichi, "Correlation between auditory-perceptual parameters and acoustic characteristics of voice in theater actors," *Middle East Journal of Rehabilitation and Health Studies*, vol. 10, no. 1, 2023.

[232] David Ross Dickson, "An acoustic study of nasality," *Journal of Speech and Hearing Research*, vol. 5, no. 2, pp. 103–111, 1962.

[233] Sara Pearsell and Daniel Pape, "The effects of different voice qualities on the perceived personality of a speaker," *Frontiers in Communication*, vol. 7, pp. 909427, 2023.

[234] Dipl-Ing Philipp Aichinger, *Diplophonic Voice*, Ph.D. thesis, PhD thesis, Medical University of Vienna, 2014.

[235] Gelin Li, Qian Hou, Chi Zhang, Zhen Jiang, and Shusheng Gong, "Acoustic parameters for the evaluation of voice quality in patients with voice disorders," *Annals of Palliative Medicine*, vol. 10, no. 1, pp. 13036–13136, 2021.

[236] Abdellah Kacha, Christophe Mertens, Francis Grenez, Sabine Skodda, and Jean Schoentgen, "On the harmonic-to-noise ratio as an acoustic cue of vocal timbre of parkinson speakers," *Biomedical Signal Processing and Control*, vol. 37, pp. 32–38, 2017.

[237] Eiji Yumoto, Wilbur J Gould, and Thomas Baer, "Harmonics-to-noise ratio as an index of the degree of hoarseness," *The journal of the Acoustical Society of America*, vol. 71, no. 6, pp. 1544–1550, 1982.

[238] Fritz Klingholz and Frank Martin, "Quantitative spectral evaluation of shimmer and jitter," *Journal of Speech, Language, and Hearing Research*, vol. 28, no. 2, pp. 169–174, 1985.

[239] B Radish Kumar, Jayashree S Bhat, and Payal Mukhi, "Vowel harmonic amplitude differences in persons with vocal nodules," *Journal of Voice*, vol. 25, no. 5, pp. 559–561, 2011.

[240] SV Narasimhan and K Vishal, "Spectral measures of hoarseness in persons with hyperfunctional voice disorder," *Journal of Voice*, vol. 31, no. 1, pp. 57–61, 2017.

[241] Christina M Esposito, Morgan Sleeper, and Kevin Schäfer, "Examining the relationship between vowel quality and voice quality," *Journal of the International Phonetic Association*, vol. 51, no. 3, pp. 361–392, 2021.

[242] Arthur S House and Kenneth N Stevens, "Analog studies of the nasalization of vowels," *Journal of Speech and Hearing Disorders*, vol. 21, no. 2, pp. 218–232, 1956.

[243] Marilyn Y Chen, *Acoustic correlates of nasality in speech*, Ph.D. thesis, Massachusetts Institute of Technology, 1996.

[244] Christer Gobl, "A preliminary study of acoustic voice quality correlates," *STL-QPSR*, vol. 4, no. 9-21, pp. 534, 1989.

[245] CarlosToshinori Ishi, Hiroshi Ishiguro, and Norihiro Hagita, "Analysis of the roles and the dynamics of breathy and whispery voice qualities in dialogue speech," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2010, pp. 1–12, 2010.

[246] Eric Bognar and Hiroya Fujisaki, "Analysis, synthesis and perception of the french nasal vowels," in *ICASSP'86. IEEE International Conference on Acoustics, Speech, and Signal Processing*. IEEE, 1986, vol. 11, pp. 1601–1604.

[247] Supraja Anand, Lisa M Kopf, Rahul Shrivastav, and David A Eddins, "Objective indices of perceived vocal strain," *Journal of Voice*, vol. 33, no. 6, pp. 838–845, 2019.

[248] Christopher Dromey, Paul Warrick, and Jonathan Irish, "The influence of pitch and loudness changes on the acoustics of vocal tremor," 2002.

[249] Youri Maryn, Marc S De Bodt, and Paul Van Cauwenberge, "Ventricular dysphonia: clinical aspects and therapeutic options," *The Laryngoscope*, vol. 113, no. 5, pp. 859–866, 2003.

