Utrecht University

# Comparison of Acoustic Feature Representation Methods for Apparent Personality Recognition


by

Yizhe Zhang


Submitted to the Game and Media Technology Program

in partial fulfillment of the requirements for the degree of

Master of Science


Graduate Program in Game and Media Technology

Utrecht University


2023

# Comparison of Acoustic Feature Representation Methods for Apparent Personality Recognition

APPROVED BY:

Dr. Heysem Kaya .....
(Thesis Supervisor)

Dr. Itir Önal Ertugrul .....
(Thesis Examiner)

DATE OF APPROVAL:

# ACKNOWLEDGEMENTS

I am thankful to my advisor, Dr. Heysem Kaya, for his guidance, support, kindness, flexibility, empathy and mentorship throughout the thesis writing process. His guidance and support were essential to my thesis.

Dr. Kaya was always willing to answer my questions and help me to solve any difficulties I encountered. He was also kind enough to reply to my emails in a timely manner. He was always willing to meet with me to discuss my thesis, even when I had to reschedule our meetings at the last minute.

I will never forget Dr. Kaya's guidance and mentorship. I am grateful for his support and I will always cherish what he has taught me.

# ABSTRACT

## Comparison of Acoustic Feature Representation Methods for Apparent Personality Recognition

This thesis examines the performance of Fisher Vector representations in classifying personality traits from audio. The Chalearn LAP First Impression dataset is used, which is a multimodal dataset. The audio modality of the dataset is focused on, and different audio feature extraction methods, including wav2vec 2.0, openSMILE, and public dimensional emotion model (PDEM), are studied for their performance on the classification task. Different encoding approaches, such as Fisher Vector, are also studied to see how they affect the performance of the classifier. The results of this thesis suggest that Fisher Vector representations are not the best choice for classifying personality traits from audio for the certain dataset. However, other feature extraction methods, such as openSMILE LLDs and PDEM, can achieve good performance on this task. The thesis also provides some insights into the selection of parameters for feature engineering and the interpretability of Fisher Vector representations.

# TABLE OF CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# LIST OF ACRONYMS/ABBREVIATIONS

| | |
|---|---|
| AI | Artificial Intelligence |
| CCC | Concordance Correlation Coefficient |
| EPQ | Eysenck Personality Questionnaire |
| FI | First Impression |
| FV | Fisher Vector |
| LLDs | Low-Level Descriptors |
| MBTI | Myers-Briggs Type Indicator |
| MDD | Major Depression Disorder |
| MFCC | Mel-Frequency Cepstral Coefficients |
| SVM | Support Vector Machine |
| 16PF | The Sixteen Personality Factor Questionnaire |

# 1. Introduction

## 1.1. Motivation and Problem Statement

It is crucial to have a thorough understanding of a candidate before they join a company. In addition to professional skills, the candidate's personality is also essential for their future career development, work environment, and the overall interests of the company. Therefore, more and more companies are introducing personality tests to determine whether candidates are suitable for the current job. However, it is very difficult to reach a convincing level through a test with a few questions. Common personality models include the MBTI (Myers-Briggs Type Indicator) [2], Big Five Personality Model [3], 16PF (The Sixteen Personality Factor Questionnaire) [4] and EPQ (Eysenck Personality Questionnaire) [5]. Therefore, employers increasingly hope for a more objective, standardized, and data-driven personality assessment of the metrics of the personality test. Machine learning offers a good opportunity here.

Moreover, recognizing personality traits or impressions can aid interpretable automatic recognition of mood disorders, such as depression and bipolar disorder. Literature works in social and medial sciences have shown strong correlations between mood and personality traits and mood disorders. Multiple studies, including [6–10], have reviewed the association of personality traits with mood disorders, such as major depressive disorder (MDD). This meta-analysis indicated a strong connection between some mental illnesses and personality, of which all disorders had a configuration of low Conscientiousness and high Neuroticism. Out of all the disorders, MDD seemed to have the strongest correlation with the Neuroticism factor. An analysis by [11] concluded a common five-factor configuration of high Neuroticism, low Conscientiousness, low Agreeableness, and low Extraversion for almost all disorders. Mood disorders, such as depression and bipolar disorder, were associated with a significantly larger negative effect size of Extraversion in comparison with the rest of the mental disorders.

Recognizing human personality traits with machine learning requires knowledge of image recognition, acoustics, natural language processing, affective computing, computational linguistics, computational paralinguistics, and other related fields [12]. In multiple state-of-the-art studies, multimodal models were found to outperform single modality models for personality five-dimensional recognition [13, 14]. The best performing system was based on audio and video modalities, followed by video only and audio only.

The European Union is the region with the strictest regulations regarding personal data in the world [15]. Personality recognition activities in Europe are subject to controls on the collection of personal data. Therefore, within the EU, the availability of data is particularly limited. It is necessary to reconsider improved methods to conduct experiments in a single modal situation.

### 1.2. Research Objectives

In this thesis, I will examine the performance of Fisher Vector representations in classifying personality traits from audio. I will be using the Chalearn Lapfi First Impression dataset, which is a multimodal dataset. I will focus on the audio modality only of the dataset and study how different audio feature extraction methods, including wav2vec 2.0, openSMILE, and public dimensional emotion model (PDEM) [16], perform on the classification task. I will also study different encoding approaches, such as Fisher Vector, to see how they affect the performance of the classifier

I will tackle the following research questions:

- Question 1: Concerning the apparent personality trait recognition task via Big Five personality traits, is there a significant difference in terms of test set accuracy performance between baseline (Interspeech 2013 Computational Paralinguistics ChallengE Setting) acoustic features and other feature representations?
  - Sub-question 1.1: Is there a significant difference in terms of test set accuracy

performance between baseline acoustic features and FV representation of the same set of openSMILE low-level descriptors (LLDs)?

  - Sub-question 1.2: Is there a significant difference in terms of test set accuracy performance between FV representations of the original (uncompressed) and the compressed (averaged per 5 consecutive, non-overlapping frames) Wav2Vec2 LLDs?
  - Sub-question 1.3: Is there a significant difference in terms of test set accuracy performance between functional representation via mean and standard deviation summarization of the Wav2Vec2 LLDs and their FV representation?
  - Sub-question 1.4: Is there a significant difference in terms of test set accuracy performance between baseline acoustic features and the FV representation of Wav2Vec2 LLDs?
  - Sub-question 1.5: Is there a significant difference in terms of test set accuracy performance between baseline acoustic features and acoustic embeddings from the public dimensional emotion model (PDEM)?

- Question 2: Concerning the apparent personality trait recognition task via Big Five personality traits, is there a significant difference in terms of test set accuracy performance between baseline acoustic features and indirect modeling via intelligible emotion primitive features (arousal, valence, and dominance) extracted from public dimensional emotion model (PDEM)?

We will use McNemar's test [17] for statistical significance analysis to answer the research (sub)-questions.

# 2. Background and Literature Review

## 2.1. Background on Acoustic Features

### 2.1.1. Paralinguistics

Paralinguistics is a study of the semantic impact of vocal features or changes [18]. The scope of paralinguistics includes the nuances of meaning given by variations in voice, intonation, stress, and rhythm. Paralanguage can be expressed intentionally or unintentionally.

Some of the most common acoustic features used in paralinguistics [19] include: pitch, loudness, tempo, and rhythm. In addition to these four basic acoustic features, there are a number of other features that can be used in paralinguistics, such as: jitter, shimme and Voice quality.

### 2.1.2. openSMILE 3.0

openSMILE 3.0 [20] is a powerful and versatile open-source toolkit for audio feature extraction and classification. openSMILE needs to configure the config file first and use the config to extract audio features. It is mainly used in speech recognition, affective computing, and music information acquisition.

