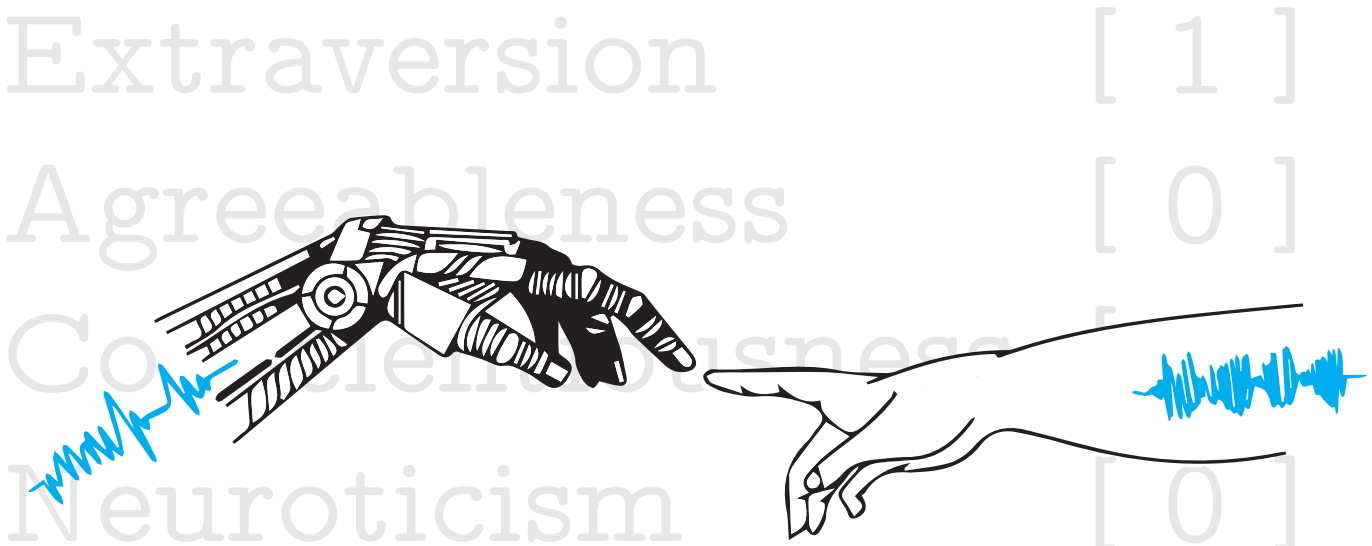
TU Delft



An Exploration of Automatic Personality Classification Using Different Speech Styles

Emile Lampe

Delft University of Technology

Faculty of Electrical Engineering, Mathematics and Computer Science
Multimedia Computing

Master Thesis

An Exploration of Automatic Personality Classification Using Different Speech Styles

Emile Lampe

Student number 4451090

Thesis committee

Dr Odette Scharenborg
Multimedia Computing Group, supervisor
Delft University of Technology

Dr Javier Marín Morales
Institute Human-Tech, supervisor
Universitat Politècnica de València

Dr Chirag Raman
Pattern Recognition and Bioinformatics Group
Delft University of Technology

June 13, 2023

Abstract

The field of speech-based Personality Computing classifies personality traits using speech data. There are two labelling methods for this: Automatic Personality Recognition (APR), using self-assessed personality scores, and Automatic Personality Perception (APP), using externally rated personality scores. Another aspect is whether the data is recorded in natural circumstances or in a controlled environment, as this influences how personality is shown. There is a lack of research into these speech styles, especially when combined with the labelling methods. Related fields have been found to be more developed in two ways. First, research from the perspective of speech styles has already been conducted and proven useful. Second, when state-of-the-art techniques are released, such as pretrained models targetting speech, these fields are often included in benchmark tests. As no personality datasets are included, this creates a knowledge gab on using these techniques for personality classification.

The influence of the labelling methods and speech styles is investigated using three datasets that target APR with controlled and natural speech, and APP with natural speech. Three types of models are used to see what personality traits can successfully be classified. The two APR datasets have not been used for personality classification before. Additionally, the models are trained using both hand-crafted features and embeddings extracted from a state-of-the-art pretrained model. The experiment on the APP and natural speech dataset indicates that the performance for 3 out of 5 traits can be improved using more effective features. The APR and controlled speech dataset was able to classify 4 out of 5 traits above chance. The APR and natural speech dataset could not well be classified. Overall, the APR datasets performed worse than the APP dataset. There were no clear patterns found between the speech styles. Furthermore, the embeddings showed better overall performance than the hand-crafted features. Future work could standardize a dataset for both labelling methods and speech styles to make direct comparison between the methods possible.

Preface

In 2018, I made a rush decision to bridge to Computer Science while being in the middle of my bachelor degree. It was the last year that this transition was allowed and the Computer Science master always intrigued me. It is a decision I cherish to this day. From the first courses of my bridging programme, the world of computing has fascinated me. I remember learning about the Turing Machine–a mathematical model invented in 1936 that was the first model that could implement any algorithm–and being amazed by how these pioneers had been building the field for such a long time. I now realize what triggers this excitement.

As Computer Scientists, we get to solve puzzles. We are on an everlasting quest towards conquering bigger puzzles, solving them more efficiently, often finding ways to use less of our own input. It has been an interesting experience exploring the puzzles that speech and personality have to offer. Combining Computer Science with psychology taught me a lot about both fields. I am curious to see how these fields will develop together in the future.

During the second semester of my master's, I was allowed to go on exchange to Milan. I did not expect the influence it would have on me. Meeting people from all over the world, learning another language, and adapting to the Italian culture broadened my world view and–not to be underestimated–significantly increased my pasta making skills. At the end of the semester, I felt like I was not finished yet. Having had a good experience with my previous rush decision, I set my focus to finding a thesis project in Valencia. This extension gave me a whole set of new experiences for which I am very grateful.

With the completion of my master's degree, there also comes an end to my time abroad. It has been a incredible experience that will stay with me for the rest of my life. Nonetheless, I feel it is the right time to return home. I look forward to spending more time with friends and family, to starting my working life in a new environment, and, undoubtedly, to solving a lot more puzzles.

Emile Lampe
Valencia, June 2023

To my father,
always in my heart

Acknowledgement

This thesis wouldn't have been possible without the contributions of numerous people to whom I am grateful. Firstly, Rianne Smits, the exchange coordinator of our faculty, who was always willing to answer my questions. Willem-Paul Brinkman was incredibly kind by connecting me to one of his colleagues at Universitat Politècnica de València. He helped this project become possible.

I am very grateful to my supervisor, Odette Scharenborg. Odette's expertise, support, and commitment have been of great value. Her enthusiasm and detailed feedback kept me motivated throughout the project. Furthermore, I would like to acknowledge the entire Multimedia Computing research group for the meetings, which were beneficial to my thesis. Additionally, I would like to thank Chirag Raman for dedicating valuable time and energy to my defence.

I would like to thank Mariano Alcañiz Raya, the head of the lab at UPV, who welcomed me to do my thesis at his university. Furthermore, it was a joy to work with the entire team from the start. Specifically, I want to thank my supervisor at UPV, Javier Marin Morales, whose tireless assistance was invaluable. His guidance helped me to stay on track, and I could always rely on him for valuable feedback on my ideas. Furthermore, I would like to thank my fellow speech researcher in our team, Lucía Gómez Zaragozá. Whether it was answering questions, discussing speech-related methodologies, or just having a simple chat, she was always happy to help.

I would also like to thank some friends. My time at TU Delft would not have been the same without my study friends, the Matricen, with whom I have spent hours and hours at the campus. I also want to thank my friend Sebastiaan Scholten, a fellow graduate of Odette, for always being willing to answer my questions and for being a fellow computer geek among my friends.

I would like to say a special thanks to my mother and sister, for being there for me whenever I needed it. They have always encouraged me to follow my heart and to do what I feel needs to be done. Such support is invaluable and will always be remembered.

Finally, I would like to thank my girlfriend, Rali, who has been my buddy in this adventure from the start. You made my time abroad more meaningful than I could have ever hoped for.

Thank you all.

Contents

1 Introduction

1.1 Personality Computing

Personality has been a long-studied subject in psychology [1–4]. The consensus on a personality model among researchers, the Big Five personality traits, is seen as an important achievement of psychological science in the 20th century [5]. These traits have been shown to predict physical and psychological health [6–9], competency in interpersonal relationships [10], and success in key aspects of life such as the work environment [11]. As a result, the field of Human-Computer Interaction started to implement personality, leading to the field of Personality Computing [12]. It combines the insights from Personality Psychology with Computer Science, and focusses on classifying, analysing, and understanding human personality [12]. A digital system tailoring its responses to personality traits could improve domains such as human resources [13], marketing [14], education [15], and healthcare [16]. Additionally, the automatic personality analysis of users can provide valuable insights into human behaviour [17], decision-making processes [18], and social dynamics [19]. Personality classification would therefore be a valuable contribution to human-computer interactions.

Personality has been defined differently across research domains [20]. To address this issue, the following definition was proposed in a recent paper [21]: *"An individual's personality is the enduring set of Traits and Styles that he or she exhibits, which characteristics represent (a) dispositions (i.e., natural tendencies or personal inclinations) of this person, and (b) ways in which this person differs from the 'standard normal person' in his or her society."* The 'enduring' element is often included in definitions [22–24], as personal characteristics are only seen as a trait if it is observed over an extended period of time [21]. The Big Five traits that are used to model personality are [25]:

- **Openness** to experience (curious, aesthetically sensitive, imaginative)
- **Conscientiousness** (organized, productive, responsible)
- **Extraversion** (sociable, assertive, energetic)
- **Agreeableness** (compassionate, respectful, trusting)
- **Neuroticism** (anxious, depressed, volatile)

These traits have been widely adopted as the standard personality framework due to their reliability [26], cross-cultural applicability [27], and real-world validity [11].

With speech-based Personality Computing, personality traits are analysed and classified using the speech signal as the primary data source. One way to do this is

by using the meaning of the words, sentences, and phrases produced by the speaker, called semantic information [12]. The other approach is without this meaning, which is called non-semantic or paralinguistic information [12, 28, 29]. In this thesis, I focus on personality classification using the paralinguistic information of speech data. It consists of all the information that conveys *how* something is being said, instead of *what*.

1.2 Recognition and Perception

Two main approaches have been proposed for automatically classifying personality traits: Automatic Personality Recognition (APR) and Automatic Personality Perception (APP) [12, 28, 30, 31]. They differ in how the personality labels are obtained. APR uses self-assessed personality scores, while APP uses personality scores obtained from external judges that rate each sample based on their perception [12]. The tasks have different aims and are therefore useful for different purposes.

Because APR labels are obtained through a self-assessed personality test, all samples belonging to a certain speaker will have the same labels. This is in line with the definition of personality being stable over time [21]. Some Personality Computing research even states that the aim of APR is to classify the "true" personality of an individual [12, 31, 32], although they acknowledge that self-assessed scores are vulnerable to bias [12]. Personality Psychology research has indicated that it is doubtful there exists a true personality [5], but has also stated that self-assessed personality scores provide important information about how an individual is likely to behave [5]. For these reasons, APR could be especially useful for tasks where the long-term personality of the user is important. In the medical field, it could be used to track the mental health [7, 8] or to present clinical treatment plans through a system that tailors its messages to the preferences and needs of a patient [16]. For consumer use, an example is matching a voice assistant's personality (VAP) to that of the user, as VAP has been shown to make the user feel more at ease and in control [33].

APP labels are based on other people's perception. The labels can vary per sample, which contradicts the definition of traits being observed over longer periods of time [21]. The aim of APP is therefore not to classify the unchanging personality of a speaker, but to predict the personality others will attribute to him [12, 31]. The perceived personality has shown to determine other people's behaviour and opinions towards a person [34]. APP could therefore be especially beneficiary to tasks where the perception of others is important. One study, for example, showed how the personalities of Hillary Clinton and Donald Trump were perceived differently during their 2016 presidential campaigns [35], while another research showed the impact of personality on perceived US presidential success [36]. APP could be used in this case to improve campaign strategies, by using perceived personality to influence public opinion.

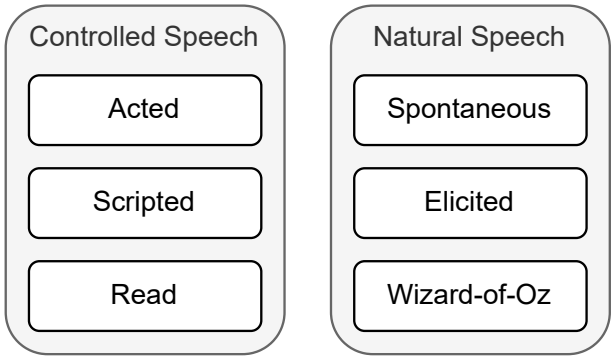


Fig. 1.1: Schematic overview of the defined speech styles: controlled and natural speech. Acted, scripted and read belong to controlled speech, while spontaneous, elicited and Wizard-of-Oz belong to natural speech.

1.3 Controlled and Natural Speech

To be able to train models that classify personality traits based on paralinguistic information, datasets with audio recordings of speakers are needed. The setting in which the audio is recorded can have a big impact on the performance of these models. The personality in speech that is recorded in natural circumstances might not be distinctive enough to successfully classify subtle differences. On the other side, speech that is recorded in a lab might influence the 'naturalness' of the personality and affect its generalizability to real-world scenarios. Existing speech datasets have been recorded in a variety of settings [28, 37–41]. As depicted in Figure 1.1, I define two umbrella terms under which these settings, called speech styles, can be categorized: *controlled* speech and *natural* speech, each with its own sub categories. A dataset can belong to multiple sub categories. Although in this thesis the focus is on personality classification, the controlled and natural speech categories are applicable to other speech-based fields, such as Speech Emotion Recognition (SER).

Controlled speech means that the speaker is given instructions on what to say or how to say it. The terms *acted*, *read*, and *scripted* speech belong to controlled speech. A dataset contains acted speech when the speakers are asked to display, and often magnify, specific characteristics in their speech. It is frequently used in SER as it can help to make emotions more prevalent in the paralinguistic information [42–46]. Scripted speech means that a scenario is defined in advance. This scenario can be determined line by line, or it can be an outline of the conversation [37, 38, 45]. Read speech means the text is literally read out loud.

Under natural speech, speech styles are categorized where speakers have not been instructed on what to say or how to say it. Natural speech includes *spontaneous*, *elicited*, and *Wizard-of-Oz* speech. Spontaneous speech is often used [42–44, 46, 47] and can refer to any situation where a speaker is free to say what he wants and how he wants. With elicited speech, certain behaviour is induced through a stimulus [48]. In the case of SER, this means provoking an emotion in the speaker, for example by asking speakers to describe pictures of happy and tragic events. During a Wizard-

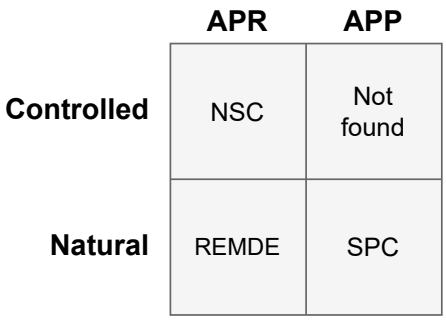


Fig. 1.2: The personality datasets that target the combinations of the labelling methods (APR and APP) and the speech styles (controlled and natural speech) for which a dataset was found. Full names: Nautilus Speaker Characterization corpus (NSC) [38], Reconocimiento EMocional para la evaluación de la DEpresión (REMDE), Speaker Personality Corpus (SPC) [28].

of-Oz conversation, the participants believe they are interacting with a computer, while in reality the response is controlled by a human operator. This enables the researchers to have some control over the conversation while still obtaining natural responses from the participant. Although the speaker's response to the Wizard-of-Oz system could be controlled through a scenario, the reviewed literature used the system to obtain spontaneous responses [42, 43, 49]. Wizard-of-Oz speech is therefore categorized as natural speech.