[250] Anne-Maria Laukkanen and Leena Rantala, "Does the acoustic voice quality index (avqi) correlate with perceived creak and strain in normophonic young adult finnish females?," *Folia Phoniatrica et Logopaedica*, vol. 74, no. 1, pp. 62–69, 2022.

[251] James W Lance, Robert S Schwab, and Elizabeth A Peterson, "Action tremor and the cogwheel phenomenon in parkinson's disease," *Brain*, vol. 86, no. 1, pp. 95–110, 1963.

[252] Marcelo Saldías O'Hrens, Christian Castro, Víctor M Espinoza, Justin Stoney, Camilo Quezada, and Anne-Maria Laukkanen, "Spectral features related to the auditory perception of twang-like voices," *Logopedics Phoniatrics Vocology*, pp. 1–18, 2024.

[253] Brad H Story, Ingo R Titze, and Eric A Hoffman, "The relationship of vocal tract shape to three voice qualities," *The Journal of the Acoustical Society of America*, vol. 109, no. 4, pp. 1651–1667, 2001.

[254] Ingo R Titze, "Acoustic interpretation of resonant voice," *Journal of voice*, vol. 15, no. 4, pp. 519–528, 2001.

[255] Brenda Kaye Scoggins Fauls, *A choral conductor's reference guide to acoustic choral music measurement: 1885 to 2007*, The Florida State University, 2008.

[256] Lucie Bailly, Nathalie Henrich Bernardoni, Frank Müller, Anna-Katharina Rohlfs, and Markus Hess, "Ventricular-fold dynamics in human phonation," *Journal of Speech, Language, and Hearing Research*, vol. 57, no. 4, pp. 1219–1242, 2014.

[257] Elizabeth Godoy, Catherine Mayo, and Yannis Stylianou, "Linking loudness increases in normal and lombard speech to decreasing vowel formant separation.," in *INTERSPEECH*, 2013, pp. 133–137.

[258] Sara Hawi, Jana Alhozami, Raneem AlQahtani, Dannah AlSafran, Maram Alqarni, and Lola El Sahmarany, "Automatic parkinson's disease detection based on the combination of long-term acoustic features and mel frequency cepstral coefficients (mfcc)," *Biomedical Signal Processing and Control*, vol. 78, pp. 104013, 2022.

[259] Rongjie Shi, Oliver Niebuhr, Wentao Gu, and Nafiseh Taghva, "The effects of loudness and smiling on timbre features: Implications for charismatic voices in mandarin, german and danish," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 11926–11930.

[260] Javier Gamboa, Félix Javier Jiménez-Jiménez, Alberto Nieto, Ignacio Cobeta, Alberto Vegas, Miguel Ortí-Pareja, Teresa Gasalla, José Antonio Molina, and Esteban García-Albea, "Acoustic voice analysis in patients with essential tremor," *Journal of Voice*, vol. 12, no. 4, pp. 444–452, 1998.

[261] Iris Meerschman, *Effect of voice training and voice therapy: content and dosage*, Ph.D. thesis, Ghent University, 2018.

[262] Yeonggwang Park, Manuel Díaz Cádiz, Kathleen F Nagle, and Cara E Stepp, "Perceptual and acoustic assessment of strain using synthetically modified voice samples," *Journal of Speech, Language, and Hearing Research*, vol. 63, no. 12, pp. 3897–3908, 2020.

[263] Lucie Bailly, Nathalie Henrich, and Xavier Pelorson, "Vocal fold and ventricular fold vibration in period-doubling phonation: Physiological description and aerodynamic modeling," *The Journal of the Acoustical Society of America*, vol. 127, no. 5, pp. 3212–3222, 2010.

[264] O Yasojima, Y Takahashi, and M Tohyama, "Resonant bandwidth estimation of vowels using clustered-line spectrum modeling for pressure speech waveforms," in *2006 IEEE International Symposium on Signal Processing and Information Technology*. IEEE, 2006, pp. 589–593.