### 2.1.3. Low-level descriptors (LLDs)

For audio signals, Low-level descriptors (LLDs) [21] are features that are extracted directly from the raw data of audio signals. They are regarded as the basis for higher-level features. Some LLDs include:

- Spectral features: These describe the frequency content of the signal. For ex-

ample, the **spectral centroid** is the average frequency of the signal, and the **spectral spread** is the range of frequencies present.

- Temporal features: These describe how the signal changes over time. For example, the **zero crossing rate** is the number of times the signal crosses the zero axis per second, and the **energy** is the average power of the signal.
- Timbral features: These describe the quality of the sound. For example, the **harmonicity** is the ratio of the harmonic frequencies to the fundamental frequency, and the **roughness** is a measure of the smoothness of the sound.

### 2.1.4. The INTERSPEECH 2013 Computational Paralinguistics Challenge

THE INTERSPEECH is a worldwide paralinguistics challenge that aims to promote research in the field of computational paralinguistics. The INTERSPEECH 2013 Computational Paralinguistics Challenge(ComparE) consists of four tasks [22]:

- Social Signal Recognition: **Recognize social signals** in speech, such as **head nods** and **body language**.
- Conflict Recognition: **Recognize conflicts** in speech, such as arguments and debates.
- Emotion Recognition: **Recognizes emotions** in speech, such as happiness, sadness, and anger.
- Autism Spectrum Disorder (ASD) Recognition: **Recognizes ASD features** in speech, such as monotone intonation and speech repetition.

The feature set used for the InterSpeech 2013 ComparE Challenge has a total of **6373 global features** and also contains **130-dimensional LLDs with temporal feature dimensions**.

### 2.1.5. Wav2vec 2.0

Wav2Vec2.0 is a pre-trained model proposed by Facebook AI Research (FAIR) in 2020. It is a general-purpose feature extractor that can be used for a variety of speech tasks, including speech recognition, speech synthesis, and speaker recognition [23].

The encoder of Wav2Vec2.0 is a transformer network. The base version of Wav2Vec2.0 has 12 transformer layers, while the large version has 24 transformer layers. The positional encoding used in Wav2Vec2.0 is a convolutional layer.

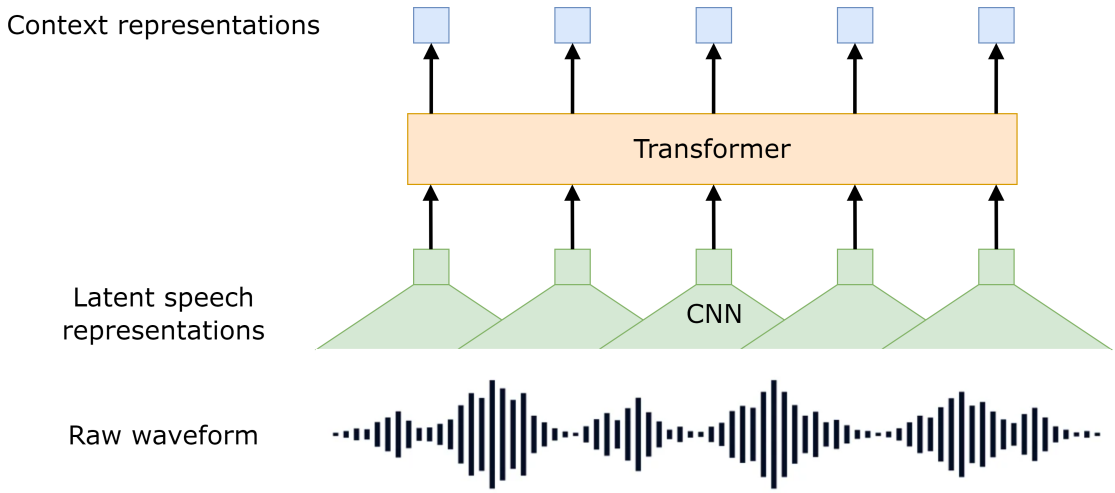


Figure 2.1: Framework for wav2vec2, adopted from [1].

### 2.1.6. Fisher Vector Representation

Fisher Vector [24] is a coding method that enables normalization for unequal feature matrices. For a segment of the speech signal, MFCC features can be extracted on each frame. The unequal length of each speech signal results in an unequal total number of frames per speech segment. When the features are fed into the network for speech recognition, the features are generally normalized into a feature matrix of uniform size. Fisher Vector is the conventional means of processing. Fisher Vector is obtained by taking partial derivatives of multiple Gaussian distributions with respect to weights, means, and variances.

Fisher vectors are a type of feature representation that are used in computer vision and machine learning. They are based on the Fisher information matrix, which is a measure of the local variance of a distribution.

Fisher vector (FV) is a superframe encoding that quantifies the gradient of background model parameters with respect to data. Given a parameterized probability model $\theta$, the expected Fisher information matrix $F(\theta)$ is the expectation of the second derivative of the log likelihood with respect to $\theta$:

$$\text{Fisher information matrix:} \quad F(\theta) = -\mathbb{E}[\frac{\partial^2 \log p(X|\theta)}{\partial \theta^2}]. \tag{2.1}$$

### 2.1.7. Public Dimensional Emotion Model

Public Dimensional Emotion Model comes from the papar named *Dawn of the transformer era in speech emotion recognition: closing the valence gap* by Johannes Wagner et al [16]. They presented this model when proposing an online expectation maximization algorithm to jointly estimate the parameters of the code as well as the parameters of the neural network. They developed an unsupervised partial detection method that combines Fisher Vector Encoding (FVE) with Convolutional Neural Networks (CNN) that can improve the recognition rate. The architecture of Wav2Vec2consists of two main parts: convolutional neural network (CNN) and transformer. During the training process, the CNN part is used to extract the features of the speech signal, and the input raw audio data is converted into a log-Mel spectrogram as the input to the transformer. The CNN uses a 14-layer architecture. The basic architecture of the transformer part consists of 12 transformer layers, each with 1024 hidden units and a parametric number of 317 M. The transformer model is pre-trained by self-supervision.

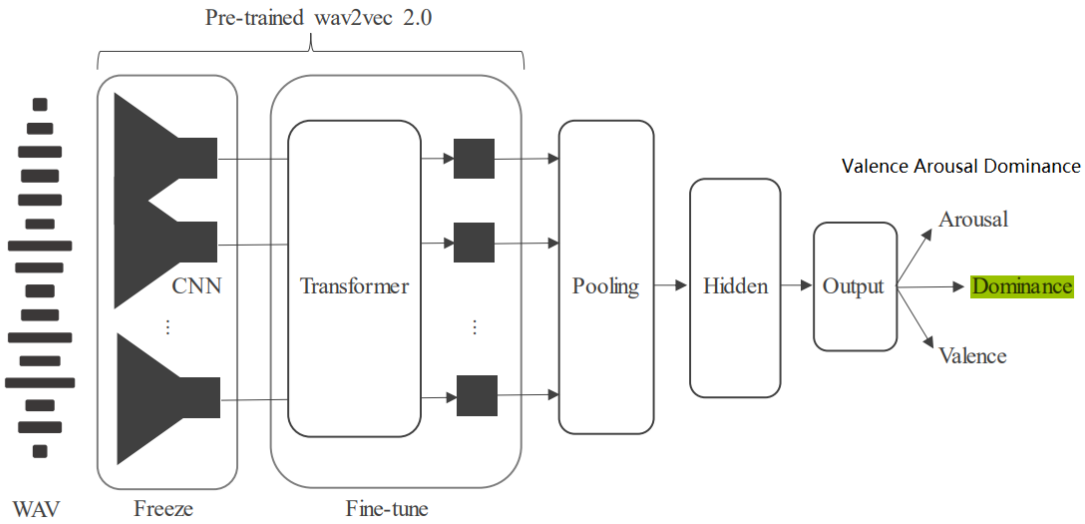


Figure 2.2: Pipeline of Public Dimensional Emotion Model

## 2.2. Background on Affective Constructs

### 2.2.1. The Big Five Model

The Big-Five model, proposed by the famous American psychologist McCrae et al [25], is widely used to describe human personality. The model describes human personality through the following five dimensions, with specific characteristics.