1.4 Combining personality classification and speech styles

In SER, there has been numerous research on the effects of controlled and natural speech [42–44, 48, 49]. In personality classification, on the other hand, not much research from a speech style perspective has been conducted. This is especially true when also taking the differences between APR and APP into account. Exploring Personality Computing using APR and APP with different speech styles could define future research directions, and identify the field's possibilities and limitations. To be able to do this, a dataset targetting each of the four combinations of labelling methods and speech styles has to be available. For this thesis, I have found two of these databases and contributed to creating a third. They are shown in Figure 1.2. A database for APP and controlled speech could not be found.

The *Nautilus Speaker Characterization* (NSC) corpus [38] is a dataset recorded in German consisting of mainly scripted and semi-spontaneous speech. For the scripted speech, the participants were asked to read the script of a specific situation out loud. The semi-spontaneous speech followed a set of scenario's. Only the scripted subset is taken into consideration. The personality scores are obtained through self-assessment. Because of the scripted speech with self-assessed personality labels, it is seen as controlled speech with APR labels.

The *Speaker Personality Corpus* (SPC) [28] from the Interspeech 2012 Speaker Trait Challenge is a dataset in French obtained from Radio Suisse Romande. The challenge provided a baseline with the dataset and called upon researchers to try to beat it. This has resulted in 11 participants, of which one was declared the winner. The samples were rated by 11 external judges, which makes it an APP dataset. It is categorized as natural speech, as the clips are from a radio channel, recorded in a real-world environment.

The *Reconocimiento EMocional para la evaluación de la DEpresión* (REMDE) is a dataset created during my thesis by a research team at Universitat Politècnica de València (UPV). I helped to conduct the experiments and with decisions on the experimental design, further discussed in Chapter 3. Participants had conversations in Spanish with a life-sized human avatar that could synthesize speech. The avatar's sentences were generated with GPT-3. As the participants were truly free in their replies, this dataset is categorized as natural speech. The participants self-assessed their personality, which makes the dataset suitable for APR.

1.5 Extracting paralinguistic features

To do personality classification based on paralinguistic features, these features must be extracted from the audio. In most Personality Computing research, this has been done using tools such as openSMILE [28, 29, 31], which extracts hand-crafted features from audio [50]. Standardized feature sets exist for this tool, making it easy to extract a large number of features from the data [51]. In recent years, deep learning models have been created, trained on large amounts of data, that can extract embeddings from the audio [52–55]. These embeddings can be used as features to train machine learning models. In 2020, the Non-Semantic Speech Benchmark (NOSS) was introduced for classification tasks based on paralinguistic and non-semantic information [52]. The benchmark includes datasets for SER [56, 57], speaker identification [58], language identification [59], speech commands [60], and dementia detection [61]. This benchmark is often used to test the performance of newly published pretrained models [52–54]. There is no database that targets personality classification included in the benchmark. Because of this, there are many more results for these pretrained models in related fields than for Personality Computing. It would be interesting to see if embeddings and pretrained models are also effective techniques for personality classification. If this is the case, it could open the door for including a personality dataset into the benchmark test.

1.6 Problem statement and research questions

Even though speech style is known to have an impact in related fields such as SER [42–44, 48, 49], research from this perspective lacks in Personality Computing. Like in the SER research, making this separation part of the experimental design could provide better insights. As there are different benefits to APR and APP [5, 13–15, 30],

it is valuable to explore each from a speech style perspective. This exploration can be done with both hand-crafted features and embedding extracted from pretrained models. This would help determine the effectiveness of embeddings for personality classification.

The aim of this thesis is to explore to which extent personality classification can be performed using different labelling methods (APR and APP), speech styles (controlled and natural speech), and feature types (hand-crafted features and embeddings). To achieve this, three types of models, further discussed in Chapter 3, will be trained on datasets that combine APR with controlled and natural speech, and APP with controlled speech. Additionally, both hand-crafted features and embeddings will be used for training the models so that they can be compared. The following research questions are defined to achieve this aim:

RQ 1 *How well can speech-based personality classification be performed using datasets with different labelling methods (APR and APP) and speech styles (controlled and natural speech)?*

It is difficult to draw meaningful conclusions from comparisons between datasets, as there are numerous external factors that can cause differences in results. The research questions are therefore divided into three sub questions that address each of the datasets. The first sub question is about APP and natural speech, for which the SPC is used. As stated in Section 1.4, this database was part of a challenge, which means that results from the baseline and challenge winner already exist. The following sub question is therefore formulated:

SQ 1.1 *For which traits can improvements be made on the challenge baseline and winner using the database for APP and natural speech (SPC)?*

The other two databases have not been used for personality classification before, which means no baseline exists. The sub questions therefore address a baseline that is defined as the chance level for each trait:

SQ 1.2 *Which traits can be classified better than chance using the database for APR and controlled speech (NSC)?*

SQ 1.3 *Which traits can be classified better than chance using the database for APR and natural speech (REMDE)?*

The second research question addresses the embeddings and hand-crafted features. It is possible to make direct comparisons here, as the models can be repeatedly trained with the same setup using both the embeddings and the hand-crafted features.

RQ 2 *How do the models perform on embeddings compared to hand-crafted features for speech-based personality classification?*

In the remainder of this thesis, Chapter 2 will discuss related work in the field of Personality Computing. Chapter 3 will discuss the methodology for the experiments. In Chapter 4, the results of the experiments will be presented. The results and limitations will be discussed in Chapter 5. Suggestions for future work are shared in 6. Finally, the research questions will be answered and concluded in Chapter 7.

Related Work

2

2.1 Research on APP and APR

The introduction of the Personality Sub Challenge of the Interspeech 2012 Speaker Trait Challenge [28] sparked a rise in APP research [62–72]. The sub challenge provided researchers with the Speaker Personality Corpus, together with an extensive feature set of 6,125 features, and a baseline set by Support Vector Machine (SVM) and Random Forest (RF) models. Each personality trait could be classified as 'Low' and 'High', therefore making it a multi-label binary classification task, where each trait is a label. In the subsequent survey of the challenge [31], the results of the challenge contributors were presented. Figure 2.1 shows that only 2 out of the 11 contributions marginally managed to surpass the baseline. Out of the 5 traits, Extraversion and Conscientiousness consistently scored highest among the participants [30], as shown in Figure 2.2. A larger circle in this graph represents a higher Unweighted Average Recall (UAR), which is the main performance metric used in the challenge. It is interesting that relative performance on the traits is so consistent among the contributors. This could have multiple reasons. It could be that the traits that scored high are more expressed externally, when compared to the other traits [5, 30]. This would make them more perceivable. Another explanation is that by chance the test set contained samples with .

The best result among the participants was achieved by incorporating numerous modulation spectrum analysis features into the existing base feature set [62]. After feature selection, the feature count in this research ranged from 6,719 for Openness to 13,425 for Extraversion. The AdaBoost [73] ensemble method was used for classification.

Since the challenge, new attempts have been made at performing APP on the SPC dataset. One study did this by training SVMs on spectrogram images [74]. Their model surpassed the baseline of the Personality sub-challenge for Agreeableness and Neuroticism traits, achieving UAR scores of respectively 64.9% and 70.8%. In another study, an asymmetric auto-encoder was used, where in each hidden layer the parameters were trained in a semi-supervised manner [75]. The paper's most notable claim is a UAR of 81.2% for Openness, making it their best classifiable trait. This is contrary to most other work making use of the SPC, where Openness often scored the lowest [28, 31]. One study used the SPC to combine paralinguistic features with linguistic and psycholinguistic features, using an SVM with a linear kernel for their models [76]. The linguistic features were obtained through part-of-speech tagging. The psycholinguistic features were obtained with the Linguistic Inquiry Word Count (LIWC) system that gives scores to words for 81 psycholinguistic categories such as 'anger'[76]. The combined feature set resulted in an average UAR

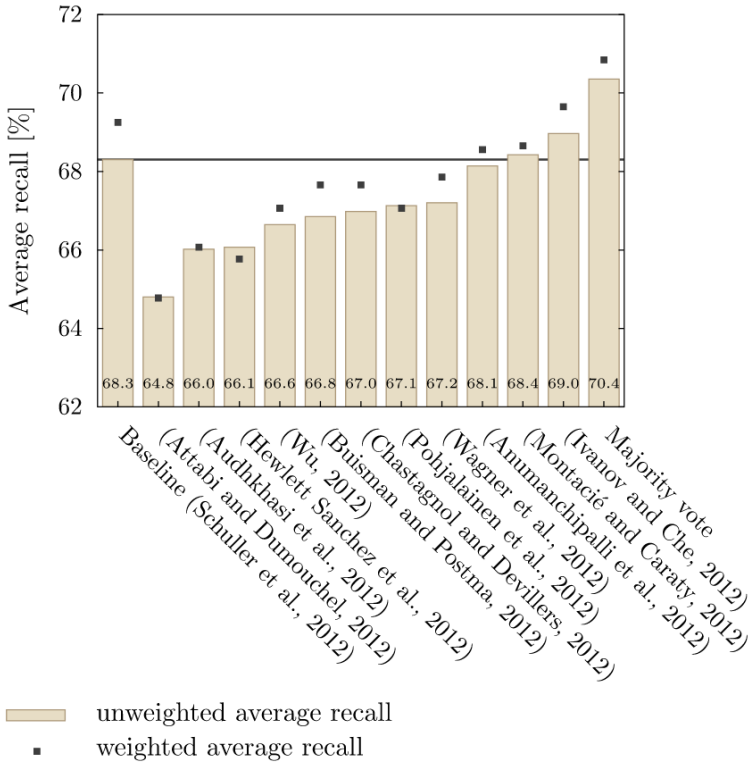


Fig. 2.1: The results from the (APP) Personality sub-challenge, taken from the INTERSPEECH 2012 Speaker Trait Challenge survey [31].

of 70.4% across the traits, while only using paralinguistic features resulted in an average UAR of 69.4%. The psycholinguistic features achieved a UAR of 58.1% and the part-of-speech tagging 51.0%. The paralinguistic features were therefore the most important contributor to the combined score.

In contrast to APP, research on speech-based APR has been relatively limited. This could be because APP is seen as harder than APP [29]. Additionally, there has not been a well-known challenge such as the Speaker Trait Challenge to attract attention to APR. One of the earlier works on APR made use of transcriptions from spoken dialogues to classify personality traits with an SVM [77]. Their findings demonstrated that linguistic features, such as word usage and sentence structure could be used to classify self-reported personality traits. Another study attempted to predict self-assessed Extraversion and Neuroticism from video-based distance between individuals in a social group interaction, using an SVM with a second-order polynomial kernel [78]. Among APR research with the highest results was a paper where Convolutional Neural Networks (CNN) were used to train on video footage of facial expressions. All traits were predicted with an accuracy higher than 90%, an average accuracy of 95.3%, and all found correlations having a p-value $<$ 0.01 [79]. Another study combined paralinguistic and visual features to perform APR [80]. Their dataset consisted of Skype calls where individuals introduced themselves in front of a camera. The highest scores were achieved on Extraversion

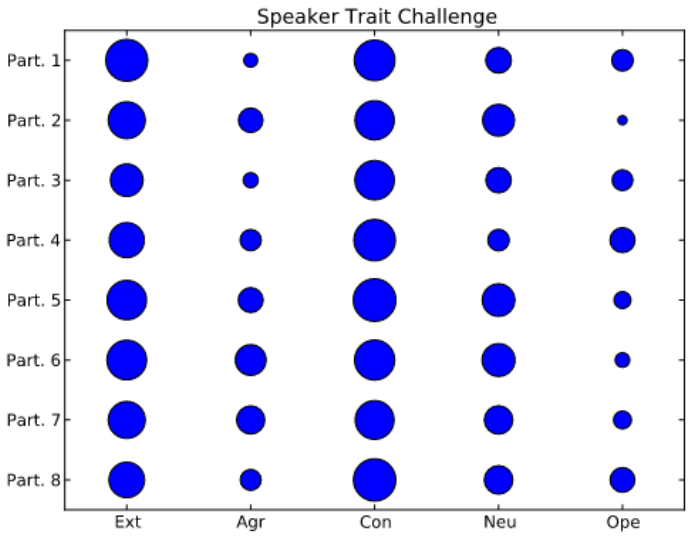


Fig. 2.2: The performance of the best 8 participants of the Personality sub-challenge per personality trait, taken from the paper 'More Personality in Personality Computing" [30]. The size of each element in the plot represents the performance that was achieved by each participant on that trait. This makes it easy to compare the relative performance of the traits across the participants, showing that Extraversion and Conscientiousness structurally were better classified than the other traits. Abbreviations: Participant (Part.), Extraversion (Ext), Agreeableness (Agr), Conscientiousness (Con), Neuroticism (Neu), Openness (Ope).

and Conscientiousness. Although the accuracy scores cannot be directly compared to the results from the Personality Sub Challenge of the Speaker Trait Challenge [28], it is interesting to note that Extraversion and Conscientiousness also there scored consistently the highest.

One study, among the first to perform APR with only paralinguistic features, used the Personable and Intelligent virtual Agents (PersIA) corpus [81]. The dataset was collected during a study on the effect of personality on user experience in a tourist call center [81]. The boostexter classifier was used for their experiments [82]. Their results include a 95.0% accuracy for Conscientiousness and a 63.0% accuracy for Extraversion, compared to respectively the 73.2% and 50.0% chance levels obtained from random drawings. In another APR research where only paralinguistic features were used, a study trained an SVM on two corpora consisting of American English and Mandarin Chinese [29]. Unlike most personality classification research that categorizes labels into 'High' and 'Low', this study introduced a third 'Medium' class. Four types of feature sets were used, both individually and in various combinations. All the traits achieved a UAR higher than the 33% chance baseline in at least one of the feature sets, although measures of significance or reliability were not provided.

2.2 Research on Controlled and Natural Speech

In Personality Computing, no work has been published yet that focusses specifically on the effects of natural and controlled speech. As stated in Chapter 1, such research

does exist for Speech Emotion Recognition. In one paper, acted speech was compared with Wizard-of-Oz speech [42]. The acted speech database was recorded for their research at TU Berlin, while the SmartKom corpus was used for the Wizard-of-Oz speech. They found that acted emotions were more easily recognized and that for acted speech the impact of feature selection was higher. Another paper compared the acted, read and Wizard-of-Oz speech [43]. The data was collected specifically for their research, and they made use of prosodic features. The best results were achieved on acted speech with a 91.5% recognition of segments containing emotions, followed by read speech with 71.6%, and finally Wizard-of-Oz speech with 70.6%. A third paper compared acted speech, from the Berlin emotional speech database, with spontaneous speech from the FAU Aibo corpus [44]. This spontaneous speech dataset was obtained from children interacting naturally with an AIBO robot. The spontaneous speech was found to be more challenging in this research. An SVM trained on the Berlin dataset achieved 75.8% UAR, while the SVM trained on the FAU Aibo corpus achieved 41.5% UAR. The FAU AIbo was also used in another paper where it was compared with acted speech from the Spanish Emotional Speech (SES) database, again finding the acted speech to achieve better results [48]. Classifying activation emotions, where happy and anger are categorized as active, while bored and sad are categorized as passive, resulted in 85% accuracy for the acted SES database and 79% for the natural AIBO database.

The research on controlled and natural speech in SER shows that models trained to classify emotions in controlled, and specifically acted, speech achieve better results than those trained on speech recorded in natural circumstances. It is interesting to see if such a clear pattern is similarly visible in personality recognition. In theory, a difference could be expected between APP and APR. APP is closer in methodology to SER, as the short segments in emotion databases are also labelled based on how they are perceived. This could mean that acted speech also makes it easier to classify perceived perception. The contrary could be true for APR. Acting could cover up the speaker's personality to show in the speech signal. As the labels are obtained from self-assessment, this could make classification harder. An important difference between SER and personality classification, is that in personality classification requires 5 traits to be classified. This means there are 32 unique combinations of personality traits. Acted speech in SER can be used to specifically amplify an emotion such as 'anger' or 'happiness', while it is not possible to do this with one of those 32 combinations.