[265] James J Ackmann, Anthony Sances, Sanford J Larson, and John B Baker, "Quantitative evaluation of long-term parkinson tremor," *IEEE Transactions on Biomedical Engineering*, , no. 1, pp. 49–56, 1977.

[266] John Charles Herbert, *Broadcast speech and the effect of voice quality on the listener: a study of the various components which categorise listener perception by vocal characteristics.*, Ph.D. thesis, University of Sheffield, 1989.

[267] Nelson Roy, "Personality and voice disorders," *Perspectives on Voice and Voice Disorders*, vol. 21, no. 1, pp. 17–23, 2011.

[268] Päivi Lukkarila, Anne-Maria Laukkanen, and Pertti Palo, "Influence of the intentional voice quality on the impression of female speaker," *Logopedics Phoniatrics Vocology*, vol. 37, no. 4, pp. 158–166, 2012.

[269] Jieun Song, Minjeong Kim, and Jaehan Park, "Acoustic correlates of perceived personality from korean utterances in a formal communicative setting," *Plos one*, vol. 18, no. 10, pp. e0293222, 2023.

[270] PerMagnus Lindborg, "Correlations between acoustic features, personality traits and perception of soundscapes," 2012.

[271] François Mairesse, Marilyn A Walker, Matthias R Mehl, and Roger K Moore, "Using linguistic cues for the automatic recognition of personality in conversation and text," *Journal of artificial intelligence research*, vol. 30, pp. 457–500, 2007.

[272] Björn Schuller, Stefan Steidl, Anton Batliner, Elmar Nöth, Alessandro Vinciarelli, Felix Burkhardt, Rob Van Son, Felix Weninger, Florian Eyben, Tobias Bocklet, et al., "A survey on perceived speaker traits: Personality, likability, pathology, and the first challenge," *Computer speech & language*, vol. 29, no. 1, pp. 100–131, 2015.

[273] Tom Hatherley Pear, "Voice and personality," 1931.

[274] Maria Dietrich and Katherine Verdolini Abbott, "Vocal function in introverts and extraverts during a psychological stress reactivity protocol," 2012.

[275] Ilse Bernadette Labuschagne and Valter Ciocca, "The perception of breathiness: Acoustic correlates and the influence of methodological factors," *Acoustical Science and Technology*, vol. 37, no. 5, pp. 191–201, 2016.

[276] Jefffery Pittam, "Listeners' evaluations of voice quality in australian english speakers," *Language and Speech*, vol. 30, no. 2, pp. 99–113, 1987.

[277] Benjamin Weiss and Benjamin Weiss, "Talker quality in passive scenarios," *Talker Quality in Human and Machine Interaction: Modeling the Listener's Perspective in Passive and Interactive Scenarios*, pp. 23–65, 2020.

[278] Teija Waaramaa, Päivi Lukkarila, Kati Järvinen, Ahmed Geneid, and Anne-Maria Laukkanen, "Impressions of personality from intentional voice quality in arabic-speaking and native finnish-speaking listeners," *Journal of Voice*, vol. 35, no. 2, pp. 326–e21, 2021.

[279] Sinae Lee, Jangwoon Park, and Dugan Um, "Speech characteristics as indicators of personality traits," *Applied Sciences*, vol. 11, no. 18, pp. 8776, 2021.

[280] Klaus R Scherer, "Voice quality analysis of american and german speakers," *Journal of Psycholinguistic Research*, vol. 3, no. 3, pp. 281–298, 1974.

[281] Saeed Saeedi, Payman Dabirmoghaddam, Mehdi Soleimani, and Mahshid Aghajanzadeh, "Relationship among five-factor personality traits and psychological distress with acoustic analysis," *Laryngoscope Investigative Otolaryngology*, vol. 8, no. 4, pp. 996–1006, 2023.

[282] Ofer Amir, Gaya Noam, Adi Primov-Fever, Ruth Epstein, Marion Alston, and Idit Gutman, "Voice disorders and personality: New steps on an old path," *Journal of Voice*, 2023.