1. Openness (O): artistry, curiosity, imagination, insight, originality, etc.
2. Conscientiousness (C): efficiency, organization, planning, reliability, responsibility, organization, planning, reliability, responsibility, thoroughness, etc.
3. Extroversion (E): positive, confident, energetic, outgoing, talkative, energetic, outgoing, talkative, etc.
4. Agreeableness (A): appreciative, kind, generous, tolerant, compassionate, generosity, tolerance, compassion, trust in others, etc.
5. Neuroticism (N): anxiety, self-pity, nervousness, sensitivity, instability, etc.

The Big Five Model is widely used in psychology and is considered to be a reliable and valid measure of personality. It is often used in research and in a variety of profes-

sional settings, including education, career counselling, and leadership development.

## 2.3. Personality Datasets

In recent years, researchers have created a number of personality recognition databases. Some of these representative databases are shown in Table 2.1. Here, we see the dominance of the western languages (mainly English), and that all corpora contain audio and video modalities while four of them include the text modality, namely the speech transcriptions. While UDIVA [26] (also collected by the ChaLearn team) has the largest FI corpus in terms of total duration, ChaLearn FI corpus [14] used in this thesis features the largest number of subjects and video clips.

Table 2.1: Overview of Publicly Available Multimodal Personality Trait and Impression Corpora.

| Corpus | Modalities | Conditions | Language | Evaluation | # Subjects | Duration, h |
|---|---|---|---|---|---|---|
| ELEA [27] | A, V | Office | French, English | Self | 148 | 10 |
| Hire Me [28] | A, V | Office | English | Self | 62 | 11 |
| YouTube vlogs [29] | A, V | In-the-Wild | English | Third-party | 442 | 48 |
| JOKER [30] | A, V | Office | English | Self | 37 | 8 |
| MHHRI [31] | A, V | Office | English | Self, familiar | 18 | 6 |
| ChaLearn FI V2 [14] | A, V, T | In-the-Wild | English | Third-party | 3060 | 41 |
| MULTISIMO [32] | A, V, T | Office | English | Self, familiar | 49 | 4 |
| UDIVA [26] | A, V, T | Office | Spanish, Catalan, English | Self, familiar | 147 | 90 |
| RoomReader [33] | A, V, T | In-the-Wild | English | Self, familiar | 118 | 8 |

## 2.4. Related Work on ChaLearn First Impression Corpus

The Chalearn LAP First Impressions dataset is a collection of 10,000 short videos (average duration 15s) of people facing and speaking in English to a camera. The videos are split into training, validation and test sets with a 3:1:1 ratio. People in videos show different gender, age, nationality, and ethnicity [34].

The videos are labeled with personality traits variables, including Extraversion, Agreeableness, Conscientiousness, Neuroticism and Openness. The annotations were generated using Amazon Mechanical Turk (AMT) and a principled procedure was

adopted to guarantee the reliability of labels.

In addition to the personality traits, the dataset also includes transcriptions of all words in the video clips and annotations indicating whether the person should be invited or not to a job interview. The labels for gender and ethnicity are also available for the First Impressions dataset. The labels were made available by Heysem Kaya and Albert Ali Salah.

In the Personality Trait challenge, we see that most approaches in the challenge utilized both audio and video modalities. The audio was often represented using handcrafted spectral features, but one team used a residual network [35]. For the video, convolutional neural networks were commonly used to learn representations. The modalities were usually fused together before being fed to regression methods like fully connected neural networks or Support Vector Regressors. One team included temporal structure by partitioning video sequences and feeding the learned audio-video representation to a recurrent Long Short Term Memory layer [36]. Many teams made semantic assumptions about the data by separating the face from the background, often through preprocessing techniques like face frontalisation. However, the NJU-LAMDA [37] did not make any semantic separation of the content. Pretrained deep models fine-tuned on the challenge dataset were commonly used. The winning team NJU-LAMDA [37] proposed two separate models for still images and audio and employed a two-step late fusion. The team evolgen [36] used a multimodal LSTM architecture to maintain temporal structure. DCC used separate auditory and visual streams with deep residual networks for each, followed by an audiovisual stream.

### 2.5. Audio-Based Personality Trait Recognition

Early audio features for automatic personality trait recognition were hand-crafted low-level descriptive features, such as prosody, voice quality, and spectral features. Mohammadi and Vinciarelli [38] used logistic regression to detect audio clips whether exceeded the average score for each of the Big-five personality traits after extracting

features of LLDs. An et al. [39] extracted 6,373 acoustic–prosodic features from audio clips and used them as input to a support vector machine (SVM) classifier to identify the Big Five personality traits. Carbonneau et al. [40] learned a discriminating feature dictionary from patches in the speech spectrograms, which were then used by an SVM classifier to classify the Big Five personality traits.

# 3. Methodology

For RQ1, I will explore the predictive performance of different acoustic feature representations for classifying Big Five personality trait impressions. The feature representations will include different LLDs (including traditional MFCC and state-of-the-art Wav2Vec2) and utterance representations (such as simple functionals and the Fisher Vectors). As shown in the figure below, first I will process the ChaLearn LAP-FI dataset, extract acoustic features and select them. At this stage, the tools for processing audio and the features selected are changeable. After that, the affective prediction model will be trained with these features.

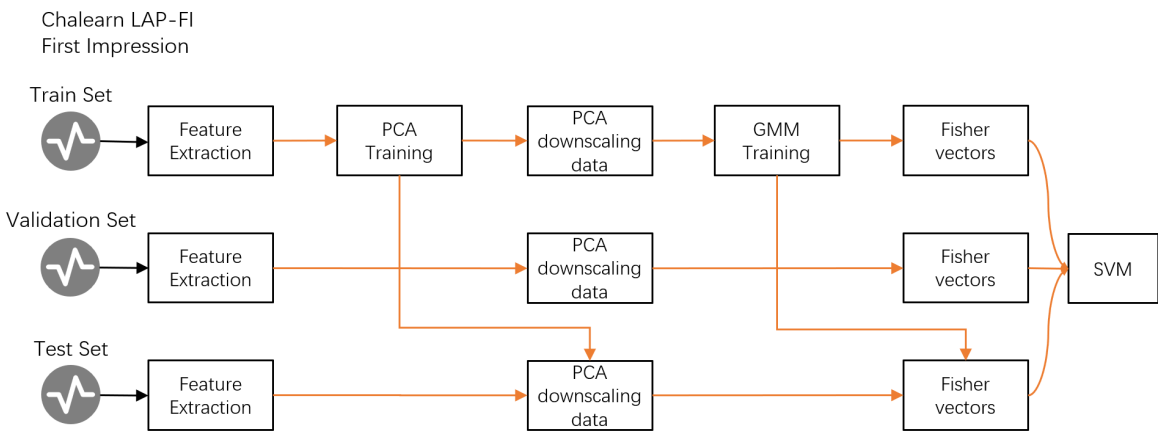


Figure 3.1: Flowchart of the FV representation based framework.