2.3 Personality Tests

Over the years, numerous personality tests have been developed to assess personality. The most important difference between the tests is the depth to which the personality is assessed. These tests consist of items, which are statements or questions, to which

I see myself as someone who...	
is reserved	is outgoing, sociable
is generally trusting	tends to find fault with others
tends to be lazy	does a thorough job
is relaxed, handles stress well	gets nervous easily
has few artistic interests	has an active imagination

Tab. 2.1: The Big Five Inventory-10 (BFI-10). All statements are to be answered on a 5-point Likert scale from “strongly disagree” to “strongly agree”.

the user can answer on a scale from ‘strongly disagree’ to ‘strongly agree’. The following is an overview of the most widely used personality tests:

Big Five Inventory-44 (BFI-44)

The BFI-44 is a test that evaluates the Big Five traits. It includes 44 statements that start with “I see myself as someone who...”, and are followed by descriptions such as “is curious about many different things” or “can be moody”. Each statement has to be answered using a 5-point Likert scale, ranging from "disagree strongly" to "agree strongly". The BFI-44 provides a balance of comprehensive insights into personality traits without being overly time-consuming to complete. Because of this, it has been used multiple times in personality classification research [83–85].

Big Five Inventory-10 (BFI-10)

The BFI-10 [32] was developed as a shorter variant of the BFI, containing only 10 statements, shown in Table 2.1. The BFI-10 provides a rapid and efficient means to assess the Big Five personality traits, particularly in situations with limited time and resources. Despite its brief nature, the BFI-10 has demonstrated significant levels of reliability and validity [32]. It is a popular choice in APP [28, 86], as the task requires each sample to be rated by multiple judges, which would be too time and resource consuming with longer tests.

NEO Personality Inventory-Revised (NEO PI-R)

The NEO PI-R [87] is an extensive self-report questionnaire developed to measure the Big Five personality traits and their six facets. Composed of 240 items, the NEO PI-R offers a comprehensive representation of an individual’s personality. While being very rigorous, it is found less often in related literature. This is thought to be due to its length.

NEO Five-Factor Inventory (NEO-FFI)

The NEO Five-Factor Inventory (NEO-FFI) [87] is a condensed version of the NEO PI-R, specifically designed to evaluate the Big Five personality traits without delving into the facets that the NEO PI-R explores. The NEO-FFI comprises 60 items and

is advantageous in situations where a shorter assessment is desirable, while still maintaining a strong focus on the primary personality dimensions. The NEO-FFI has also been used in personality classification research [29].

Methodology

3

The research question formulated in Chapter 1 is the following: *Can personality classification be performed on each of the labelling method and speech style combinations?* To answer this question, an experiment will be conducted for each of the datasets belonging to the combinations. The code for this methodology can be found on GitHub [1].

The first step will be preprocessing the audio samples so that their format is suitable for the experiments (Section 3.1). In the next step, features will be extracted from the data. This will be done using both classical hand-crafted features and using a pretrained deep learning model (Section 3.2). After this, the dataset is divided into a training and test set in a way that is suitable to our task (Section 3.3). The next step is to process the feature with standardization and feature selection (Section 3.4). Then, three selected models are trained on the features. The models and their hyperparameters will be discussed in Section 3.5. Finally, the results of the models will be evaluated using the chosen evaluation metrics. The choice for these metrics is explained in Section 3.6.

Each personality trait will be seen as its own binary classification task. This means that there will be 5 models, one for each trait. The personality scores are categorized into scoring either 'Low' or 'High' on a trait. The samples will be divided into a training and test set. After this division, each trait will have a separate pipeline, with its own selected features and model hyperparameters. Furthermore, the two types of feature sets will also be trained on separately.

3.1 Datasets

3.1.1 Speaker Personality Corpus

The Speaker Personality Corpus (SPC) from the Interspeech 2012 Speaker Trait Challenge [28] was used for the combination of APP and natural speech. The SPC consists of audio recordings from the French-speaking Swiss radio channel Radio Suisse Romande. It includes 322 individuals, speaking for approximately 10 seconds per sample, for a total of roughly 1 hour and 40 minutes. The samples have a frequency range of 8kHz with a 32-bit depth and 1-channel. There are 264 male speakers in the dataset and 59 women. No further processing on the audio samples was performed so that the results could be compared to earlier obtained results on the challenge from related work.

As this dataset is intended for APP, external judges rated the labels. For this dataset, 11 external judges rated each sample using the BFI-10 personality test. The judges did not speak French, ensuring that the semantics would not influence their

[1] https://github.com/emilelampe/speech-automatic-personality-recognition

Trait	Label	Calc.	IS12	Diff.
Openness	O	222	247	25
	NO	418	393	25
Conscientious	C	413	290	123
	NC	227	350	123
Extraversion	E	337	320	17
	NE	303	320	17
Agreeableness	A	370	323	47
	NA	270	317	47
Neuroticism	N	306	318	12
	NN	334	322	12

Tab. 3.1: The distribution of the binary labels calculated following the Challenge instructions compared to the distribution as stated in their paper. Calc. is the number of samples calculated in this thesis for each class. IS12 is the number of samples calculated in the challenge paper. Diff. is the difference between the two.

perception. From these 11 different tables of personality scores, the binary labels were calculated following the instructions from the original Speaker Trait Challenge: "Each clip is labelled to be above average (X) for a given trait X $\in$ O, C, E, A, N if at least six judges (the majority) assign to it a score higher than their average for the same trait; otherwise, it is labelled NX" [28], where NX means scoring below average on that trait.

Following these instructions, however, resulted in a label distribution different from the one reported in the Challenge paper. Despite various interpretations of the instructions, contacting the creator of the SPC, and contacting contributors to the challenge, the original label set could not be obtained. As a result, the labels created following the exact instructions in the challenge paper were used. Table 3.1 illustrates the differences in distribution between the labels as obtained when using the instructions and those stated in the Challenge paper. Conscientiousness, in particular, has a significantly different label distribution, with the majority and minority classes inverted.

3.1.2 REMDE

The REMDE dataset is used to explore APR with natural speech. As discussed in Chapter 1, this dataset is currently being developed at the Universitat Politècnica de València (UPV), where I conducted my research for this thesis. The team's objective is to detect depression symptoms in participants using data recorded through conversations between participants and a digital avatar. I helped to develop the dataset as part of this thesis. Because I am the first of the team to work with the audio data and because the audio was unprocessed, the REMDE dataset required extensive preprocessing. This section will therefore discuss in greater detail what preprocessing steps were taken.

Context of the database

The dataset consists of speech belonging to participants that have had spontaneous conversations with a digital avatar. The avatar is a life-sized representation of either a male or female character. It communicates using speech synthesis with text generated by GPT-3. This connection to GPT-3 allows the character's response, and therefore the conversation, to be spontaneous. The avatar can exhibit various emotional states, such as happiness, sadness, anger, and neutrality, and adjusts its choice of words and body language based on the selected emotion. It can also be dressed formal and casual. Each participant engaged in six conversations with the avatar, alternating between the male and female characters. The emotional states and ways of dressing were distributed equally among the participants. Each participant encountered all emotional states at least once in their six conversations. Figure 3.1 shows the interaction between a participant and the avatar. The avatar in the figure is female, has a neutral emotional state, and is dressed formally.

Throughout the interactions, multi-modal data was recorded from the participant, including speech audio, eye-tracking, electroencephalogram (EEG, electrical activity of the brain), electrocardiogram (ECG, electrical activity of the heart), and electrodermal activity (EDA, electrical conductance of the skin). The study included 54 healthy participants and 50 clinically depressed individuals. Only the data from healthy participants was used in this research. This was done to eliminate the potential influence of depression on speech patterns and personality expression. The dataset is recorded in Spanish, with an average participant age of 31.9 years, ranging from 18 to 54 years old. There were 28 male and 26 female participants.

Fig. 3.1: A person wearing the recording devices, who is interacting with the neutral-state, formally dressed female avatar.

Recording setup

Speech audio was recorded using the SYNCO G1 (A1) wireless system at 44.1kHz/32 bit and 1-channel. The audio was recorded in a controlled recording environment.

Lucía Gómez Zaragozá–the researcher responsible for speech analysis in the UPV depression project–and I conducted the experiments. We determined that the optimal microphone location was the forehead. This decision was based on a paper [88] that demonstrated that this location was the least obstructive to the frequency spectrum, when compared to a microphone that was placed at mouth height 1 meter in front of the participant. Figure 3.2 shows the frequency spectra for different microphone positions [88]. It can be seen that the other positions result in a relatively big drop in the high-frequency range. To verify that the forehead is the best location, we recorded multiple samples with the different microphone positions. During playback, the samples with the microphone placed on the forehead had the clearest sound. This solidified our decision to use the forehead as the microphone position. The initial number of samples was 3,730.

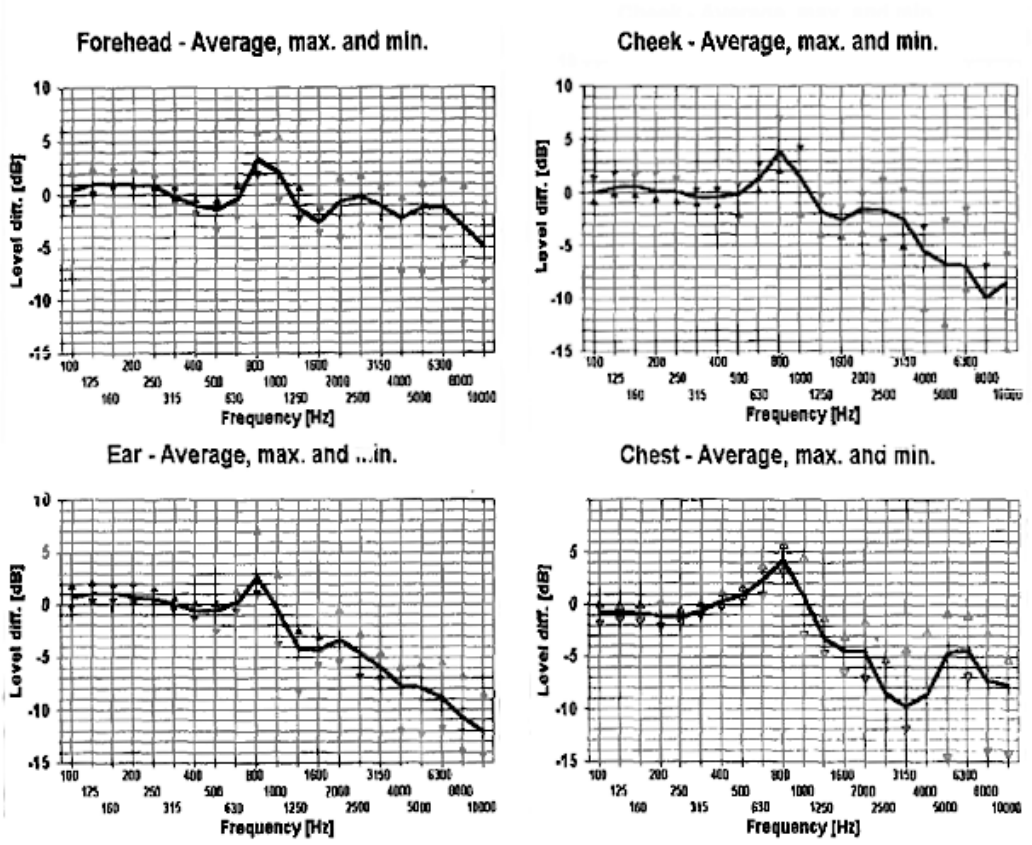


Fig. 3.2: The frequency spectrum with different microphone locations. The paper [88] reported that the forehead resulted in the least amount of spectral change.

Detecting samples with a high level of noise

Upon inspection of the recorded audio, some conversations turned out to contain a constant noise signal, which is thought to be due to microphone problems. The loudness of the noise varied per conversation. The noise in some conversations was so loud that the samples were marked as unusable. Because the high number of samples made it not feasible to check every sample manually, I developed an

algorithm to filter out samples with a high level of noise using the Root Mean Square (RMS) energy. The RMS is the average power output of a signal. When there is constant noise, the average power output will be higher. The algorithm calculates the RMS of each sample and filters out the samples that have an RMS above a certain threshold. After some experimentation with samples known to contain both acceptable and unacceptable levels of noise, the RMS threshold was set to be 0.35. All samples with an RMS above 0.35 were removed from the dataset. This brought the initial number of samples down from 3,730 to 3,665. A selection of the removed samples were examined and all of them were confirmed to have a high level of noise.

Reduction of remaining noise

In the next step of the audio preprocessing, non-stationary noise reduction was explored to remove any noise of lower but still notable levels. The non-stationary noise reduction algorithm, from the PyPi package *noisereduce* [89], creates a time and frequency smoothed mask from the spectrogram of an audio signal. This mask selectively reduces noise that persists on longer timescales than the target signal. This means that noise is removed that is present for longer durations than the fluctuating speech signal. Although it successfully removed noise and performed better than other noise reduction techniques, it also affected the quality of the speech signal. Especially the lower ends were sometimes filtered out. An exploration was conducted with models trained on features from audio with and without noise reduction. The results from the noisy samples showed higher standard deviations in the cross validation when compared to the noise-reduced samples. This instability is suspected to be caused by the noise. Therefore, the samples with noise reduction were chosen for the experiments.

Normalization, segmentation, and concatenation

Next, all the samples were normalized using peak normalization, so that the effect of external influences, such as the exact placement of the microphone, were minimized. The audio files often contained long silent beginnings or endings. Because of this, the start and end of the samples were cut off if there was more than 1200ms of audio with an amplitude lower than -30dB. The silences between words within a response were kept, as these could be indicators of personality traits. The samples from each conversation were then concatenated with an algorithm that optimally creates combinations of samples between 10 and 15 seconds. This duration was based on the 10-second duration used in the Speaker Personality Corpus. A window of 5 seconds was used so that complete utterances could be concatenated, instead of samples being cut off in the middle of a word. To retain a natural speech style, only samples from the same conversation were combined. They were also concatenated in sequential order. This process resulted in a total of 1,249 samples of an average duration of 13.3 seconds. The total duration of the samples is 4.6 hours. There are 23 clips on average per speaker.

Labelling the samples

To obtain the personality scores of the participants, the BFI-44 self-assessment questionnaire was chosen as it combines a thorough analysis while being relatively short. The BFI-44 outputs personality scores on a continuous scale. To make it suitable for a classification task, the personality scores were divided into the categories 'Low' and 'High' for each personality trait, consistent with most related work in this field [28, 29, 31, 62, 65]. Ideally, this division occurs in the middle between the highest and lowest possible score, as this score represents the neutral point where a person scores neither high nor low on a trait. This way, both classes have an equal range in which a personality score could fall into its category. However, the raw personality scores did not distribute evenly around this neutral point, resulting in highly imbalanced binary classes. Therefore, the median was used as the binary threshold. Figure 3.3 shows the distribution of the personality scores. In the figure it can be seen that if 0, the neutral midpoint, would have been used as the threshold to divide the scores into two classes, the label distribution would have been very imbalanced. The median, which is shown in the figure as the point where the colours change, provides a balanced distribution when used as the binary threshold and was therefore used.

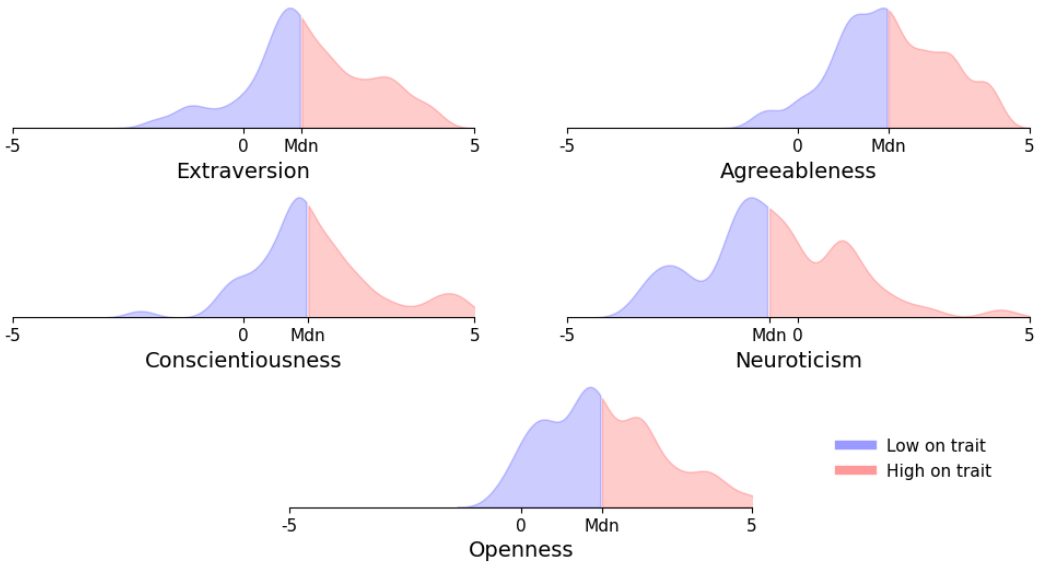


Fig. 3.3: The distribution of the personality scores of the participants in the REMDE dataset. The binary threshold is set at the median. The colours represent the two different classes. It can be seen that if 0, the neutral midpoint, would have been used as the binary threshold, the label distribution would have bee very imbalanced.