[283] Charles F Diehl, Richard White, and Kenneth W Burk, "Voice quality and anxiety," *Journal of Speech and Hearing Research*, vol. 2, no. 3, pp. 282–285, 1959.

[284] GS Claridge, *Studies in Experimental Psychopathology: Objective Behavioural and Psychophysiological Correlates of Personality*, University of Glasgow (United Kingdom), 1971.

[285] Tim Polzehl, Sebastian Möller, and Florian Metze, "Modeling speaker personality using voice.," in *Interspeech*, 2011, pp. 2369–2372.

[286] Ikuko Patricia Yuasa, "Creaky voice: A new feminine voice quality for young urban-oriented upwardly mobile american women?," *American Speech*, vol. 85, no. 3, pp. 315–337, 2010.

[287] Brett Welch, *Investigating the Psychological Aspects of Voice and Communication via Contemporary Personality Science*, Ph.D. thesis, University of Pittsburgh, 2024.

[288] Linda Diane Kobitisch, "Experimental study of the relationship between perceived nasality and judgments of personality," 1971.

[289] Brad Story, "Physical modeling of voice and voice quality," in *proc. Voqual'03, Voice Quality: Functions, analysis and synthesis, ISCA workshop*, 2003.

[290] Víctor J Rubio, David Aguado, Doroteo T Toledano, and María Pilar Fernández-Gallego, "Feasibility of big data analytics to assess personality based on voice analysis," *Sensors*, vol. 24, no. 22, pp. 7151, 2024.

[291] Lea Tylečková, Zuzana Prokopová, and Radek Skarnitzl, "The effect of voice quality on hiring decisions," *AUC PHILOLOGICA*, vol. 2017, no. 3, pp. 109–120, 2017.

[292] Agaath MC Sluijter and Vincent J Van Heuven, "Spectral balance as an acoustic correlate of linguistic stress," *The Journal of the Acoustical society of America*, vol. 100, no. 4, pp. 2471–2485, 1996.

[293] Sofia Holmqvist, Pekka Santtila, Elisabeth Lindström, Eeva Sala, and Susanna Simberg, "The association between possible stress markers and vocal symptoms," *Journal of voice*, vol. 27, no. 6, pp. 787–e1, 2013.

[294] Ingo R Titze, Christine C Bergan, Eric J Hunter, and Brad Story, "Source and filter adjustments affecting the perception of the vocal qualities twang and yawn," *Logopedics Phoniatrics Vocology*, vol. 28, no. 4, pp. 147–155, 2003.

[295] Maëva Garnier, Nathalie Henrich Bernardoni, Michèle Castellengo, David Sotiropoulos, and Danièle Dubois, "Characterisation of voice quality in western lyrical singing: From teachers' judgements to acoustic descriptions," *Journal of interdisciplinary music studies*, vol. 1, no. 2, pp. 62–91, 2007.

[296] Julia Stern, Christoph Schild, Benedict C Jones, Lisa M DeBruine, Amanda Hahn, David A Puts, Ingo Zettler, Tobias L Kordsmeyer, David Feinberg, Dan Zamfir, et al., "Do voices carry valid information about a speaker's personality?," *Journal of Research in Personality*, vol. 92, pp. 104092, 2021.

[297] Donna Erickson, Shigeto Kawahara, Albert Rilliard, Ryoko Hayashi, Toshiyuki Sadanobu, Yongwei Li, Hayato Daikuhara, João De Moraes, and Kerrie Obert, "Cross cultural differences in arousal and valence perceptions of voice quality," *Speech Prosody 2020*, pp. 720–724, 2020.

[298] Sara Pearsell, *Effects of Acoustic Speech Variation on Personality Trait Perception*, Ph.D. thesis, 2024.

[299] Delia Lorenz, Daniel Schwieger, Hans Moises, and Günther Deuschl, "Quality of life and personality in essential tremor patients," *Movement disorders: official journal of the Movement Disorder Society*, vol. 21, no. 8, pp. 1114–1118, 2006.

[300] Simon M Breil, Sarah Osterholz, Steffen Nestler, and Mitja D Back, "13 contributions of nonverbal cues to the accurate judgment of personality traits," *The Oxford handbook of accurate personality judgment*, pp. 195–218, 2021.