I will mainly compare the Wav2Vec2signal representation with Fisher vector encoding and a standard set of openSMILE features.

Combining the Wav2Vec2 signal representation with Fisher Vector (FV) encoding can take advantage of the strengths of both methods. The idea behind this combination is to use Wav2Vec2 to learn a set of fixed-dimensional representations (vectors) of audio segments that capture the underlying semantics of the audio signal, and then use the FV encoding to further compact and enhance the information in the Wav2Vec2

representations.

At the same time, we also used the LLDs extracted from the configuration of INERSPEECH 2013 ComparE to train the GMM model to generate the Fisher Vectors.

For Public Dimensional Emotion Model (PDEM), I not only extracted the final output, which is the intelligent emotion primitive features (arousal, valence, and dominance), but I also extracted its hidden states to be used as embedding for the classification task.

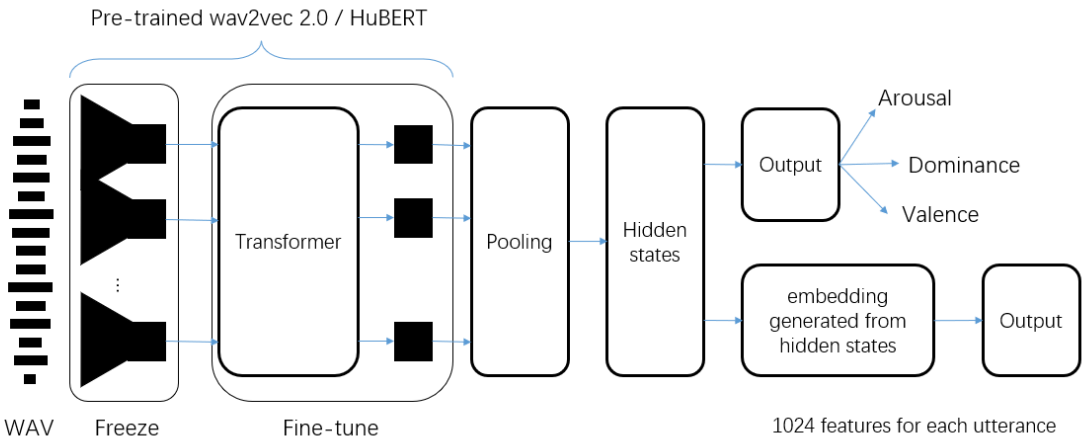


Figure 3.2: Flowchart of the representations from PDEM model.

## 3.1. Data Normalization

In data analysis, the different sources of data usually lead to the difference in the scale of the data. In order to make these data comparable, we need to introduce data preprocessing techniques to eliminate these differences. Normalization is to subtract by the minimum value divided by the variable range.

Z-Score normalization is a common method of data processing. It enables data of different scales to be transformed into a certain range for comparison. The range sets the mean is 0 and the standard deviation is 1. This can help to improve the performance like classifying and prevent over-fitting.

To z-score normalize a value $x$, we should first subtract the mean $\mu$ of the training data from $x$. Then we need to divide the resulting value by the standard deviation $\sigma$ of the data. This will give us a new value, $z(x)$, which has a mean of 0 and a standard deviation of 1.

$$\text{Z-score normalization:} \quad z(x) = \frac{x - \mu}{\sigma} \tag{3.1}$$

Z-score normalization can help to improve the performance of machine learning algorithms by making the data more consistent. This can help the algorithm to learn more effectively and to avoid over-fitting the training data.

In this experiment, I will use two data preprocessing methods, including Z-Score Normalization and L2 Normalization before doing the classification.

The first way is to only use Z-Score Normalization. Then I will combine Z-Score Normalization and L2 Normalization. Finally, the result with the higher accuracy of two methods on the test set will be taken.

# 4. Experimental Validation

In this chapter, I will first start by describing my pre-feature extraction and processing, dataset processing, and model training process in Sec. 6.1. Then I will present the results of all models trained for ChaLearn LAPFI First Impression Dataset Recognition in Sec. 6.2. In Sec. 6.3 I will discuss the performance of different models on personality recognition performances. For the sake of brevity and consistency, coding convention used for the generated feature sets as well as in the subsequent tables showing experimental results is in Table 4.1.

| Model and Parameter Names | Alias |
|---|---|
| Wav2Vec Uncompressed PCA400 GMM50 | W2VU_FV_PCA400_GMM50 |
| Wav2Vec Compressed PCA300 GMM50 | W2VC_FV_PCA300_GMM50 |
| Wav2Vec Compressed PCA400 GMM50 | W2VC_FV_PCA400_GMM50 |
| Wav2Vec Mean+Standard Deviation | W2V2_MS |
| OpenSMILE 130-dimensional Feature LLDs | IS13_LLD_MS |
| OpenSMILE 6373-dimensional Functionals (Baseline) | IS13_FUN |
| Public Dimensional Emotion Model VAD | PDEM-VAD |
| Public Dimensional Emotion Model Embeddings | PDEM-EBD |
| Valence Arousal Dominance | VAD |
| Average | AVG |
| Embedding | EBD |
| Openness | OPEN |
| Conscientiousness | CONS |
| Extraversion | EXTR |
| Agreeableness | AGRE |
| Neuroticism | NEUR |

Table 4.1: Model, Parameter Names and Their Corresponding Alias.

### 4.1. Experiment Setup and Results

I used the ChaLearn Looking at People (LAP) FI First Impression Dataset [14] to run different training on the big five dimensions of Openness, Conscientiousness, Extroversion, Agreeableness, and Neuroticism. The dataset consists of 10000 video clips average about 15 seconds in length. The dataset is divided into a training set of 6000 audios, a validation set of 2000 audios and a test set of 2000 audios. The original labels of each dimension cover a range from 0 to 1, which I binarized into 0 or 1 based on average of each dimension to do binary classification.

For the extraction of Wav2Vec2 LLDs, I obtain 768-dimensional embeddings from 32 ms of raw audio (waveform) with 10 ms steps. This process gives a time-sequenced set of LLDs for the whole clip. For each 15-second audio clip from ChaLearn FI corpus, this generates a sequence of approximately 1530 LLDs. To avoid memory issues in FV modeling, first, I subsample these taking one every 10 LLDs. Then these subsamples are combined into one large feature matrix. After that, I pre-train this feature matrix to generate the respective PCA model and GMM models. The PCA model is used to reduce the decorrelate and dimensionality of the original acoustic LLDs. After that, the PCA-reduced data per audio clip and the trained GMM model together used to generate the corresponding Fisher Vector features used for classification.

I used features extracted by openSMILE using a standard feature configuration used in the INTERSPEECH 2013 Computational Paralinguistics Challenge [41] as the baseline for my experiments. This contains 130 LLDs (65 raw and 65 temporal derivatives) that are summarized by 54 functionals, which in total yield 6373 functional features representing an audio utterance. For comparison under controlled settings, I also used the 130-dimensional LLDs from the same openSMILE configuration for FV representation. As shown earlier in Figure 3.1, FV representation for Wav2Vec2 follows similar steps.

Also, I used an external pre-training model Public Dimensional Emotion Model

as a reference. It extracts Valence, Arousal and Dominance (VAD) emotion primitives.