3.1.3 Nautilus Speaker Characterization Corpus

The Nautilus Speaker Characterization (NSC) corpus [38, 90] is used to explore controlled speech APR. It comprises scripted telephone conversations in German. The conversations were recorded under controlled conditions to ensure high audio quality. The samples were recorded at 48kHz/16 bit and 1-channel. I downsampled

the audio to 41kHz so that it matches the frequency range of the REMDE dataset. The age of the participants ranges between 18 and 35 years old. 117 of the speakers are male, and 156 speakers are female. Personality trait scores were obtained using the German version of the BFI-44, which includes a 45th German question, and the BFI-10, depending on the recording session. A total of 181 participants completed the BFI-44, while 92 speakers completed the BFI-10.

The documentation states the database contains no noise [38]. The algorithm that was used on the REMDE dataset to detect samples with high levels of noise was used and, as expected, no samples were detected above the threshold. The audio samples were then normalized, segmented, and concatenated in the same way as the REMDE dataset, resulting in samples of a duration between 10 and 15 seconds. This resulted in 3354 samples with a total of 12.0 hours of speech.

Figure 3.4 shows that, just as with the REMDE dataset, using the neutral midpoint as the binary threshold would have resulted in an imbalanced label distribution. Therefore, the median was used as the threshold to divide the scores into the binary classes 'Low' and 'High'.

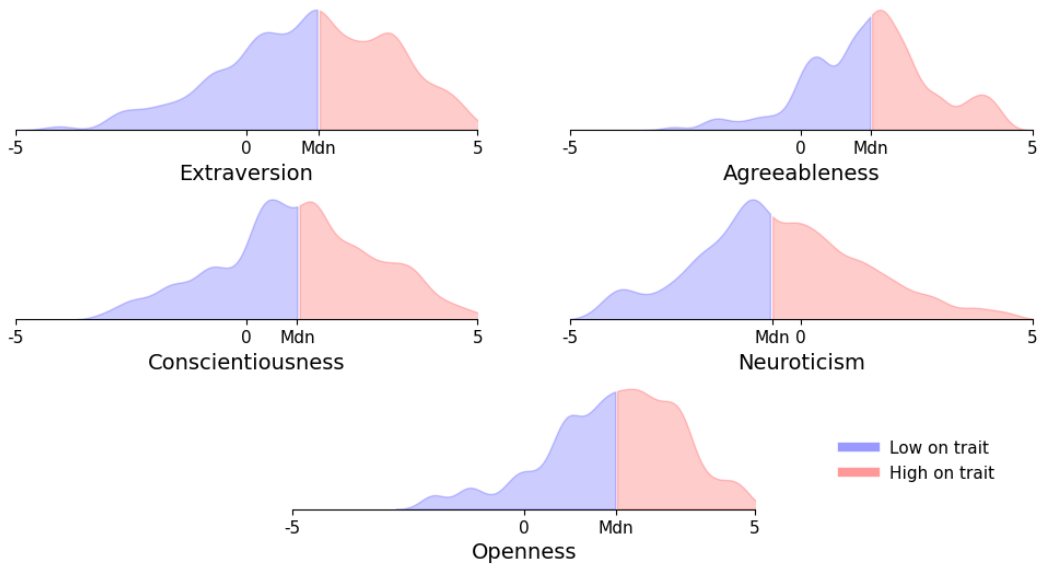


Fig. 3.4: The distribution of the personality scores of the participants in the NSC corpus. The binary threshold is set at the median. The colours represent the two different classes. It can be seen that if 0, the neutral midpoint, would have been used as the binary threshold, the label distribution would have been very imbalanced.

3.2 Feature Extraction

Feature extraction was done in two ways: through the classical approach using hand-crafted features [51], and using a pretrained deep learning model [53]. For the classically extracted features, the tool openSMILE [50] was used. This tool provides multiple standard feature sets. Two of these were investigated: the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) [51], comprising 88 features, and

Model	VoxCeleb [58]	VoxForge [59]	SpeechCommands [60]	CREMA-D [56]
openSMILE (ComParE) [91]	2.5	78.0	36.5	53.7
TRILL [52]	13.8	84.5	77.6	65.7
Wav2Vec2 [55]	17.9	98.5	95.0	77.4
CAP12 [93]	**51.0**	**99.7**	**97.1**	**88.2**
TRILLsson [53]	46.2	**99.7**	93.9	86.1

Tab. 3.2: A comparison of the different pretrained models for datasets in the NOSS benchmark. A higher score means better performance. Only datasets for the benchmark are included that were used across the publications of these models. CAP12 achieves the highest score on all the datasets. TRILLsson, a distilation from CAP12, achieves comparable results while being a fraction of the size.

the Computational Paralinguistics Challenge (ComParE) [91], comprising 6,373 features. Upon exploration, eGeMAPS was found to outperform ComParE in almost all cases while greatly reducing training time. Consequently, the eGeMAPS feature set was used for the experiments. eGeMAPS includes a set of low-level acoustic descriptors (LLDs), such as fundamental frequency (F0), loudness, spectral slope, spectral flux, Mel-frequency cepstral coefficients (MFCCs), and jitter and shimmer measures. It is widely used in related fields such as SER [91, 92].

For the deep learning model, 4 pretrained models were compared. These models are TRILL [52], FRILL [54], Wav2Vec2 [55], and TRILLsson [53]. The best model was selected based on datasets in the NOSS benchmark, discussed in Chapter 1, that were used in all publications of the models. The results for these models on the datasets are presented in Table 3.2. It shows that CAP12 achieved the highest score on all datasets. CAP12 is a massive model trained on the YT-U dataset, containing 900M+ hours of footage from YouTube [93]. It is not publicly available. The creators distilled a smaller, publicly available model from CAP12, called TRILLsson [53]. Distillation means that knowledge got transferred from a larger model into a smaller one. TRILLsson was built as a faster, smaller, and mobile-friendly distillation. The model on average performs 90-96% as well as CAP12, despite being 1%-15% the size and trained using only 6% the data. Because of these competitive results using a small model, this model was selected as the embedding extractor for this thesis. TRILLsson outputs 1024 embeddings for each sample. These embeddings are then used as features for the classification models described in Section 3.5.

3.3 Training and Test Splits of the Databases

Given that there are five personality traits to be classified from the data, this task can be considered a multi-label problem. This means that each sample has multiple non-exclusive labels. Standard algorithms for dividing data into training and test sets fail to maintain pairwise balance between labels, which leads to bad generalization [94]. Therefore, dedicated multi-label algorithms are needed to split the data. Additionally, the tasks in this thesis require all samples belonging to a speaker not to be shared between the training and test set. Such a division could cause the model to train on

the characteristics of the speaker rather than the personality traits. Although tools exist to split data into multi-label stratified training and test sets, they do not ensure speaker-independency between training and test sets. Similarly, there exists a tool to create training and test sets that ensures speaker-independece, but this only works for single-label tasks.

To that end, I developed a PyPi package, named *maestros*[2] (MULTI-LABEL STRATIFIED GROUP SPLITS). It makes use of multistrat for splitting multi-label data into train and test sets, but also guarantees speaker-independence. In addition to splitting data according to these requirements, the package also generates stratification reports and charts, allowing for easy examination of the stratification quality of the data sets.

Figure 3.5 to 3.7 display the charts created by *maestros* after performing a multi-label stratified speaker-independent split. Figure 3.5 shows that there were many more men than women in the dataset, which could impact the results. The traits show slight imbalances, but no extreme imbalance. Between the complete, training, and test set, each trait and gender is well stratified. Figure 3.4 shows that the NSC corpus is almost perfectly balanced and stratified. This is the result of using the median as the binary threshold. The traits for the REMDE dataset in Figure 3.7 show that the training and test sets are slightly less well-balanced and stratified compared to the NSC corpus. This is because the dataset has fewer speakers, making it harder to create a perfect stratification for all traits. Specifically, the balance of Neuroticism in the test set diverges from the complete and training set.

Besides the personality traits, gender was also included in the stratification process of all the datasets to ensure that the male-to-female ratio in the training and test sets is similar to that of the complete data set. This is important as gender is known to influence the paralinguistic information of speech.

3.4 Feature preprocessing

After the data was split into a development and a test set, a different pipeline for each trait was created. This resulted in 5 different single-label binary classification tasks. The first step in the pipeline was standardization. After this, Principal Component Analysis (PCA) was explored. PCA is a technique to reduce data complexity by transforming many related features into fewer, uncorrelated components, while retaining important information. PCA was tried by creating components that retained both 95% and 99% variability. However, the results of the exploration indicated that excluding PCA performed better. This step was therefore not included in the final experiments. Then, feature selection using cross-validated Recursive Feature Elimination (RFE) was explored. This algorithm uses cross-validation to find the optimal number of features to be included. It was done with 5-fold speaker-independent cross validation. Just as with PCA< this step showed to be detrimental

[2] https://github.com/emilelampe/maestros

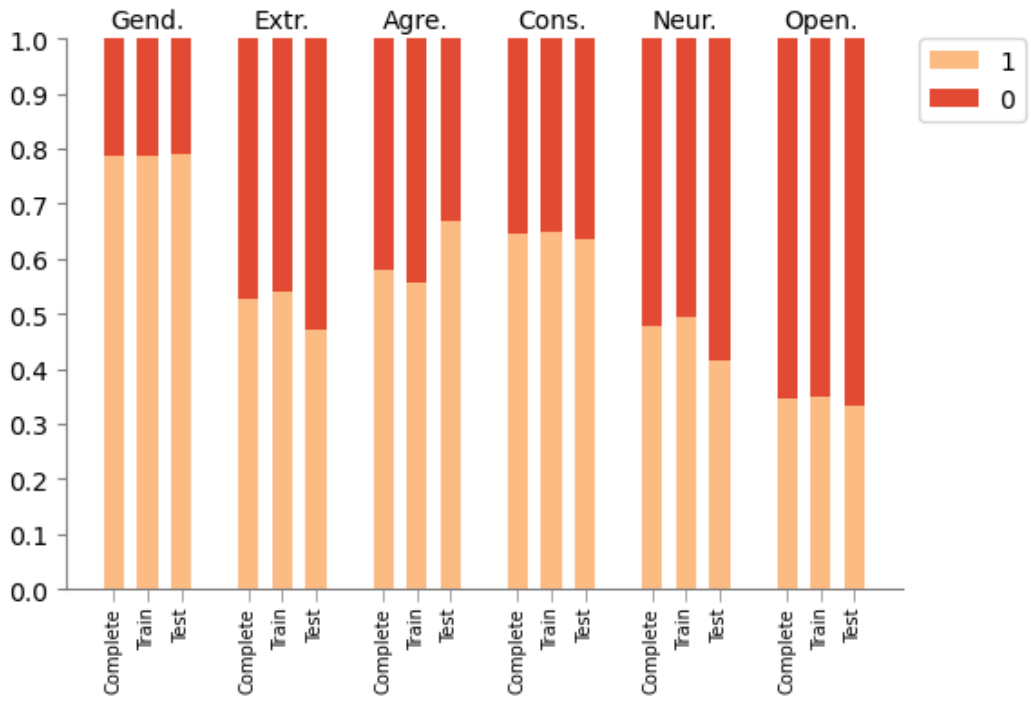


Fig. 3.5: A chart of the SPC dataset stratification. The chart, generated by *maestros*, illustrates the stratification of the training and test set compared to the whole dataset. Gender is included in the stratification.

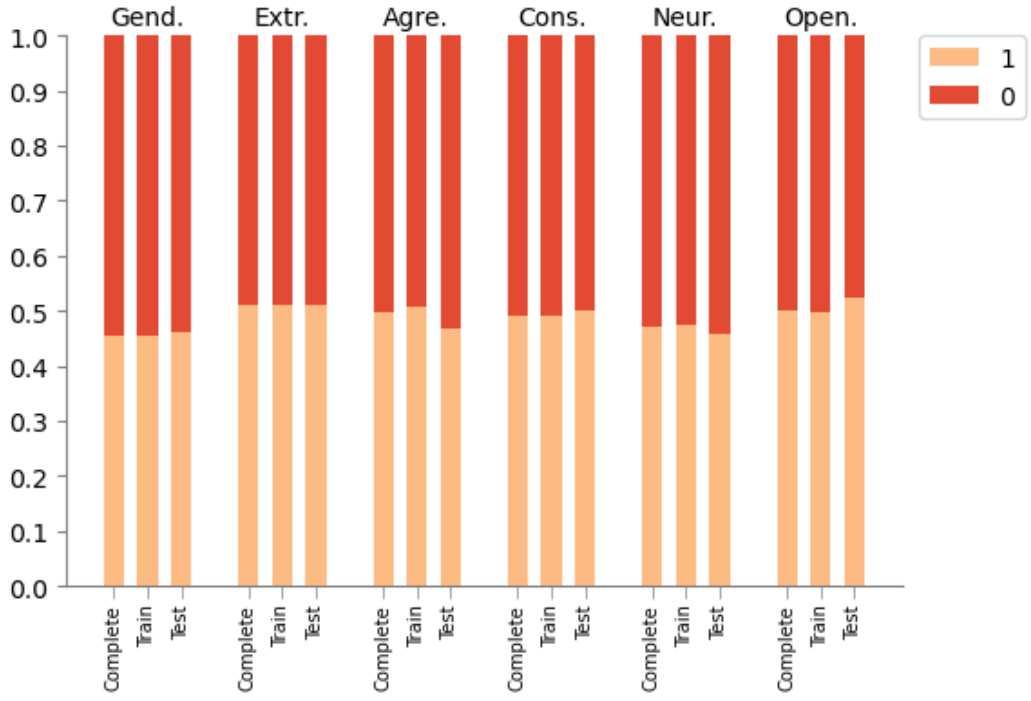


Fig. 3.6: A chart of the NSC corpus stratification. The chart, generated by *maestros*, illustrates the stratification of the training and test set compared to the whole dataset. Gender is included in the stratification.

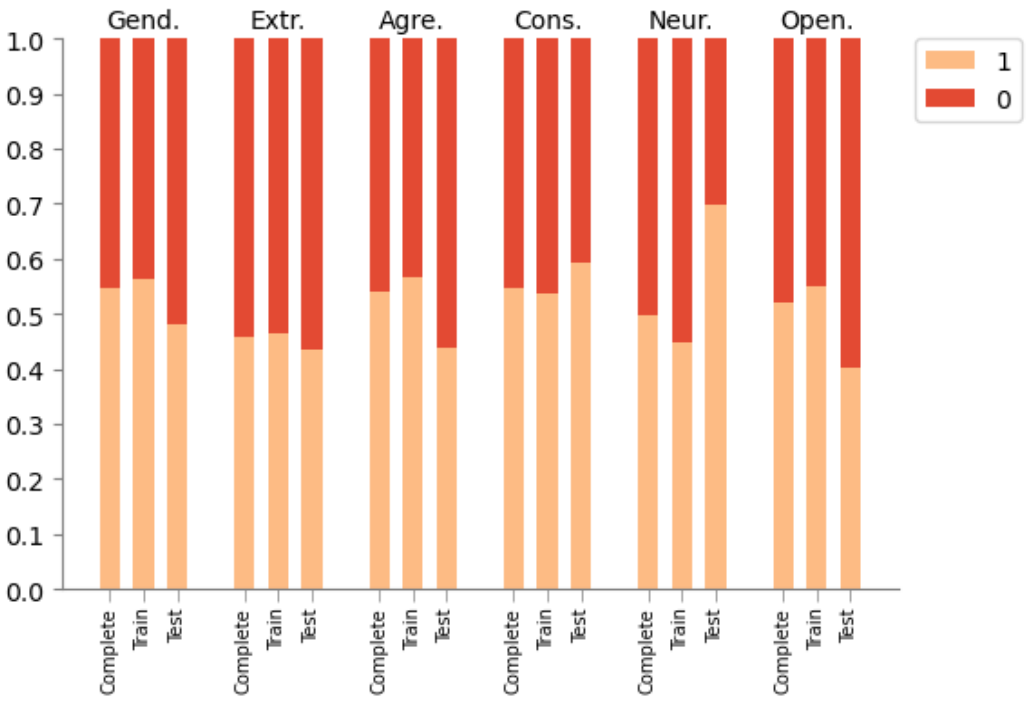


Fig. 3.7: A chart of the REMDE corpus stratification. The chart, generated by *maestros*, illustrates the stratification of the training and test set compared to the whole dataset. Gender is included in the stratification.

to results in the cross validation. Therefore, this step was also excluded from the final experiments.