[301] James Hillenbrand, Ronald A Cleveland, and Robert L Erickson, "Acoustic correlates of breathy vocal quality," *Journal of Speech, Language, and Hearing Research*, vol. 37, no. 4, pp. 769–778, 1994.

[302] James Hillenbrand and Robert A Houde, "Acoustic correlates of breathy vocal quality: Dysphonic voices and continuous speech," *Journal of Speech, Language, and Hearing Research*, vol. 39, no. 2, pp. 311–321, 1996.

[303] Adam Stráník, Roman Čmejla, and Jan Vokřál, "Acoustic parameters for classification of breathiness in continuous speech according to the grbas scale," *Journal of Voice*, vol. 28, no. 5, pp. 653–e9, 2014.

[304] S Prytz and B Frökjaer-Jensen, "Longtime average spectra analyses of normal and pathological voices," *Folia Phoniatr*, vol. 28, pp. 280, 1976.

[305] H Yoshida, Y Furuya, K Shimodaira, T Kanazawa, R Kataoka, and K Takahashi, "Spectral characteristics of hypernasality in maxillectomy patients 1," *Journal of Oral Rehabilitation*, vol. 27, no. 8, pp. 723–730, 2000.

[306] Youri Maryn, Marc Leblans, Andrzej Zarowski, and Julie Barkmeier-Kraemer, "Objective acoustic quantification of perceived voice tremor severity," *Journal of Speech, Language, and Hearing Research*, vol. 62, no. 10, pp. 3689–3705, 2019.

[307] Jun Shao, Julia K MacCallum, Yu Zhang, Alicia Sprecher, and Jack J Jiang, "Acoustic analysis of the tremulous voice: assessing the utility of the correlation dimension and perturbation parameters," *Journal of communication disorders*, vol. 43, no. 1, pp. 35–44, 2010.

[308] Julián Villegas, Seunghun J Lee, Jeremy Perkins, and Konstantin Markov, "Psychoacoustic features explain creakiness classifications made by naive and non-naive listeners," *Speech Communication*, vol. 147, pp. 74–81, 2023.

[309] Nicole Hildebrand-Edgar, *Creaky voice: An interactional resource for indexing authority*, Ph.D. thesis, 2016.

[310] Ned W Bowler, "A fundamental frequency analysis of harsh vocal quality," *Communications Monographs*, vol. 31, no. 2, pp. 128–134, 1964.

[311] Elizabeth U Grillo and Katherine Verdolini, "Evidence for distinguishing pressed, normal, resonant, and breathy voice qualities by laryngeal resistance and vocal efficiency in vocally trained subjects," *Journal of Voice*, vol. 22, no. 5, pp. 546–552, 2008.

[312] Irena Yanushevskaya, Christer Gobl, and Ailbhe Ní Chasaide, "Voice quality in affect cueing: does loudness matter?," *Frontiers in psychology*, vol. 4, pp. 52044, 2013.

[313] Yunyang Zeng, Joseph Konan, Shuo Han, David Bick, Muqiao Yang, Anurag Kumar, Shinji Watanabe, and Bhiksha Raj, "Taploss: A temporal acoustic parameter loss for speech enhancement," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[314] Arsha Nagrani, Joon Son Chung, Weidi Xie, and Andrew Zisserman, "Voxceleb: Large-scale speaker verification in the wild," *Computer Speech & Language*, vol. 60, pp. 101027, 2020.

[315] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[316] Daniel Nettle and Lars Penke, "Personality: bridging the literatures from human psychology and behavioural ecology," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 365, no. 1560, pp. 4043–4050, 2010.

[317] Yngwie Asbjørn Nielsen, Stefan Pfattheicher, and Isabel Thielmann, "How much can personality predict prosocial behavior?," *European Journal of Personality*, p. 08902070241251516, 2023.

[318] CI Nass, "Wired for speech: How voice activates and advances the human-computer relationship," 2005.