### 4.1.1. Fisher Vector vs. Baseline Functionals on openSMILE LLDs

To answer research sub-question 1 (RSQ1), we compared IS13_FUN (baseline) features with the FV representation of the same set of LLDs, namely, IS_13LLD_FV. The results are shown in Table 4.2. In terms of validation set accuracy, the IS_13LLD_FV encoding achieves a validation accuracy of 69.40% for OPEN, 70.90% for CONS, 68.80% for EXTR, 63.95% for AGRE, and 69.45% for NEUR. The average validation accuracy across all emotion dimensions is 68.50%. On the other hand, the IS13FUN encoding achieves slightly lower validation accuracy, with scores of 69.05% for OPEN, 67.55% for CONS, 68.15% for EXTR, 62.15% for AGRE, and 67.65% for NEUR. The average validation accuracy for this encoding is 66.91%. In terms of test set accuracy, the IS13_LLD_FV encoding achieves accuracies of 70.85% for OPEN, 69.15% for CONS, 69.35% for EXTR, 64.70% for AGRE, and 70.60% for NEUR. The average test accuracy for this encoding is 68.93%. Similarly, the IS_13_FUN encoding achieves accuracies of 68.70% for OPEN, 70.00% for CONS, 69.30% for EXTR, 64.50% for AGRE, and 69.05% for NEUR on the test set. The average test accuracy for this encoding is 68.31%. Among the encoding techniques, the IS_13_LLD_FV encoding achieves higher accuracy on both the validation and test sets, compared to the IS_13_FUN encoding. It demonstrates the state-of-the-art level of classification in this evaluation.

| Validation Set | OPEN | CONS | EXTR | AGRE | NEUR | AVG |
|---|---|---|---|---|---|---|
| IS113_LLD_FV | 69.40 | 70.90 | 68.80 | 63.95 | 69.45 | 68.50 |
| IS13_FUN | 69.05 | 67.55 | 68.15 | 62.15 | 67.65 | 66.91 |
| Test Set | OPEN | CONS | EXTR | AGRE | NEUR | AVG |
| IS13_LLD_FV | 70.85 | 69.15 | 69.35 | 64.70 | 70.60 | 68.93 |
| IS13_FUN | 68.70 | 70.00 | 69.30 | 64.50 | 69.05 | 68.31 |

Table 4.2: Validation and test set accuracy (%) performance of different encoding on openSMILE.

### 4.1.2. Effect of Wav2vec2 LLD Compression in FV Modeling

The classification results comparing compression options for Wav2Vec2 embeddings are shown in Table 4.3. The W2VU_FV_PCA400_GMM50 model demonstrated superior accuracy in predicting the various personality dimensions, particularly excelling in the Neuroticism dimension. However, it had the least accurate predictions in the Agreeableness dimension. On the other hand, the W2VC_FV_PCA400_GMM50 model had slightly lower accuracy scores overall compared to the W2VU_FV_PCA400_GMM50 model. However, it showcased more consistent performance across all personality dimensions. In summary, the W2VU_FV_PCA400_GMM50 model displayed higher accuracy in most dimensions except for Agreeableness, while the W2VC_FV_PCA400_GMM50 model depicted more stable performance overall across the personality dimensions.

| Validation set | OPEN | CONS | EXTR | AGRE | NEUR | AVG |
|---|---|---|---|---|---|---|
| W2VU_FV_PCA400_GMM50 | 64.55 | 67.15 | 65.45 | 59.35 | 66.60 | 64.62 |
| W2VC_FV_PCA400_GMM50 | 63.45 | 66.80 | 63.70 | 61.50 | 64.75 | 64.04 |
| Test set | OPEN | CONS | EXTR | AGRE | NEUR | AVG |
| W2VU_FV_PCA400_GMM50 | 66.15 | 68.88 | 65.75 | 62.50 | 66.95 | 66.05 |
| W2VC_FV_PCA400_GMM50 | 65.55 | 68.80 | 64.10 | 63.35 | 66.70 | 65.70 |

Table 4.3: Validation and test set accuracy (%) performance of compressed and uncompressed Wav2Vec2 LLDs with FV encoding.

### 4.1.3. Fisher Vector vs. Mean+Std. Functionals on Wav2Vec2

Table 4.4 represents the performance evaluation results for two different utterance representation methods, namely FV and simple functionals (mean + std.), applied on Wav2Vec2 LLDs. Each row represents a specific trait model, including W2VU_FV_PCA400_GMM50 and W2V2_MS. The columns display the validation accuracy and test accuracy. Based on the validation accuracy, the W2VU_FV_PCA400_GMM50 and W2V2MS models have varying performance across different personality trait dimensions. The W2V2MS model performs better in the OPEN, CONS, and EXTR

trait dimensions, with validation accuracies of 65.50%, 68.85%, and 65.20% respectively. In the AGRE and NEUR emotion dimensions, the W2VU_FV_PCA400_GMM50 model has slightly higher performance. Overall, the average accuracy indicates that the W2V2MS model has better performance, with an average accuracy of 65.66%. In terms of test accuracy, the performance trends of the two models across different trait dimensions are similar to the validation accuracy. The W2VU_FV_PCA400_GMM50 model has slightly higher test accuracies in the OPEN and AGRE emotion dimensions compared to the W2V2_MS model. However, the W2V2_MS model performs slightly better in the CONS, EXTR, and NEUR emotion dimensions. Overall, the average accuracy indicates that the W2V2_MS model has slightly better performance on the test set, with an average accuracy of 66.19%. In conclusion, the W2V2_MS model shows relatively better performance on this dataset, with a higher average accuracy. The overall results show that the high-dimensional and computationally complex FV representation performs either on par with or (mostly) poorer compared to a simple use of two functionals on Wav2Vec2.

| Validation Set | OPEN | CONS | EXTR | AGRE | NEUR | AVG |
|---|---|---|---|---|---|---|
| W2VU_FV_PCA400_GMM50 | 64.55 | 67.15 | 65.45 | 59.35 | 66.60 | 64.62 |
| W2V2_MS | 64.50 | 65.90 | 65.20 | 61.95 | 67.20 | 64.95 |
| Test Set | OPEN | CONS | EXTR | AGRE | NEUR | AVG |
| W2VU_FV_PCA400_GMM50 | 66.15 | 68.88 | 65.75 | 62.50 | 66.95 | 66.05 |
| W2V2_MS | 66.15 | 69.30 | 65.35 | 63.65 | 67.55 | 66.40 |

Table 4.4: Validation and test set accuracy (%) performance of Fisher Vector vs. mean+std. functionals on Wav2Vec2.

### 4.1.4. Wav2Vec2 Fisher Vector vs. openSMILE Baseline

Table 4.5 presents the validation and test set accuracy performance of different encoding techniques, including W2VU_FV_PCA400_GMM50, W2V2_MS, and IS13_FUN, on the Wav2vec vs Baseline task. In terms of validation set accuracy, the W2VU_FV_PCA400_GMM50 encoding achieves a validation accuracy of 64.55% for OPEN, 67.15% for CONS, 65.45% for EXTR, 59.35% for AGRE, and 66.60% for

NEUR. The average validation accuracy across all emotion dimensions is 64.62%. The W2V2MS encoding performs slightly better, with validation accuracies of 65.50% for OPEN, 68.85% for CONS, 65.20% for EXTR, 62.25% for AGRE, and 66.50% for NEUR. The average validation accuracy for this encoding is 65.66%. The IS13FUN encoding achieves a higher validation accuracy compared to the two Wav2vec encodings, with scores of 69.05% for Based on these results, it appears that IS13_FUN performs the best in terms of accuracy scores for most dimensions on both the validation and testing sets. W2V2_MS also demonstrates fairly high accuracy scores, while W2VU_FV_PCA400_GMM50 has relatively lower scores. It is seen that the combination of fisher vector, Chalearn LAP-FI and Wav2vec is not suitable for binary classification.

| Validation Set | OPEN | CONS | EXTR | AGRE | NEUR | AVG |
|---|---|---|---|---|---|---|
| W2VU_FV_PCA400_GMM50 | 64.55 | 67.15 | 65.45 | 59.35 | 66.60 | 64.62 |
| W2V2_MS | 64.50 | 65.90 | 65.20 | 61.95 | 67.20 | 64.95 |
| IS13_FUN | 69.05 | 67.55 | 68.15 | 62.15 | 67.65 | 66.91 |
| Test Set | OPEN | CONS | EXTR | AGRE | NEUR | AVG |
| W2VU_FV_PCA400_GMM50 | 66.15 | 68.88 | 65.75 | 62.50 | 66.95 | 66.05 |
| W2V2_MS | 66.15 | 69.30 | 65.35 | 63.65 | 67.55 | 66.40 |
| IS13_FUN | 68.70 | 70.00 | 69.30 | 64.50 | 69.05 | 68.31 |

Table 4.5: Validation and test set accuracy (%) performance of different encoding on Wav2vec vs. Baseline.