3.5 Classifiers

This section will discuss the three models that were chosen for the experiments: Support Vector Machines (SVM) with an RBF kernel, Random Forest (RF) classifiers, and k-Neirest Neighbors (kNN). An explanation of the models is given, together with research where the models were used. Additionally, the chosen range of hyperparameters that were explored for the cross validation development phase are presented.

3.5.1 Support Vector Machine

Given a set of training data, the goal of an SVM is to find the optimal hyperplane that best separates the data points into their respective classes. The optimal hyperplane is the one that maximizes the margin between the two classes, defined as the distance between the hyperplane and the closest data points from each class, known as support vectors. SVMs have been a popular choice in earlier work on personality classification. It was, for example, one of the models for the baseline in the Personality Sub Challenge of the Interspeech 2012 Speaker Trait Challenge [28].

SVMs work with kernels, which are functions that compute the similarity between data points in a transformed space. The linear kernel uses a linear function to

separate data points in a feature space. The polynomial kernel is a non-linear kernel that allows SVM to learn more complex decision boundaries. Figure 3.8 shows how an SVM with a polynomial kernel manages to non-linearly separate data points. For the experiments, an SVM with a Radial Basis Function (RBF) kernel is used. The RBF kernel is a Gaussian function that computes the similarity between input vectors. It maps the input data into an infinite-dimensional space, where the data points can become linearly separable [95].

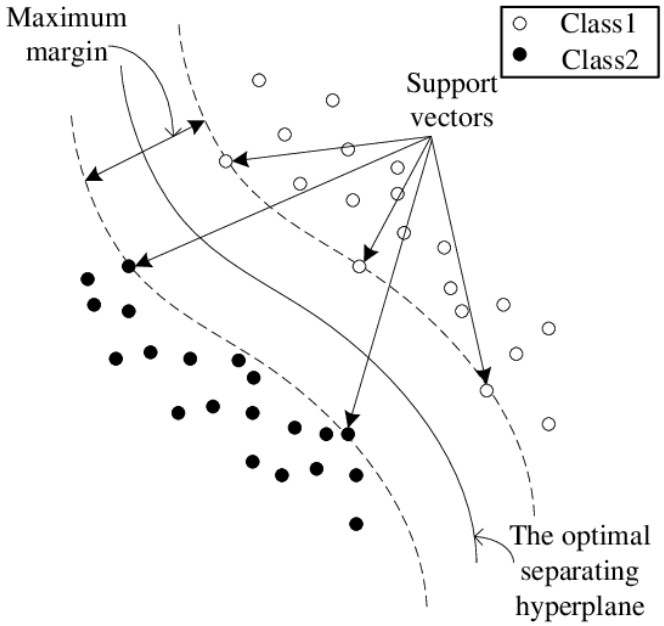


Fig. 3.8: Schematic representation of a Support Vector Machine with a polynomial kernel [96].

The SVm with RBF kernel has 2 hyperparameters that can be tuned: the C and γ (gamma) hyperparameters. The C hyperparameter controls the trade-off between maximizing the margin and minimizing the classification error. A small value of C creates a wider margin, allowing some misclassifications, which can result in a more general model. A large value of C aims to minimize the classification error, even at the expense of a narrower margin, which can lead to overfitting. In simpler terms, C determines the balance between finding the best possible separation between classes and avoiding overfitting. The γ hyperparameter in the RBF kernel influences the shape of the decision boundary, with smaller values leading to a more flexible boundary and larger values resulting in a more rigid one.

The following values for the hyperparameters are chosen to explore during the cross validation:

- C: $[10^{-3}, 10^{-2}, \ldots, 10^{6}]$ (10 values, logarithmically spaced)
- γ: $[10^{-7}, 10^{-6}, \ldots, 10^{2}]$ (10 values, logarithmically spaced)

3.5.2 Random Forest

Random Forest (RF) is an ensemble learning method used for both classification and regression tasks [97]. It consists of multiple decision trees, and the final output is determined by aggregating the predictions from each tree. In the case of classification, the majority vote from all trees is considered the final prediction.

The decision trees in a random forest are constructed using a random subset of the training data, and at each node, a random subset of features is considered for splitting. This randomization process results in a diverse set of trees, reducing overfitting and improving the generalization ability of the model. Figure 3.9 illustrates a Random Forest classifier, where multiple decision trees are combined to make the final prediction. Just as SVM, the RF model has been used for the baseline of the Interspeech 2012 Speaker Trait Challenge [28].

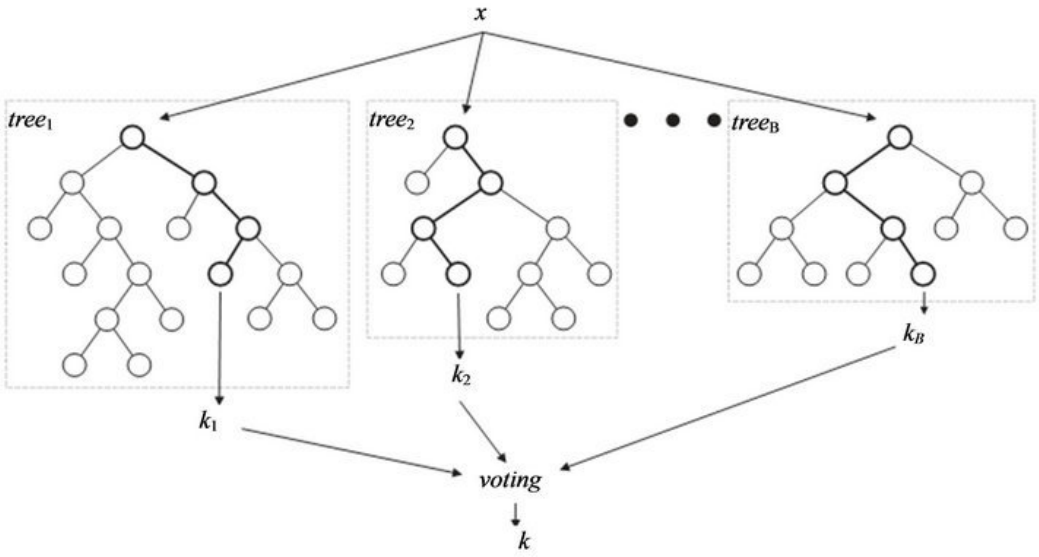


Fig. 3.9: Schematic representation of a Random Forest classifier.

There are multiple hyperparameters that can be tuned in RF models. Two of the most often used are the number of trees in the forest and the maximum tree depth. The number of trees affects the ensemble's ability to generalize and reduce overfitting. A larger number of trees generally leads to better performance and lower variance but increases the risk of overfitting and computational complexity. The maximum tree depth controls the complexity of the individual decision trees in the ensemble. A shallow tree may underfit the data, while a deep tree may overfit. By limiting the maximum depth, the model's capacity to overfit is reduced, resulting in a more generalizable model.

The following hyperparameters have been explored for the Random Forest models:

- Number of trees in the forest: $[10, 100, 250, 500, 1000]$.
- Maximum tree depth: $[2, 4, 8, 16, None]$.

3.5.3 k-Nearest Neighbors

The k-Nearest Neighbors (kNN) algorithm is a non-parametric learning method used for classification and regression tasks [98]. In the classification context, the kNN algorithm assigns a new data point to the majority class of its k nearest neighbours in the feature space.

The distance between data points can be calculated using various distance metrics, such as Euclidean, Manhattan, or Minkowski distance. Both the choice for the distance metric and the number of neighbours can affect the performance of the kNN algorithm. A smaller number of neighbours may lead to a more flexible decision boundary, which is prone to overfitting. In contrast, a larger number of neighbours results in a smoother decision boundary, which may underfit the data. The optimal number of neighbours balances the trade-off between overfitting and underfitting. Figure 3.10 shows a k-Nearest Neighbors classifier, where a new data point is classified based on the majority class of its k nearest neighbors.

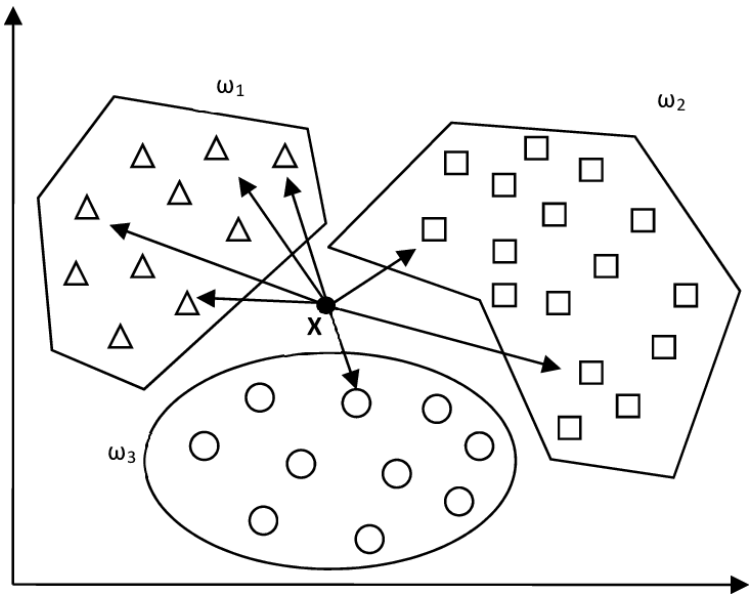


Fig. 3.10: Schematic representation of a k-Nearest Neighbors classifier.

For the kNN models, the following hyperparameter was tuned:

- Number of neighbors: $[1, 2, 3, ..., 20]$.

3.6 Evaluation Metrics

The performance of the trained classifiers will be evaluated using two metrics: the main metric Unweighted Average Recall (UAR) and the Area Under the Receiver Operating Characteristic curve (AUC ROC). Unweighted Average Recall is a metric that is often used for personality classification [28, 29, 62]. It is also known as the balanced accuracy in the case of binary classes. This metric, defined as $Sensitivity * 0.5 + Specificity * 0.5$, works well for imbalanced data as it gives equal

importance to both the positive and the negative class, regardless of their size. When the data is perfectly balanced, the UAR is the same as the accuracy.

The AUC ROC measures the performance of a classification model across all possible classification thresholds, quantifying the trade-off between sensitivity (true positive rate) and specificity (true negative rate). The AUC ROC ranges from 0 to 1, where a higher value indicates better classification performance. A value of 0.5 represents a random classifier, while a value of 1 indicates perfect classification.

AUC ROC is a suitable evaluation metric for comparing different models and datasets because it is robust to class imbalance, invariant to the choice of classification threshold, and provides an aggregate measure of performance across all possible thresholds. By incorporating AUC ROC alongside UAR, a more comprehensive explanation of the classifier's performance can be presented.

For the cross-validation of the development phase, the mean and standard deviation of the evaluation metrics are reported. During this phase, the best model from the different hyperparameters is selected for evaluation on the test set. Often in literature, this selection is based on the highest performance according to the primary evaluation metric, which in this case is the UAR. However, instances were observed during exploration where models with similar UAR scores had significantly different standard deviations. A high standard deviation during the development phase is indicative of a model's instability when encountering new data, which could be a result of overfitting or an inability to generalize effectively. Consequently, a high score on the test set may not truly reflect the model's real-world performance, as it may not consistently maintain this level of performance on other unseen data. Therefore, I consider both the mean performance and the standard deviation when selecting the optimal model for evaluation on the test set. This approach ensures that the chosen model not only has a high UAR, but is also stable and robust when applied to new data.

To address this issue, a ranking score that incorporates both the mean of the UAR and the standard deviation was used, calculated as $\bar{x} - 0.5 \times \sigma$, where $\bar{x}$ is the average UAR and σ is the standard deviation. This ranking score penalizes models with high standard deviations, favouring more stable models. Table 3.3 presents an example where, based on the mean UAR alone, Model 1 would have been selected. However, by incorporating the ranking score, the more stable Model 2 is chosen for evaluation.

For the test set, we will use a statistical technique called bootstrapping to obtain more reliable estimates of the evaluation metrics. Bootstrapping is a resampling method that involves repeatedly sampling with replacement to create multiple datasets of the same size. In this case, multiple test sets are created. By analysing the evaluation metrics obtained from these bootstrapped datasets, we can estimate the sampling distribution of the metrics, and in turn, derive confidence intervals around their estimates. Confidence intervals provide a range within which the true value of the metric is likely to fall, given a specified level of confidence. This is chosen to be

Model	Mean UAR	Std Dev UAR	Ranking score
Model 1	0.678	0.18	0.588
Model 2	0.665	0.06	0.635

Tab. 3.3: An example illustrating the impact of the adjusted ranking score: Model 2 is selected as the best model, considering both the UAR performance and stability, instead of Model 1, which would have been chosen based solely on the mean UAR.

95%, which is an often used standard in literature. If the lower confidence interval is higher than the baseline score to which the model is compared, the results can called significant.

In this study, we will bootstrap the test set 1,000 times and report the average of the bootstrapped evaluation metrics, together with the lower and upper bounds of the confidence interval. This should provide a more rigorous image of how well the different datasets and approaches perform.

Results

4

In this chapter, the results that were obtained from the experiments will be presented. Each dataset has a large table where for each personality trait the results from the evaluation metrics, the type of feature set and the model with the selected hyperparameters are presented. The evaluation metrics of the results include the average UAR and average AUC ROC, as explained in Section 3.6, for both the development and test phase. The average of the development phase is derived from the 5-fold cross validation, while the average in the test set is derived from the bootstrapped scores. Additionally, the standard deviation of the cross validation is presented, while the test phase includes the confidence intervals from the bootstrapped test set. For each trait, the model that performed best is highlighted with a grey background colour in the table. The best performing model is the model with the highest score for $\bar{x}$ UAR $+ 0.1 \times \bar{x}$ AUC ROC for the test set, where $\bar{x}$ is the average of the metric. This makes the UAR the main performance metric, but includes the AUC ROC in selecting the best model when models achieve similar UAR scores.

The results are also accompanied by graphs of the test set results. Each graph includes error bars, which show the 95% confidence interval of the bootstrapped test set. Furthermore, some graphs contain horizontal lines on the bars. These lines are the average UAR of the cross validation during the development phase. The inclusion of the average UAR from the cross validation provides additional information on the stability of the model. Additionally, a dashed line is included on the 0.5 score line, as this is the chance level for UAR.

4.1 APP & Natural Speech

The results in Table 4.1 show that for APP and natural speech, Extraversion was the best classifiable trait, with a best achieved UAR of 0.795 and an average UAR of 0.750. Openness was the least well classifiable trait with the best model having a UAR of 0.559, where its lower confidence bound is 0.47 which is below the chance level of 0.5. Figure 4.1 shows the best models for each trait for both the eGeMAPS features and the embeddings. The average UAR scores obtained from the cross validation in the development phase, shown with the horizontal lines on the bars, are all better for models trained on the embeddings. In general, they show similar performance compared to the test set, which indicates that the models are stable when being exposed to new data.