[319] Lingyan Zhang and Yun Wang, "Digital personality in the voice interaction products design," in *Design Studies and Intelligence Engineering*, pp. 191–204. IOS Press, 2024.

[320] Mohammad Amin Kuhail, Mohamed Bahja, Ons Al-Shamaileh, Justin Thomas, Amina Alkazemi, and Joao Negreiros, "Assessing the impact of chatbot-human personality congruence on user behavior: A chatbot-based advising system case," *IEEE Access*, 2024.

[321] Richard W Robins, Oliver P John, and Avshalom Caspi, "The typological approach to studying personality," *Methods and models for studying the individual*, pp. 135–160, 1998.

[322] Oliver P John, Richard W Robins, and Lawrence A Pervin, *Handbook of personality: Theory and research*, 2010.

[323] Sanne MA Lamers, Gerben J Westerhof, Viktória Kovács, and Ernst T Bohlmeijer, "Differential relationships in the association of the big five personality traits with positive mental health and psychopathology," *Journal of Research in Personality*, vol. 46, no. 5, pp. 517–524, 2012.

[324] Gordon Claridge and Caroline Davis, *Personality and psychological disorders*, Routledge, 2013.

[325] Brent W Roberts, Nathan R Kuncel, Rebecca Shiner, Avshalom Caspi, and Lewis R Goldberg, "The power of personality: The comparative validity of personality traits, socioeconomic status, and cognitive ability for predicting important life outcomes," *Perspectives on Psychological science*, vol. 2, no. 4, pp. 313–345, 2007.

[326] Oliver P John, "Towards a taxonomy of personality descriptors," in *Personality psychology: Recent trends and emerging directions*, pp. 261–271. Springer, 1989.

[327] Raymond B Cattell, *Personality and mood by questionnaire.*, Jossey-Bass, 1973.

[328] Hans Eysenck and Glenn Wilson, "Know your own personality.," 1976.

[329] Jerry S Wiggins, "Personality structure," *Annual review of psychology*, vol. 19, no. 1, pp. 293–350, 1968.

[330] Bryan Stroube, "Literary freedom: Project gutenberg," *XRDS: Crossroads, The ACM Magazine for Students*, vol. 10, no. 1, pp. 3–3, 2003.

[331] Gordon W Allport and Henry S Odbert, "Trait-names: A psycho-lexical study.," *Psychological monographs*, vol. 47, no. 1, pp. i, 1936.

[332] LE ENTICKNAP, "Handbook for the 16 personality factor questionnaire-cattell, rb, saunders, dr, stice, g," 1958.

[333] Warren T Norman, "2800 personality trait descriptors–normative operating characteristics for a university population.," 1967.

[334] Warren T Norman, "Toward an adequate taxonomy of personality attributes: Replicated factor structure in peer nomination personality ratings.," *The journal of abnormal and social psychology*, vol. 66, no. 6, pp. 574, 1963.

[335] LR Goldberg, "From ace to zombie: Some explorations in the language of personality," *Advances in personality assessment/Lawrence Erlbaum Associates*, 1982.

[336] Edgar F Borgatta, "The structure of personality characteristics," *Behavioral science*, vol. 9, no. 1, pp. 8–17, 1964.

[337] John M Digman and Naomi K Takemoto-Chock, "Factors in the natural language of personality: Re-analysis, comparison, and interpretation of six major studies," *Multivariate behavioral research*, vol. 16, no. 2, pp. 149–170, 1981.

[338] Paul T Costa Jr and Robert R McCrae, "Domains and facets: Hierarchical personality assessment using the revised neo personality inventory," *Journal of personality assessment*, vol. 64, no. 1, pp. 21–50, 1995.

[339] Martin Gerlach and Francesc Font-Clos, "A standardized project gutenberg corpus for statistical analysis of natural language and quantitative linguistics," *Entropy*, vol. 22, no. 1, pp. 126, 2020.

[340] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al., "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[341] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al., "Opt: Open pre-trained transformer language models," *arXiv preprint arXiv:2205.01068*, 2022.

[342] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *The Journal of Machine Learning Research*, vol. 21, no. 1, pp. 5485–5551, 2020.