### 4.1.5. Public Dimensional Emotion Model vs. openSMILE Features

In terms of validation accuracy, the performance of the models varies across different emotion dimensions. Among them, the PDEM-EBD model achieves the highest validation accuracy in the OPEN and AGRE emotion dimensions, with accuracies of 70.60% and 71.80% respectively. For the EXTR and CONS emotion dimensions, the PDEM-EBD and PDEM-VAD models perform equally well. In the NEUR emotion dimension, the PDEM-EBD and IS113LLDFV models have similar performance. Re-

garding the test accuracy, the results follow a similar trend as the validation accuracy. The PDEM-EBD model still performs the best in the OPEN and AGRE emotion dimensions, with test accuracies of 70.10% and 71.80% respectively. For the EXTR and CONS emotion dimensions, the PDEM-EBD and IS13FUN models have similar performance. In the NEUR emotion dimension, the PDEM-EBD and IS13FUN models perform equally well. Overall, the PDEM-EBD model demonstrates the highest performance across multiple emotion dimensions, with relatively high test accuracy.

| Validation Set | OPEN | CONS | EXTR | AGRE | NEUR | AVG |
|---|---|---|---|---|---|---|
| IS113_LLD_FV | 69.40 | 70.90 | 68.80 | 63.95 | 69.45 | 68.50 |
| IS13_FUN | 69.05 | 67.55 | 68.15 | 62.15 | 67.65 | 66.91 |
| PDEM-VAD | 65.30 | 64.35 | 66.60 | 60.40 | 67.10 | 64.75 |
| PDEM-EBD | 70.60 | 71.20 | 70.80 | 65.40 | 70.75 | 69.75 |
| Test Set | OPEN | CONS | EXTR | AGRE | NEUR | AVG |
| IS13_LLD_FV | 70.85 | 69.15 | 69.35 | 64.70 | 70.60 | 68.93 |
| IS13_FUN | 68.70 | 70.00 | 69.30 | 64.50 | 69.05 | 68.31 |
| PDEM-VAD | 65.00 | 63.10 | 64.90 | 60.45 | 67.00 | 64.09 |
| PDEM-EBD | 70.10 | 72.05 | 71.25 | 65.60 | 71.80 | 70.16 |

Table 4.6: Validation and test set accuracy (%) performance of different features from PDEM and openSMILE.

## 4.2. Statistical tests

As the results of the different feature sets are discussed, now I will use McNemar's test to compare systems in each personality trait to answer my research questions. The tables provide an overview of the statistical tests performed for each personality trait, including Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism.

| Comparison for Openness | $p$ | $\chi^2$ | Opponent Test |
|---|---|---|---|
| RSQ1- Baseline vs. IS13_LLD_FV | <0.05 | 15.84 | 70.85 |
| RSQ2- CONP vs. UNCONP | <0.05 | 34.40 | 66.15 |
| RSQ3- W2V2_MS vs. Wav2Vec2_FV | 0.668 | 0.1875 | 66.15 |
| RSQ4- Baseline vs. Wav2Vec2_FV | <0.05 | 24.43 | 66.15 |
| RSQ5- Baseline vs. PDEM-EBD | <0.05 | 19.00 | 70.10 |
| RQ2- Baseline vs. PDEM-VAD | <0.05 | 12.34 | 65.00 |

Table 4.7: Overview of statistical tests performed for Openness with the corresponding test set accuracy (%) performances. The baseline (IS13_FUN) test performance for this dimension is 68.70%.

| Comparison for Conscientiousness | $p$ | $\chi^2$ | Opponent Test |
|---|---|---|---|
| RSQ1- Baseline vs. IS13_LLD_FV | <0.05 | 20.75 | 70.90 |
| RSQ2- CONP vs. UNCONP | 0.059 | 3.58 | 68.80 |
| RSQ3- W2V2_MS vs. Wav2Vec2_FV | <0.05 | 11.42 | 68.88 |
| RSQ4- Baseline vs. Wav2Vec2_FV | 0.257 | 1.14 | 68.88 |
| RSQ5- Baseline vs. PDEM-EBD | <0.05 | 19.00 | 71.20 |
| RQ2- Baseline vs. PDEM-VAD | <0.05 | 12.34 | 64.35 |

Table 4.8: Overview of statistical tests performed for Conscientiousness with the corresponding test set accuracy (%) performances. The baseline (IS13_FUN) test performance for this dimension is 70.00%.

## 4.3. Discussion

NOw, I will answer the question proposed at the beginning. Question 1: Concerning the apparent personality trait recognition task via Big Five personality traits, is there a significant difference in terms of test set accuracy performance between baseline (Interspeech 2013 Computational Paralinguistics ChallengE Setting) acoustic features and other feature representations?

- Sub-question 1.1: Is there a significant difference in terms of test set accuracy

| Comparison for Extraversion | $p$ | $\chi^2$ | Opponent Test |
|---|---|---|---|
| RSQ1- Baseline vs. IS13_LLD_FV | 0.611 | 0.26 | 69.35 |
| RSQ2- CONP vs. UNCONP | 0.328 | 0.9575 | 64.10 |
| RSQ3- W2V2_MS vs. Wav2Vec2_FV | 0.1 | 1.106 | 65.75 |
| RSQ4- Baseline vs. Wav2Vec2_FV | <0.05 | 12.54 | 65.75 |
| RSQ5- Baseline vs. PDEM-EBD | 0.552 | 0.36 | 71.25 |
| RQ2- Baseline vs. PDEM-VAD | <0.05 | 7.32 | 64.90 |

Table 4.9: Overview of statistical tests performed for Extraversion with the corresponding test set accuracy (%) performances. The baseline (IS13_FUN) test performance for this dimension is 69.30%.

| Comparison for Agreeableness | $p$ | $\chi^2$ | Opponent Test |
|---|---|---|---|
| RSQ1- Baseline vs. IS13_LLD_FV | 0.602 | 0.28 | 64.70 |
| RSQ2- CONP vs. UNCONP | 0.102 | 2.94 | 62.50 |
| RSQ3- W2V2_MS vs. Wav2Vec2_FV | <0.05 | 17.191 | 63.75 |
| RSQ4- Baseline vs. Wav2Vec2_FV | 0.809 | 0.06 | 63.75 |
| RSQ5- Baseline vs. PDEM-EBD | 0.436 | 0.62 | 65.60 |
| RQ2- Baseline vs. PDEM-VAD | 0.153 | 2.03 | 60.45 |

Table 4.10: Overview of statistical tests performed for Agreeableness with the corresponding test set accuracy (%) performances. The baseline (IS13_FUN) test performance for this dimension is 64.50%.

performance between baseline acoustic features and FV representation of the same set of openSMILE low-level descriptors (LLDs)?