APP & Natural Speech

Speaker Personality Corpus

Extraversion			Development ($\bar{x} \pm s.d.$)		Test ($\bar{x} \pm c.i.$)	
Model	*Hyperparameters*	*Features*	*UAR*	*AUC ROC*	*UAR*	*AUC ROC*
SVM	$C = 1000,\ gamma = 1e-05$	eGeMAPS	0.767 ± 0.03	0.827 ± 0.05	$0.75 \pm (0.68, 0.82)$	$0.849 \pm (0.78, 0.91)$
RF	$max\ depth = 16,\ trees = 1000$	eGeMAPS	0.748 ± 0.05	0.82 ± 0.06	$0.795 \pm (0.72, 0.86)$	$0.868 \pm (0.8, 0.93)$
kNN	$neighbors = 1$	eGeMAPS	0.685 ± 0.03	0.75 ± 0.04	$0.728 \pm (0.66, 0.79)$	$0.84 \pm (0.77, 0.91)$
SVM	$C = 100000,\ gamma = 1e-05$	Embeddings	0.782 ± 0.01	0.844 ± 0.02	$0.751 \pm (0.68, 0.83)$	$0.854 \pm (0.79, 0.92)$
RF	$max\ depth = 16,\ trees = 1000$	Embeddings	0.77 ± 0.06	0.849 ± 0.07	$0.752 \pm (0.68, 0.82)$	$0.848 \pm (0.78, 0.91)$
kNN	$neighbors = 1$	Embeddings	0.745 ± 0.04	0.821 ± 0.06	$0.72 \pm (0.65, 0.8)$	$0.79 \pm (0.71, 0.87)$

Agreeableness			Development ($\bar{x} \pm s.d.$)		Test ($\bar{x} \pm c.i.$)	
Model	*Hyperparameters*	*Features*	*UAR*	*AUC ROC*	*UAR*	*AUC ROC*
SVM	$C = 10,\ gamma = 0.001$	eGeMAPS	0.654 ± 0.04	0.695 ± 0.07	$0.664 \pm (0.58, 0.75)$	$0.724 \pm (0.63, 0.81)$
RF	$max\ depth = none,\ trees = 500$	eGeMAPS	0.639 ± 0.04	0.681 ± 0.05	$0.665 \pm (0.58, 0.75)$	$0.728 \pm (0.63, 0.83)$
kNN	$neighbors = 7$	eGeMAPS	0.614 ± 0.04	0.628 ± 0.03	$0.597 \pm (0.51, 0.69)$	$0.683 \pm (0.58, 0.78)$
SVM	$C = 10,\ gamma = 0.001$	Embeddings	0.688 ± 0.06	0.734 ± 0.05	$0.701 \pm (0.61, 0.78)$	$0.778 \pm (0.68, 0.86)$
RF	$max\ depth = 8,\ trees = 500$	Embeddings	0.71 ± 0.05	0.768 ± 0.04	$0.657 \pm (0.57, 0.74)$	$0.742 \pm (0.64, 0.83)$
kNN	$neighbors = 1$	Embeddings	0.62 ± 0.03	0.675 ± 0.04	$0.627 \pm (0.54, 0.72)$	$0.692 \pm (0.59, 0.8)$

Conscientiousness			Development ($\bar{x} \pm s.d.$)		Test ($\bar{x} \pm c.i.$)	
Model	*Hyperparameters*	*Features*	*UAR*	*AUC ROC*	*UAR*	*AUC ROC*
SVM	$C = 100,\ gamma = 0.01$	eGeMAPS	0.665 ± 0.04	0.758 ± 0.05	$0.621 \pm (0.53, 0.71)$	$0.714 \pm (0.62, 0.81)$
RF	$max\ depth = none,\ trees = 500$	eGeMAPS	0.704 ± 0.02	0.775 ± 0.03	$0.691 \pm (0.6, 0.77)$	$0.756 \pm (0.66, 0.84)$
kNN	$neighbors = 8$	eGeMAPS	0.644 ± 0.02	0.708 ± 0.03	$0.585 \pm (0.51, 0.67)$	$0.682 \pm (0.58, 0.78)$
SVM	$C = 100,\ gamma = 0.001$	Embeddings	0.769 ± 0.01	0.85 ± 0.04	$0.673 \pm (0.59, 0.76)$	$0.783 \pm (0.7, 0.86)$
RF	$max\ depth = none,\ trees = 500$	Embeddings	0.769 ± 0.02	0.847 ± 0.03	$0.651 \pm (0.56, 0.74)$	$0.76 \pm (0.67, 0.84)$
kNN	$neighbors = 6$	Embeddings	0.73 ± 0.03	0.802 ± 0.03	$0.704 \pm (0.62, 0.79)$	$0.707 \pm (0.61, 0.81)$

Neuroticism			Development ($\bar{x} \pm s.d.$)		Test ($\bar{x} \pm c.i.$)	
Model	*Hyperparameters*	*Features*	*UAR*	*AUC ROC*	*UAR*	*AUC ROC*
SVM	$C = 10,\ gamma = 0.001$	eGeMAPS	0.681 ± 0.06	0.735 ± 0.04	$0.692 \pm (0.61, 0.77)$	$0.792 \pm (0.71, 0.87)$
RF	$max\ depth = 8,\ trees = 10$	eGeMAPS	0.665 ± 0.04	0.712 ± 0.07	$0.724 \pm (0.65, 0.8)$	$0.748 \pm (0.66, 0.83)$
kNN	$neighbors = 1$	eGeMAPS	0.661 ± 0.05	0.712 ± 0.04	$0.667 \pm (0.59, 0.74)$	$0.733 \pm (0.64, 0.82)$
SVM	$C = 10,\ gamma = 0.0001$	Embeddings	0.721 ± 0.02	0.788 ± 0.03	$0.731 \pm (0.64, 0.82)$	$0.825 \pm (0.74, 0.9)$
RF	$max\ depth = none,\ trees = 500$	Embeddings	0.724 ± 0.04	0.791 ± 0.03	$0.732 \pm (0.65, 0.81)$	$0.793 \pm (0.71, 0.87)$
kNN	$neighbors = 9$	Embeddings	0.646 ± 0.04	0.702 ± 0.06	$0.631 \pm (0.55, 0.71)$	$0.698 \pm (0.6, 0.79)$

Openness			Development ($\bar{x} \pm s.d.$)		Test ($\bar{x} \pm c.i.$)	
Model	*Hyperparameters*	*Features*	*UAR*	*AUC ROC*	*UAR*	*AUC ROC*
SVM	$C = 100000,\ gamma = 1e-05$	eGeMAPS	0.564 ± 0.02	0.573 ± 0.04	$0.463 \pm (0.39, 0.54)$	$0.518 \pm (0.42, 0.62)$
RF	$max\ depth = none,\ trees = 500$	eGeMAPS	0.544 ± 0.01	0.6 ± 0.03	$0.512 \pm (0.46, 0.57)$	$0.585 \pm (0.49, 0.69)$
kNN	$neighbors = 1$	eGeMAPS	0.518 ± 0.02	0.551 ± 0.03	$0.525 \pm (0.46, 0.59)$	$0.524 \pm (0.42, 0.62)$
SVM	$C = 1000,\ gamma = 0.0001$	Embeddings	0.59 ± 0.02	0.621 ± 0.04	$0.559 \pm (0.47, 0.65)$	$0.541 \pm (0.43, 0.65)$
RF	$max\ depth = 4,\ trees = 10$	Embeddings	0.568 ± 0.04	0.651 ± 0.06	$0.537 \pm (0.47, 0.61)$	$0.59 \pm (0.48, 0.7)$
kNN	$neighbors = 1$	Embeddings	0.552 ± 0.05	0.61 ± 0.09	$0.542 \pm (0.48, 0.62)$	$0.613 \pm (0.51, 0.72)$

Tab. 4.1: The results of the APP and natural speech experiment. The Development metrics show the mean and standard deviation of the cross-validation. The Test metrics show the mean of the bootstrapped test set with the lower and higher bound of the confidence interval. The first three models of each trait are trained on the eGeMAPS features and the second three models on the embeddings.

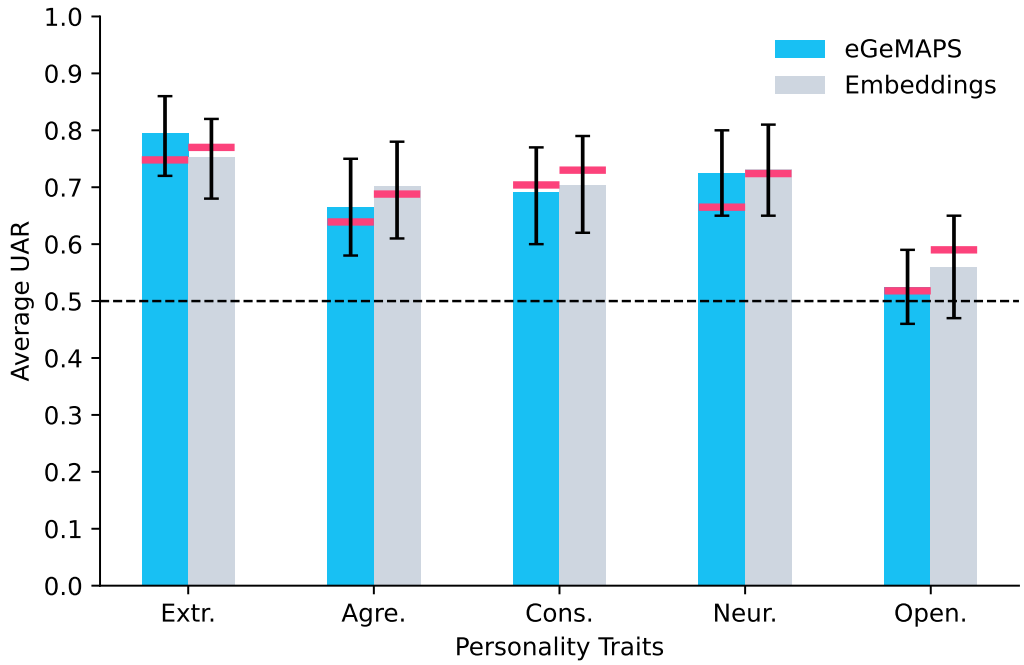


Fig. 4.1: The average UAR of the models for the eGeMAPS and the embeddings feature set for the SPC dataset. The error bar represents the average 95% confidence interval of the bootstrap. The horizontal line on each bar is the average UAR of the cross validation during the development phase. The dotted line is the chance level of 0.50.

Table 4.2 shows the results of the best models for both the eGeMAPS and embeddings feature sets, compared to baseline and challenge winner of the Speaker Trait Challenge [28]. The best models of this experiment outperform the baseline and challenge winner on Extraversion, Agreeableness, and Neuroticism. These results were tested as significant using a one sample t-test ($\alpha = 0.05$). Only for Neuroticism, both the embeddings and the eGeMAPS features would have outperformed the challenge baseline and winner. The good performance on Extraversion in line with earlier achieved results on the SPC. Furthermore, all models achieve relatively low results on Openness.

Methods	*Extraversion*	*Agreeableness*	*Conscientiousness*	*Neuroticism*	*Openness*
Best IS12 baseline	0.762 (0.841)	0.642 (0.667)	**0.801 (0.845)**	0.659 (0.718)	**0.59 (0.674)**
IS12 winner	0.765 (NA)	0.672 (NA)	0.77 (NA)	0.692 (NA)	0.564 (NA)
Best eGeMAPS	**0.795 (0.868)**	0.665 (0.728)	0.691 (0.756)	0.724 (0.748)	0.525 (0.524)
Best Embedding	0.752 (0.848)	**0.701 (0.778)**	0.704 (0.707)	**0.731 (0.825)**	0.559 (0.541)

Tab. 4.2: A comparison between UAR of the Interspeech 2012 Speaker Trait Challenge and the results obtained here. The 'Best IS12 baseline' represents the best UAR from both the SVM and RF models of the baseline. The 'IS12 winner' is the contribution to the challenge that was selected as the winner of the challenge, due to having the highest *average* UAR. The 'Best eGeMAPS' are the best scores obtained in this thesis using the eGeMAPS feature set, while the 'Best Embeddings' are the best obtained results using the pretrained model. The highest results are bolded.

Trait	eGeMAPS UAR ($\bar{x} \pm c.i.$)	Embeddings UAR ($\bar{x} \pm c.i.$)	Normality Test (p-value)	Significance Test (p-value)
Extraversion	$0.795 \pm (0.72, 0.86)$	$0.752 \pm (0.68, 0.82)$	$0.34,\ 0.45$	$<< 0.001$
Agreeableness	$0.665 \pm (0.57, 0.75)$	$0.701 \pm (0.61, 0.78)$	$0.99,\ 0.06$	$<< 0.001$
Conscientiousness	$0.691 \pm (0.60, 0.77)$	$0.704 \pm (0.62, 0.79)$	$0.09,\ 0.74$	$<< 0.001$
Neuroticism	$0.724 \pm (0.65, 0.82)$	$0.731 \pm (0.64, 0.81)$	$0.58,\ 0.08$	$<< 0.001$
Openness	$0.525 \pm (0.46, 0.59)$	$0.559 \pm (0.47, 0.65)$	$0.30,\ 0.12$	$<< 0.001$

Tab. 4.3: The results of a significance test for the SPC between the bootstrapped UAR scores of the best models trained on the eGeMAPS features and on the embeddings. For the normality test, a p-value higher than 0.05 means that the null-hypothesis that the bootstrapped values have a normal distribution cannot be rejected. The significance test, done with a paired t-test, shows that the difference between the bootstrapped results is significant. For both tests, $\alpha = 0.05$.

A significance test was performed on the scores obtained from the bootstrapped test set, to see if the best model using one type of feature was significantly better than the other for each trait, shown in Table 4.3. As both models were tested on the same bootstrapped test set values, a paired t-test was conducted. This test assumes a normal distribution, which is why the scores from the bootstrapped were also tested for normality. The resulting p-values were all far below 0.001, from which we can conclude that the best performing models for each trait outperformed the best performing models using the other feature type. As can be seen in the table, the best model trained on the embeddings outperformed the best model trained on the eGeMAPS features 4 out of 5 times.

In most cases, the performance of the different types of models was comparable between the two types of features. Interesting exceptions are the kNN models that classify Conscientiousness. In Table 4.1, it can be seen that the kNN model trained on eGeMAPS features scored worse than the SVM and RF trained on those features, with an average UAR of 0.585 and a lower confidence bound of 0.51, which is only slightly above chance. In contrast, the kNN model trained on the embeddings obtained the best UAR results of any of the models trained for Conscientiousness, with a UAR of 0.704 and a lower confidence bound of 0.62. This model also scored high in the development phase, with an average UAR of 0.730, which makes it more likely that the high test score cannot be contributed to luck.

4.2 APR & Controlled Speech

Table 4.5 shows the results of the models trained on the Nautilus Speaker Characterization corpus for APR and controlled speech. The best results are displayed in Figure 4.2. It shows that for all traits except Openness, a lower confidence bound above the chance level of 0.5 was achieved. To make sure the results are significantly above chance, the best performing models for each trait were tested using a one sample t-test. Table 4.4 shows that only Openness did not score significantly above the chance level. The best performing models shown in Figure 4.2 were also tested

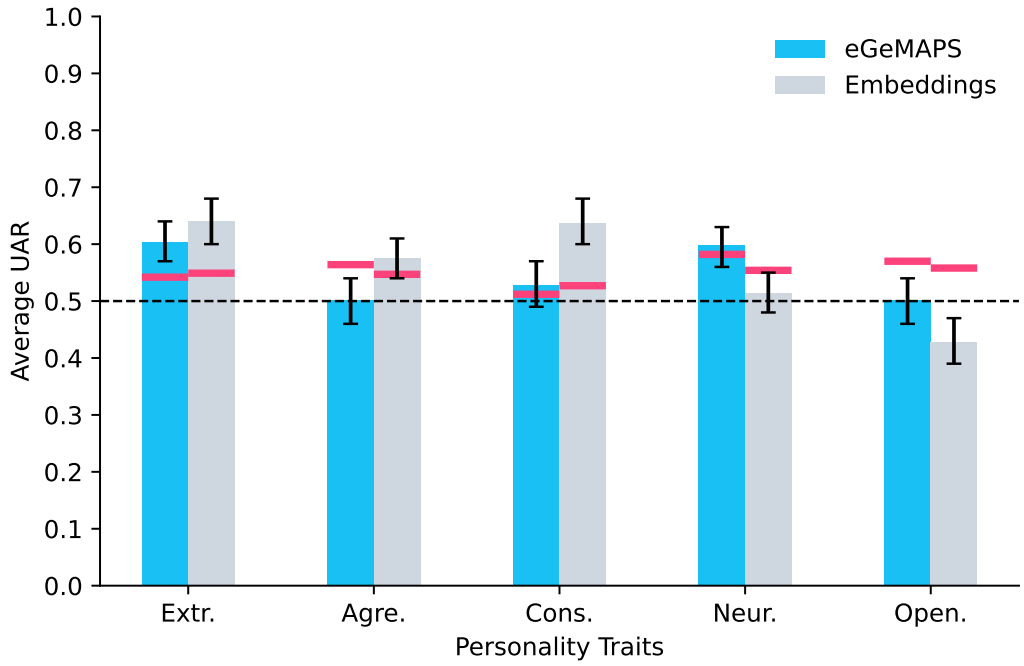


Fig. 4.2: The average UAR of the models for the eGeMAPS and the embeddings feature set for the NSC dataset. The error bar represents the average 95% confidence interval of the bootstrap. The dotted line is the chance level of 0.50.

Trait	Feature Set	UAR ($\bar{x} \pm c.i.$)	Significance Test (p-value)
Extraversion	Embeddings	$0.64 \pm (0.6, 0.68)$	$<< 0.001$
Agreeableness	Embeddings	$0.576 \pm (0.54, 0.61)$	$<< 0.001$
Conscientiousness	Embeddings	$0.637 \pm (0.6, 0.68)$	$<< 0.001$
Neuroticism	eGeMAPS	$0.598 \pm (0.56, 0.63)$	$<< 0.001$
Openness	eGeMAPS	$0.501 \pm (0.46, 0.54)$	> 0.05

Tab. 4.4: A test to see if the achieved scores for the NSC corpus are significantly above the chance level of 0.5 for $\alpha = 0.05$.

for significance compared to each other, just as with the SPC models. The results for this are in 4.6, which confirms that the best models were significantly better and that all models passed the test for normality.

As can be seen, the highest scores were achieved with an SVM trained on the embeddings for Extraversion and Conscientiousness with an UAR of respectively 0.640 and 0.637. It should be noted that the UAR during the development phase, shown with a horizontal line on the bar, was considerably lower for those models, with an average UAR of 0.541 and 0.527.

APR & Controlled Speech

Nautilus Speaker Characterization Corpus

Extraversion			Development ($\bar{x} \pm s.d.$)		Test ($\bar{x} \pm c.i.$)	
Model	*Hyperparameters*	*Features*	*UAR*	*AUC ROC*	*UAR*	*AUC ROC*
SVM	$C = 10000,\ gamma = 1e-07$	eGeMAPS	0.577 ± 0.04	0.584 ± 0.04	$0.54 \pm (0.5, 0.58)$	$0.569 \pm (0.52, 0.61)$
RF	$max\ depth = none,\ trees = 500$	eGeMAPS	0.542 ± 0.02	0.54 ± 0.02	$0.603 \pm (0.57, 0.64)$	$0.609 \pm (0.57, 0.65)$
kNN	$neighbors = 1$	eGeMAPS	0.54 ± 0.03	0.537 ± 0.04	$0.584 \pm (0.55, 0.62)$	$0.592 \pm (0.55, 0.64)$
SVM	$C = 100,\ gamma = 0.001$	Embeddings	0.549 ± 0.04	0.575 ± 0.07	$0.64 \pm (0.6, 0.68)$	$0.673 \pm (0.63, 0.72)$
RF	$max\ depth = 16,\ trees = 10$	Embeddings	0.545 ± 0.03	0.543 ± 0.03	$0.543 \pm (0.51, 0.58)$	$0.578 \pm (0.53, 0.62)$
kNN	$neighbors = 9$	Embeddings	0.53 ± 0.02	0.536 ± 0.03	$0.527 \pm (0.49, 0.57)$	$0.545 \pm (0.5, 0.59)$

Agreeableness			Development ($\bar{x} \pm s.d.$)		Test ($\bar{x} \pm c.i.$)	
Model	*Hyperparameters*	*Features*	*UAR*	*AUC ROC*	*UAR*	*AUC ROC*
SVM	$C = 100,\ gamma = 0.001$	eGeMAPS	0.564 ± 0.07	0.586 ± 0.08	$0.487 \pm (0.45, 0.52)$	$0.494 \pm (0.45, 0.54)$
RF	$max\ depth = 4,\ trees = 10$	eGeMAPS	0.564 ± 0.03	0.567 ± 0.03	$0.501 \pm (0.46, 0.54)$	$0.537 \pm (0.49, 0.58)$
kNN	$neighbors = 7$	eGeMAPS	0.514 ± 0.03	0.515 ± 0.04	$0.441 \pm (0.4, 0.48)$	$0.439 \pm (0.39, 0.48)$
SVM	$C = 100,\ gamma = 1e-05$	Embeddings	0.541 ± 0.02	0.573 ± 0.04	$0.567 \pm (0.53, 0.61)$	$0.588 \pm (0.54, 0.64)$
RF	$max\ depth = 2,\ trees = 500$	Embeddings	0.547 ± 0.02	0.575 ± 0.04	$0.576 \pm (0.54, 0.61)$	$0.564 \pm (0.52, 0.61)$
kNN	$neighbors = 3$	Embeddings	0.567 ± 0.03	0.579 ± 0.04	$0.528 \pm (0.49, 0.57)$	$0.54 \pm (0.5, 0.59)$

Conscientiousness			Development ($\bar{x} \pm s.d.$)		Test ($\bar{x} \pm c.i.$)	
Model	*Hyperparameters*	*Features*	*UAR*	*AUC ROC*	*UAR*	*AUC ROC*
SVM	$C = 100000,\ gamma = 1e-05$	eGeMAPS	0.531 ± 0.04	0.532 ± 0.05	$0.497 \pm (0.46, 0.54)$	$0.464 \pm (0.42, 0.51)$
RF	$max\ depth = 2,\ trees = 10$	eGeMAPS	0.512 ± 0.02	0.504 ± 0.05	$0.528 \pm (0.49, 0.57)$	$0.572 \pm (0.53, 0.62)$
kNN	$neighbors = 7$	eGeMAPS	0.507 ± 0.02	0.501 ± 0.02	$0.457 \pm (0.42, 0.49)$	$0.458 \pm (0.41, 0.5)$
SVM	$C = 100,\ gamma = 1e-05$	Embeddings	0.527 ± 0.03	0.548 ± 0.07	$0.637 \pm (0.6, 0.68)$	$0.62 \pm (0.58, 0.67)$
RF	$max\ depth = 2,\ trees = 10$	Embeddings	0.526 ± 0.01	0.533 ± 0.02	$0.589 \pm (0.55, 0.63)$	$0.601 \pm (0.56, 0.65)$
kNN	$neighbors = 1$	Embeddings	0.544 ± 0.01	0.562 ± 0.02	$0.577 \pm (0.54, 0.62)$	$0.589 \pm (0.55, 0.63)$

Neuroticism			Development ($\bar{x} \pm s.d.$)		Test ($\bar{x} \pm c.i.$)	
Model	*Hyperparameters*	*Features*	*UAR*	*AUC ROC*	*UAR*	*AUC ROC*
SVM	$C = 10000,\ gamma = 1e-05$	eGeMAPS	0.582 ± 0.05	0.604 ± 0.04	$0.598 \pm (0.56, 0.63)$	$0.616 \pm (0.57, 0.66)$
RF	$max\ depth = 16,\ trees = 10$	eGeMAPS	0.517 ± 0.02	0.529 ± 0.03	$0.564 \pm (0.52, 0.6)$	$0.585 \pm (0.54, 0.63)$
kNN	$neighbors = 2$	eGeMAPS	0.52 ± 0.01	0.523 ± 0.01	$0.556 \pm (0.52, 0.6)$	$0.605 \pm (0.56, 0.65)$
SVM	$C = 100000,\ gamma = 1e-05$	Embeddings	0.594 ± 0.04	0.633 ± 0.05	$0.454 \pm (0.42, 0.49)$	$0.443 \pm (0.4, 0.49)$
RF	$max\ depth = 16,\ trees = 500$	Embeddings	0.554 ± 0.04	0.578 ± 0.07	$0.514 \pm (0.48, 0.55)$	$0.529 \pm (0.48, 0.57)$
kNN	$neighbors = 2$	Embeddings	0.547 ± 0.02	0.562 ± 0.02	$0.498 \pm (0.47, 0.53)$	$0.494 \pm (0.46, 0.54)$

Openness			Development ($\bar{x} \pm s.d.$)		Test ($\bar{x} \pm c.i.$)	
Model	*Hyperparameters*	*Features*	*UAR*	*AUC ROC*	*UAR*	*AUC ROC*
SVM	$C = 10000,\ gamma = 1e-05$	eGeMAPS	0.57 ± 0.03	0.596 ± 0.03	$0.501 \pm (0.46, 0.54)$	$0.499 \pm (0.46, 0.54)$
RF	$max\ depth = 4,\ trees = 500$	eGeMAPS	0.583 ± 0.04	0.601 ± 0.05	$0.37 \pm (0.33, 0.41)$	$0.361 \pm (0.32, 0.4)$
kNN	$neighbors = 1$	eGeMAPS	0.561 ± 0.01	0.569 ± 0.03	$0.423 \pm (0.39, 0.46)$	$0.38 \pm (0.34, 0.43)$
SVM	$C = 100,\ gamma = 0.0001$	Embeddings	0.558 ± 0.03	0.57 ± 0.04	$0.428 \pm (0.39, 0.47)$	$0.408 \pm (0.37, 0.45)$
RF	$max\ depth = 4,\ trees = 250$	Embeddings	0.564 ± 0.02	0.584 ± 0.03	$0.381 \pm (0.34, 0.42)$	$0.388 \pm (0.34, 0.43)$
kNN	$neighbors = 1$	Embeddings	0.523 ± 0.02	0.516 ± 0.03	$0.411 \pm (0.37, 0.45)$	$0.383 \pm (0.34, 0.43)$

Tab. 4.5: The results of the APR and controlled speech experiment. The Development metrics show the mean and standard deviation of the cross-validation. The Test metrics show the mean of the bootstrapped test set with the lower and higher bound of the confidence interval. The first three models of each trait are trained on the eGeMAPS features and the second three models on the embeddings.

Trait	eGeMAPS UAR ($\bar{x} \pm c.i.$)	Embeddings UAR ($\bar{x} \pm c.i.$)	Normality Test (p-value)	Significance Test (p-value)
Extraversion	$0.603 \pm (0.57, 0.64)$	$0.64 \pm (0.60, 0.68)$	$0.70,\ 0.96$	$<< 0.001$
Agreeableness	$0.501 \pm (0.46, 0.54)$	$0.576 \pm (0.54, 0.61)$	$0.52,\ 0.39$	$<< 0.001$
Conscientiousness	$0.528 \pm (0.49, 0.57)$	$0.637 \pm (0.60, 0.68)$	$0.80,\ 0.96$	$<< 0.001$
Neuroticism	$0.598 \pm (0.56, 0.63)$	$0.514 \pm (0.48, 0.55)$	$0.60,\ 0.85$	$<< 0.001$
Openness	$0.501 \pm (0.46, 0.54)$	$0.428 \pm (0.39, 0.47)$	$0.23,\ 0.68$	$<< 0.001$

Tab. 4.6: The results of a significance test for the NSC corpus between the bootstrapped UAR scores of the best models trained on the eGeMAPS features and on the embeddings. For the normality test, a p-value higher than 0.05 means that the null-hypothesis that the bootstrapped values have a normal distribution cannot be rejected. The significance test, done with a paired t-test, shows that the difference between the bootstrapped results is significant. For both tests, $\alpha = 0.05$.

4.3 APR & Natural Speech

Table 4.8 shows the results of the models trained on the REMDE corpus for APR and Natural Speech. The results are mostly around or below the chance level. Furthermore, multiple models have a UAR of 0.5 with equivalent lower and upper confidence bounds on the test set. This is caused by the model guessing only one class for all the test samples, which results in an UAR of 0.5. Table 4.7 shows all models except Conscientiousness scored significantly above chance in the evaluation phase. However, they also obtained a low average UAR and standard deviation of the cross validation. The highest UAR is achieved by the kNN trained on embeddings for Agreeableness with a UAR of 0.616. Also here, development UAR for this model is low with a score of 0.409 and a high standard deviation of 0.11, indicating an unstable model and thereby making the achieved test score unreliable.

Trait	Feature Set	UAR ($\bar{x} \pm c.i.$)	Significance Test (p-value)
Extraversion	Embeddings	$0.505 \pm (0.44, 0.58)$	< 0.001
Agreeableness	Embeddings	$0.616 \pm (0.54, 0.69)$	< 0.001
Conscientiousness	Embeddings	$0.5 \pm (0.5, 0.5)$	–
Neuroticism	Embeddings	$0.53 \pm (0.46, 0.6)$	< 0.001
Openness	Embeddings	$0.549 \pm (0.48, 0.62)$	< 0.001

Tab. 4.7: A test to see if the achieved scores for the REMDE corpus are significantly above the chance level of 0.5 for $\alpha = 0.05$. The dash means that no significance test could be performed, as all the scores were the same as the value that was being tested for (0.5).

APR & Natural Speech

REMDE Corpus

Extraversion			Development ($\bar{x} \pm s.d.$)		Test ($\bar{x} \pm c.i.$)	
Model	*Hyperparameters*	*Features*	*UAR*	*AUC ROC*	*UAR*	*AUC ROC*
SVM	$C = 100,\ gamma = 1e-06$	eGeMAPS	0.503 ± 0	0.517 ± 0.14	$0.5 \pm (0.5, 0.5)$	$0.47 \pm (0.38, 0.56)$
RF	$max\,depth = none,\ trees = 10$	eGeMAPS	0.477 ± 0.14	0.488 ± 0.2	$0.444 \pm (0.37, 0.52)$	$0.435 \pm (0.35, 0.52)$
kNN	$neighbors = 2$	eGeMAPS	0.512 ± 0.1	0.491 ± 0.13	$0.421 \pm (0.35, 0.5)$	$0.405 \pm (0.33, 0.49)$
SVM	$C = 1000,\ gamma = 0.01$	Embeddings	0.506 ± 0.01	0.489 ± 0.14	$0.5 \pm (0.5, 0.5)$	$0.58 \pm (0.49, 0.67)$
RF	$max\,depth = none,\ trees = 10$	Embeddings	0.512 ± 0.11	0.505 ± 0.2	$0.505 \pm (0.44, 0.58)$	$0.476 \pm (0.39, 0.56)$
kNN	$neighbors = 4$	Embeddings	0.517 ± 0.08	0.506 ± 0.16	$0.49 \pm (0.42, 0.57)$	$0.515 \pm (0.43, 0.6)$

Agreeableness			Development ($\bar{x} \pm s.d.$)		Test ($\bar{x} \pm c.i.$)	
Model	*Hyperparameters*	*Features*	*UAR*	*AUC ROC*	*UAR*	*AUC ROC*
SVM	$C = 0,\ gamma = 0.0001$	eGeMAPS	0.5 ± 0	0.351 ± 0.13	$0.5 \pm (0.5, 0.5)$	$0.541 \pm (0.44, 0.63)$
RF	$max\,depth = none,\ trees = 10$	eGeMAPS	0.408 ± 0.1	0.376 ± 0.13	$0.519 \pm (0.45, 0.59)$	$0.509 \pm (0.41, 0.61)$
kNN	$neighbors = 3$	eGeMAPS	0.457 ± 0.13	0.432 ± 0.16	$0.572 \pm (0.5, 0.65)$	$0.583 \pm (0.49, 0.67)$
SVM	$C = 10,\ gamma = 0.1$	Embeddings	0.5 ± 0	0.497 ± 0.01	$0.5 \pm (0.5, 0.5)$	$0.5 \pm (0.5, 0.5)$
RF	$max\,depth = 8,\ trees = 10$	Embeddings	0.446 ± 0.11	0.404 ± 0.16	$0.495 \pm (0.42, 0.57)$	$0.508 \pm (0.42, 0.6)$
kNN	$neighbors = 1$	Embeddings	0.409 ± 0.11	0.409 ± 0.11	$0.616 \pm (0.54, 0.69)$	$0.616 \pm (0.54, 0.69)$