OpenSmile LLD fisher vector is slightly better than Baseline in terms of average performance. where there is a statistically significant advantage in Openness, Conscienciousness.

- Sub-question 1.2: Is there a significant difference in terms of test set accuracy performance between FV representations of the original (uncompressed) and the compressed (averaged per 5 consecutive, non-overlapping frames) Wav2Vec2 LLDs?

| Comparison for Neuroticism | $p$ | $\chi^2$ | Opponent Test |
|---|---|---|---|
| RSQ1- Baseline vs. IS13_LLD_FV | 0.63 | 0.22 | 70.60 |
| RSQ2- CONP vs. UNCONP | <0.05 | 5.33 | 66.70 |
| RSQ3- W2V2_MS vs. Wav2Vec2_FV | <0.05 | 2.463 | 66.95 |
| RSQ4- Baseline vs. Wav2Vec2_FV | <0.05 | 6.65 | 66.95 |
| RSQ5- Baseline vs. PDEM-EBD | <0.05 | 23.27 | 71.80 |
| RQ2- Baseline vs. PDEM-VAD | <0.05 | 4.29 | 67.00 |

Table 4.11: Overview of statistical tests performed for Neuroticisms with the corresponding test set accuracy (%) performances. The baseline (IS13_FUN) test performance for this dimension is 69.05%.

There is not much difference in average performance between the two classifiers used with or without compression of training data. Only on Openness, Uncompressed Model performs slightly better.

- Sub-question 1.3: Is there a significant difference in terms of test set accuracy performance between functional representation via mean and standard deviation summarization of the Wav2Vec2 LLDs and their FV representation?

  There is not much difference in average performance between the two classifiers. However, for Conscienciousness, Agreebleness and Neuroticism, Wav2vec vectors perform not as good as mean and standard deviation summarization. In other personality dimensions, there is no significant difference.

- Sub-question 1.4: Is there a significant difference in terms of test set accuracy performance between baseline acoustic features and the FV representation of Wav2Vec2 LLDs?

  There is a significant difference in terms of test set accuracy performance between baseline acoustic features and the FV representation of Wav2Vec2 LLDs. The FV representation of Wav2Vec2 LLDs outperforms the baseline acoustic features in all five personality traits, with statistically significant differences in Openness, Conscientiousness, Agreebleness and Neuroticism.

- Sub-question 1.5: Is there a significant difference in terms of test set accuracy performance between baseline acoustic features and acoustic embeddings from

the public dimensional emotion model (PDEM)?

There is a significant difference in terms of test set accuracy performance between baseline acoustic features and acoustic embeddings from the public dimensional emotion model (PDEM). The PDEM acoustic embeddings outperform the baseline acoustic features in all five personality traits, with statistically significant differences in Openness, Conscientiousness and Neuroticism.

- Question 2: Concerning the apparent personality trait recognition task via Big Five personality traits, is there a significant difference in terms of test set accuracy performance between baseline acoustic features and indirect modeling via intelligible emotion primitive features (arousal, valence, and dominance) extracted from public dimensional emotion model (PDEM)?

  For every trait except agreeableness, PDEM is statistically significantly better. The average performance of PDEM is much better.

# 5. Conclusion

This thesis focuses on several audio feature engineering methods, mainly on fisher vector, to predict the personality five dimensions. Based on my experimental results, for Chalearn Lapfi First Impression audio data and Wav2vec specialization extraction, the performance using fisher vector did not meet expectations. For Chalearn lapfi's single-modal audio divergence, openSMILE low level discriptors(LLDs) based on Inter-Speech 2013 configuration, and fisher-vector-based embeddings extracted by external model Personal dimensional emotion modal can reach the state-of-the-art level. Meanwhile, this thesis also explores the differences in the selection of parameters in some feature engineering practical experience. Also, I explore whether it can be used as an interpretable for the mood prediction of personality five-dimensional model based on valence, arousal and dominance model.

## 5.1. Limitations and future work

The one of methods I tested which should be the main breakthrough, combing Fisher Vectors and SVM performed well on other datasets [42], but not on this the current data set Chalearn Lap FI. This suggests that the combination of these techniques may not be universally applicable and that further research is needed to understand the factors that contribute to its performance.

One possible explanation for the poor performance on this dataset is the Fisher Vector technique may not be well-suited for this particular dataset, as it is designed for tasks such as image classification, where the features are more evenly distributed.

Despite the poor performance of this dataset, we believe that the method we tested has the potential to be a powerful tool for classification. Further research is needed to address the limitations that we have identified and to improve its performance on a wider range of datasets.

This project is a pre-task for a large project aimed at finding an interpretable machine learning intermediate feature to address the diagnosis of mood disorders and depression. However, existing automated depression severity prediction methods often rely on deep learning and black-box models that lack explanatory and interpretable considerations. In order to meet clinicians' needs to understand the decision-making process of models, we need to build interpretable models with competitive performance. Therefore, we introduce the Big Five personality traits as an interpretable intermediate. After finding the intermediate, more in-depth analyses and explorations can be conducted to understand the relationship between these traits and depression. Further empirical research could lead to a better understanding of the role of these features in the diagnosis of depression and validate their predictive power for depression. Beyond depression diagnosis, extending the application of these intermediate traits to other mental health domains or medical treatments. For example, explore potential applications of intermediate features in the diagnosis of anxiety disorders, schizophrenia, or other mental health problems. Algorithms and models can be further improved for intermediate features to increase their performance and accuracy. This may include the use of more advanced machine learning methods, model integration or deep learning techniques to further optimise the diagnosis of depression.

# REFERENCES


1. Krebbers, D., H. Kaya and A. Karpov, "Multi-level Fusion of Fisher Vector Encoded BERT and Wav2vec 2.0 Embeddings for Native Language Identification", S. R. M. Prasanna, A. Karpov, K. Samudravijaya and S. S. Agrawal (Editors), *Speech and Computer*, pp. 391–403, Springer International Publishing, Cham, 2022.

2. Schneider, B., D. B. Smith, S. Taylor and J. Fleenor, "Personality and organizations: A test of the homogeneity of personality hypothesis.", *Journal of Applied Psychology*, Vol. 83, No. 3, p. 462, 1998.

3. Goldberg, L. R., "The development of markers for the Big-Five factor structure.", *Psychological assessment*, Vol. 4, No. 1, p. 26, 1992.

4. Cattell, H. E. and A. D. Mead, "The sixteen personality factor questionnaire (16PF)", *The SAGE handbook of personality theory and assessment*, Vol. 2, pp. 135–159, 2008.

5. Davies, M. F., C. C. French and E. Keogh, "Self-deceptive enhancement and impression management correlates of EPQ-R dimensions", *The Journal of Psychology*, Vol. 132, No. 4, pp. 401–406, 1998.

6. Dennis, B., M. Charney, J. C. Nelson, D. M. Quinlan *et al.*, "Personality traits and disorder in depression", *American Journal of Psychiatry*, Vol. 138, p. 1601, 1981.

7. Cloninger, C. R., D. M. Svrakic and T. R. Przybeck, "Can personality assessment predict future depression? A twelve-month follow-up of 631 subjects", *Journal of affective disorders*, Vol. 92, No. 1, pp. 35–44, 2006.

8. Quilty, L. C., M. Sellbom, J. L. Tackett and R. M. Bagby, "Personality trait predictors of bipolar disorder symptoms", *Psychiatry Research*, Vol. 169, No. 2,

pp. 159–163, 2009.

9. Klein, D. N., R. Kotov and S. J. Bufferd, "Personality and depression: Explanatory models and review of the evidence", *Annual Review of Clinical Psychology*, Vol. 7, pp. 269–295, 2011.

10. Kotov, R., W. Gamez, F. Schmidt and D. Watson, "Linking "Big" personality traits to anxiety, depressive, and substance use disorders: a meta-analysis.", *Psychological bulletin*, Vol. 136, No. 5, p. 768, 2010.

11. Malouff, J. M., E. B. Thorsteinsson and N. S. Schutte, "The relationship between the five-factor model of personality and symptoms of clinical disorders: A meta-analysis", *Journal of psychopathology and behavioral assessment*, Vol. 27, No. 2, pp. 101–114, 2005.

12. Zhao, X., Z. Tang and S. Zhang, "Deep Personality Trait Recognition: A Survey", *Frontiers in Psychology*, p. 2390, 2022.

13. Kaya, H. and A. A. Salah, "Multimodal personality trait analysis for explainable modeling of job interview decisions", *Explainable and Interpretable Models in Computer Vision and Machine Learning*, pp. 255–275, 2018.

14. Escalante, H. J., H. Kaya, A. A. Salah, S. Escalera, Y. Güçlütürk, U. Güçlü, X. Baró, I. Guyon, J. C. S. J. Junior, M. Madadi, S. Ayache, E. Viegas, F. Gürpınar, A. S. Wicaksana, C. C. S. Liem, M. A. J. van Gerven and R. van Lier, "Modeling, Recognizing, and Explaining Apparent Personality From Videos", *IEEE Transactions on Affective Computing*, Vol. 13, No. 2, pp. 894–911, 2022.

15. Goddard, M., "The EU General Data Protection Regulation (GDPR): European regulation that has a global impact", *International Journal of Market Research*, Vol. 59, No. 6, pp. 703–705, 2017.

16. Wagner, J., A. Triantafyllopoulos, H. Wierstorf, M. Schmitt, F. Burkhardt, F. Ey-

ben and B. W. Schuller, "Dawn of the Transformer Era in Speech Emotion Recognition: Closing the Valence Gap", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–13, 2023.

17. Kim, S. and W. Lee, "Does McNemar's test compare the sensitivities and specificities of two diagnostic tests?", *Statistical methods in medical research*, Vol. 26, No. 1, pp. 142–154, 2017.

18. Schuller, B., S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. MüLler and S. Narayanan, "Paralinguistics in speech and language—state-of-the-art and the challenge", *Computer Speech & Language*, Vol. 27, No. 1, pp. 4–39, 2013.

19. Roach, P., R. Stibbard, J. Osborne, S. Arnfield and J. Setter, "Transcription of prosodic and paralinguistic features of emotional speech", *Journal of the International Phonetic Association*, Vol. 28, No. 1-2, pp. 83–94, 1998.

20. Eyben, F., M. Wöllmer and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor", *Proceedings of the 18th ACM international conference on Multimedia*, pp. 1459–1462, 2010.

21. Wang, J.-C., J.-F. Wang, K. W. He and C.-S. Hsu, "Environmental sound classification using hybrid SVM/KNN classifier and MPEG-7 audio low-level descriptor", *The 2006 IEEE international joint conference on neural network proceedings*, pp. 1731–1735, IEEE, 2006.

22. Schuller, B., S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi *et al.*, "The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism", *Proceedings INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France*, 2013.

23. Baevski, A., Y. Zhou, A. Mohamed and M. Auli, "wav2vec 2.0: A Framework for

Self-Supervised Learning of Speech Representations", *Advances in Neural Information Processing Systems*, Vol. 33, pp. 12449–12460, 2020.

24. Lapuschkin, S., A. Binder, G. Montavon, K.-R. Muller and W. Samek, "Analyzing classifiers: Fisher vectors and deep neural networks", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2912–2920, 2016.

25. McCrae, R. R. and O. P. John, "An introduction to the five-factor model and its applications", *Journal of personality*, Vol. 60, No. 2, pp. 175–215, 1992.

26. Palmero, C., J. Selva, S. Smeureanu, J. Junior, C. Jacques, A. Clapés, A. Moseguí, Z. Zhang, D. Gallardo, G. Guilera *et al.*, "Context-aware personality inference in dyadic scenarios: Introducing the udiva dataset", *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1–12, 2021.

27. Sanchez-Cortes, D., O. Aran and D. Gatica-Perez, "An audio visual corpus for emergent leader analysis", *Workshop on multimodal corpora for machine learning: taking stock and road mapping the future, ICMI-MLMI*, 2011.

28. Nguyen, L. S., D. Frauendorfer, M. S. Mast and D. Gatica-Perez, "Hire me: Computational inference of hirability in employment interviews based on nonverbal behavior", *IEEE transactions on multimedia*, Vol. 16, No. 4, pp. 1018–1031, 2014.

29. Biel, J.-I. and D. 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.

30. Devillers, L., S. Rosset, G. D. Duplessis, M. A. Sehili, L. Béchade, A. Delaborde, C. Gossart, V. Letard, F. Yang, Y. Yemez *et al.*, "Multimodal data collection of human-robot humorous interactions in the joker project", *2015 international conference on affective computing and intelligent interaction (ACII)*, pp. 348–354, IEEE, 2015.

31. Celiktutan, O., E. Skordos and H. 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.

32. Koutsombogera, M. and C. Vogel, "Modeling collaborative multimodal behavior in group dialogues: The MULTISIMO corpus", *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

33. Reverdy, J., S. O. Russell, L. Duquenne, D. Garaialde, B. R. Cowan and N. Harte, "RoomReader: A Multimodal Corpus of Online Multiparty Conversational Interactions", *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pp. 2517–2527, 2022.

34. Ponce-López, V., B. Chen, M. Oliu, C. Corneanu, A. Clapés, I. Guyon, X. Baró, H. J. Escalante and S. Escalera, "Chalearn lap 2016: First round challenge on first impressions-dataset and results", *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14*, pp. 400–418, Springer, 2016.

35. Güçlütürk, Y., U. Güçlü, M. A. van Gerven and R. van Lier, "Deep impression: Audiovisual deep residual networks for multimodal apparent personality trait recognition", *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14*, pp. 349–358, Springer, 2016.

36. Subramaniam, A., V. Patel, A. Mishra, P. Balasubramanian and A. Mittal, "Bimodal first impressions recognition using temporally ordered deep audio and stochastic visual features", *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14*, pp. 337–348, Springer, 2016.

37. Zhang, C.-L., H. Zhang, X.-S. Wei and J. Wu, "Deep bimodal regression for appar-

ent personality analysis", *European conference on computer vision*, pp. 311–324, Springer, 2016.

38. Mohammadi, G. and A. 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.

39. An, G., S. I. Levitan, R. Levitan, A. Rosenberg, M. Levine and J. Hirschberg, "Automatically Classifying Self-Rated Personality Scores from Speech.", *Interspeech*, pp. 1412–1416, 2016.

40. Carbonneau, M.-A., E. Granger, Y. Attabi and G. Gagnon, "Feature learning from spectrograms for assessment of personality traits", *IEEE Transactions on Affective Computing*, Vol. 11, No. 1, pp. 25–31, 2017.

41. Schuller, B., S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, M. Mortillaro, H. Salamin, A. Polychroniou, F. Valente and S. Kim, "The INTERSPEECH 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism", *Proc. Interspeech 2013*, pp. 148–152, 2013.

42. Gosztolya, G., "Using the fisher vector representation for audio-based emotion recognition", *Acta Polytechnica Hungarica*, Vol. 17, No. 6, pp. 7–23, 2020.