Conscientiousness			Development ($\bar{x} \pm s.d.$)		Test ($\bar{x} \pm c.i.$)	
Model	*Hyperparameters*	*Features*	*UAR*	*AUC ROC*	*UAR*	*AUC ROC*
SVM	$C = 0,\ gamma = 0.0001$	eGeMAPS	0.5 ± 0	0.41 ± 0.13	$0.5 \pm (0.5, 0.5)$	$0.379 \pm (0.3, 0.46)$
RF	$max\,depth = 2,\ trees = 10$	eGeMAPS	0.42 ± 0.09	0.364 ± 0.15	$0.447 \pm (0.41, 0.49)$	$0.341 \pm (0.26, 0.42)$
kNN	$neighbors = 2$	eGeMAPS	0.452 ± 0.07	0.432 ± 0.09	$0.445 \pm (0.38, 0.51)$	$0.409 \pm (0.34, 0.48)$
SVM	$C = 10,\ gamma = 0.01$	Embeddings	0.504 ± 0.01	0.475 ± 0.09	$0.5 \pm (0.5, 0.5)$	$0.411 \pm (0.33, 0.5)$
RF	$max\,depth = 4,\ trees = 250$	Embeddings	0.485 ± 0.09	0.463 ± 0.11	$0.394 \pm (0.34, 0.45)$	$0.272 \pm (0.21, 0.35)$
kNN	$neighbors = 2$	Embeddings	0.503 ± 0.05	0.467 ± 0.07	$0.366 \pm (0.3, 0.44)$	$0.409 \pm (0.33, 0.49)$

Neuroticism			Development ($\bar{x} \pm s.d.$)		Test ($\bar{x} \pm c.i.$)	
Model	*Hyperparameters*	*Features*	*UAR*	*AUC ROC*	*UAR*	*AUC ROC*
SVM	$C = 0,\ gamma = 0.0001$	eGeMAPS	0.5 ± 0	0.481 ± 0.1	$0.5 \pm (0.5, 0.5)$	$0.61 \pm (0.53, 0.69)$
RF	$max\,depth = 2,\ trees = 500$	eGeMAPS	0.497 ± 0.1	0.433 ± 0.11	$0.498 \pm (0.46, 0.53)$	$0.458 \pm (0.38, 0.54)$
kNN	$neighbors = 2$	eGeMAPS	0.424 ± 0.09	0.407 ± 0.11	$0.497 \pm (0.44, 0.56)$	$0.489 \pm (0.42, 0.57)$
SVM	$C = 10,\ gamma = 0.1$	Embeddings	0.5 ± 0	0.511 ± 0.01	$0.5 \pm (0.5, 0.5)$	$0.5 \pm (0.5, 0.5)$
RF	$max\,depth = 2,\ trees = 500$	Embeddings	0.497 ± 0.09	0.448 ± 0.19	$0.424 \pm (0.37, 0.48)$	$0.529 \pm (0.44, 0.62)$
kNN	$neighbors = 1$	Embeddings	0.448 ± 0.11	0.448 ± 0.11	$0.53 \pm (0.46, 0.6)$	$0.53 \pm (0.46, 0.6)$

Openness			Development ($\bar{x} \pm s.d.$)		Test ($\bar{x} \pm c.i.$)	
Model	*Hyperparameters*	*Features*	*UAR*	*AUC ROC*	*UAR*	*AUC ROC*
SVM	$C = 100,\ gamma = 1e-06$	eGeMAPS	0.503 ± 0	0.416 ± 0.11	$0.507 \pm (0.46, 0.55)$	$0.48 \pm (0.4, 0.56)$
RF	$max\,depth = 8,\ trees = 10$	eGeMAPS	0.461 ± 0.08	0.443 ± 0.09	$0.411 \pm (0.35, 0.48)$	$0.325 \pm (0.25, 0.41)$
kNN	$neighbors = 6$	eGeMAPS	0.46 ± 0.07	0.411 ± 0.08	$0.4 \pm (0.33, 0.48)$	$0.372 \pm (0.29, 0.46)$
SVM	$C = 100000,\ gamma = 1e-05$	Embeddings	0.549 ± 0.09	0.536 ± 0.11	$0.478 \pm (0.41, 0.55)$	$0.531 \pm (0.45, 0.61)$
RF	$max\,depth = 8,\ trees = 10$	Embeddings	0.524 ± 0.08	0.526 ± 0.13	$0.549 \pm (0.48, 0.62)$	$0.564 \pm (0.48, 0.64)$
kNN	$neighbors = 2$	Embeddings	0.499 ± 0.05	0.508 ± 0.09	$0.453 \pm (0.39, 0.51)$	$0.457 \pm (0.38, 0.53)$

Tab. 4.8: The results of the APR and natural speech experiment. The Development metrics show the mean and standard deviation of the cross-validation. The Test metrics show the mean of the bootstrapped test set with the lower and higher bound of the confidence interval. The first three models of each trait are trained on the eGeMAPS features and the second three models on the embeddings.

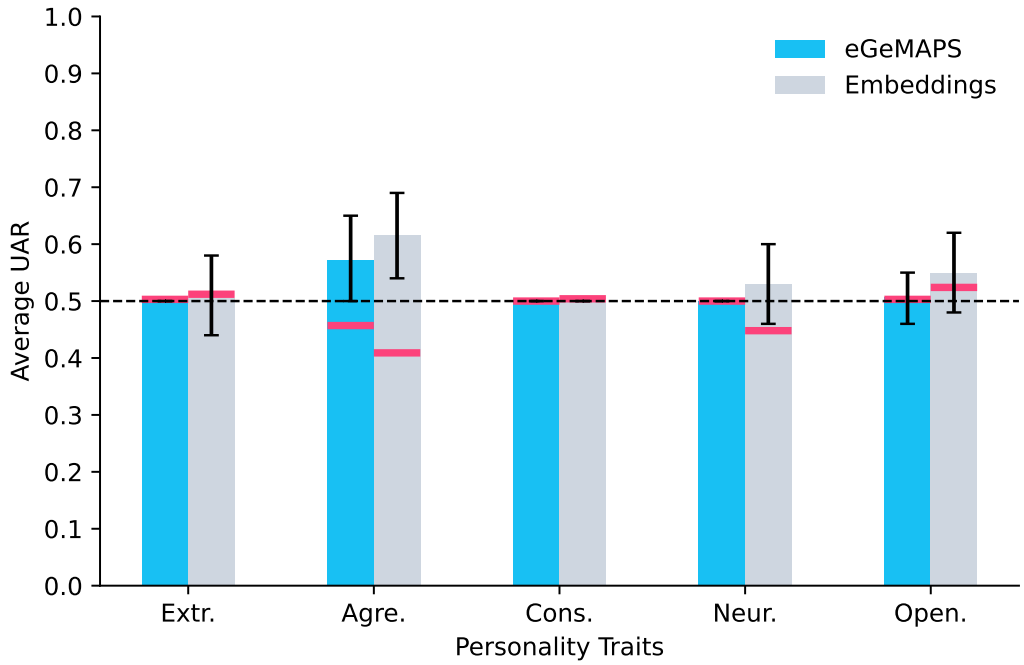


Fig. 4.3: The average UAR of the models for the eGeMAPS and the embeddings feature set for the REMDE dataset. The error bar represents the average 95% confidence interval of the bootstrap. The dotted line is the chance level of 0.50.

4.4 The datasets together

Figure 4.4 shows the best overall models for each trait and dataset. The models trained on the SPC dataset for APP and natural speech style clearly achieved the best overall results. It is interesting that there is no clear pattern when it comes to most of the traits. They all scored relatively low on Openness. However, for the SPC dataset, the highest score was achieved for Extraversion, while for the REMDE dataset, this score could not be classified better than chance.

Table 4.9 shows for how many traits each feature types won over the other type. This is shown for the different datasets and for the models. The left number counts the number of traits for the eGeMAPS feature and the right number for the embeddings. When both models did not have a significant score, they were not included. The table shows that the models trained on the embeddings outperformed those on the eGeMAPS features for all datasets. Furthermore, the embeddings worked better across the datasets for the SVM and the kNN models, while it performed equal for the RF models.

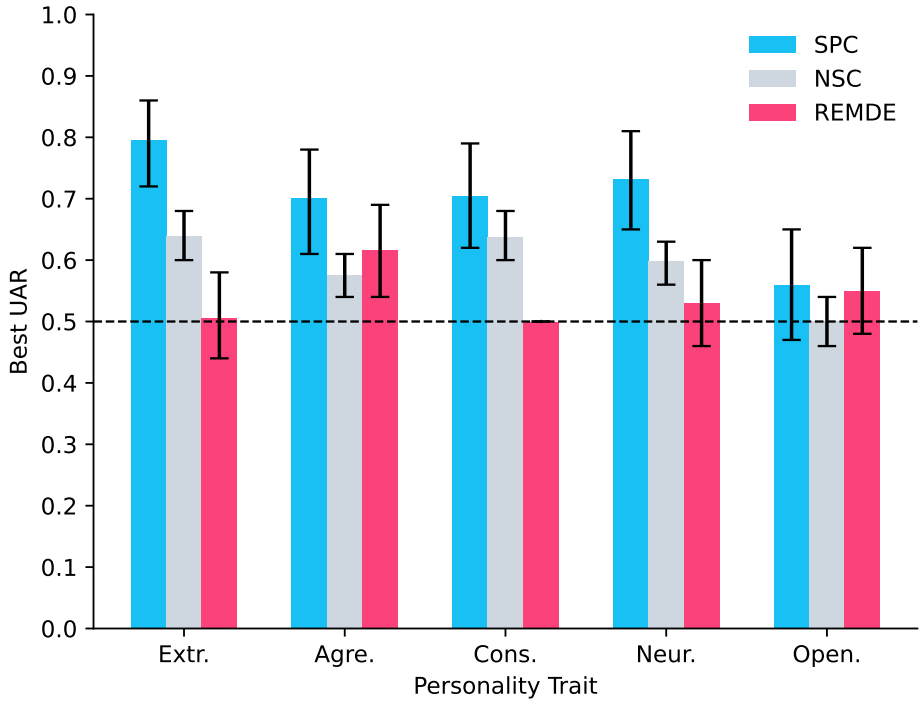


Fig. 4.4: The best achieved result on the personality traits for each of the datasets. The error bar represents the 95% confidence interval of the bootstrap. The dotted line is the chance level of 0.50.

	SPC	NSC	REMDE	*Total*
SVM	(0, 5)	(2, 3)	-	(2, 8)
RF	(3, 2)	(2, 3)	-	(5, 5)
kNN	(2, 3)	-	(0, 2)	(2, 5)
Total	(5, 10)	(4, 6)	(0, 2)	**(9, 18)**

Tab. 4.9: The number of personality traits where each feature type won over the other. The left number is for the eGeMAPS features and the right number for the embeddings. Only traits where at least one of the two models was significant above chance ($p<0.05$) are included.

Discussion and limitations

5

5.1 Discussion of results

In this thesis, automatic personality classification is explored by combining different labelling methods (APR and APP) and different speech styles (controlled and natural speech). Three types of models were trained on four datasets to explore each of the combinations. Furthermore, both hand-crafted features and embeddings extracted from a pretrained model were used and compared. In this chapter, I will discuss the results that were presented in Chapter 4.

Out of all the experiments, the APP and natural speech experiment trained on the SPC obtained the best results. As was shown in Table 4.2, Extraversion, Agreeableness, and Neuroticism achieved a better UAR than the baseline and challenge winner of the Interspeech 2012 Speaker Trait Challenge [28]. The SVM and RF that were used in this experiment had also been used for the challenge baseline [28]. The increase in performance compared to the Speaker Trait Challenge could therefore be caused by the implementation of more effective features. The challenge was released in 2012 and provided an extensive feature set of 6,125 features [28]. In 2015, the eGeMAPS feature set was released with a relatively low number of 88 features [51]. These features were selected based on their success in past literature and their theoretical significance. This seems to have resulted in increased performance on the SPC corpus compared to the challenge baseline. The embeddings, of which Table 4.3 showed it significantly outperformed the eGeMAPS feature set on 4 out of 5 personality traits, obtained even better results. The embeddings are from a pretrained deep learning model released in 2021 [53] that achieved near state-of-the-art results on related paralinguistic classification tasks. The results obtained in this experiment indicate that the deep learning model is also an effective feature extractor for perceived personality classification based on paralinguistic information.

In Figure 2.2, it could be seen that all the participants of the Personality Sub Challenge from the Interspeech 2012 Speaker Trait Challenge [28] achieved the best results on Extraversion and Conscientiousness. In related work on personality classification, Extraversion is often among the traits on which the highest performance is achieved [29]. This is also the case with the models trained on the SPC corpus. For a trait to be successfully classified, it must be perceptible [5]. Extraversion, which manifests itself in characteristics such as being talkative, is expressed *externally*, while characteristics such as fantasy proneness, which are related to Openness, are expressed *internally* [5]. This could contribute to why Extraversion often performs well on personality classification tasks.

While the results for Extraversion are in line with the results of the challenge contributions, Conscientiousness scored worse in this experiment (0.701 in the best

embedding model versus 0.801 in the challenge baseline), as is shown in Table 4.2. This is most likely explained by the different label distribution that was calculated for the SPC in this thesis. Table 3.1 showed that the difference between the labels is 127 for each class of Conscientiousness. This makes the majority class in this experiment 'High' on Conscientiousness, while the majority class in the challenge distribution was 'Low' on Conscientiousness.

Although this big difference in the label distribution resulted in a lower score for Conscientiousness compared to the challenge baseline, a score of 0.701 was still achieved, thereby outperforming Agreeableness and Openness in the same experiment. This raises the question of how, with such a big difference in the label distribution, both Conscientiousness models can achieve a score above 0.7. It can be explained by the impact that a 'hard' binary threshold has on the label distribution. The raw personality scores have to be converted to binary scores at some point in the label calculation. A hard cut-off is used, which causes everything below the threshold to be labelled as 'Low' on the trait, while everything above the threshold is labelled as 'High' on the trait. If a large amount of the personality scores for a trait have values close to the binary threshold, slight differences in where the threshold is set could have a relatively big impact on the binary label distribution. At the same time, the range in which a personality score will be categorized as one of the two classes mostly stays the same. Therefore, a change in the value of the binary threshold can have a relatively big impact on the label distribution, while good results for that trait can still be achieved.

Table 4.4 showed that in the APR and controlled speech experiment with the NSC corpus, scores significantly higher than chance were obtained for all traits except Openness. The best score was obtained for Extraversion. This further supports the explanation that traits related to external expression (such as Extraversion) are more suitable for classification tasks based on paralinguistic information than traits related to internal expression (such as Openness) [5]. The significance test between the models trained on the eGeMAPS features and the embeddings, of which the results are in Table 4.6, showed that the embeddings outperformed the eGeMAPS features for 3 out of 5 traits. The embeddings therefore also performed better for the experiment with APR and controlled speech.

In Figure 4.2 of the APR and controlled speech experiment, it can be seen that for all traits except Neuroticism, at least one of the models had a cross validation score, shown with the horizontal line on the bar, that is out of the bounds of the confidence interval for the test score. Many of the models therefore performed considerably different during the development phase when compared to the evaluation phase. This could indicate that the model is unstable when seeing new data or that the samples in the test set were either favourable or unfavourable to that specific trait. The instability can be further examined by looking at the standard deviation of the cross validation. Table 4.5 shows that the standard deviations for the best performing models in the development phase were low. The highest standard deviation out of

all the models that showed inconsistency between the development and evaluation phase, presented in Figure 4.2, is 0.04. These standard deviations go against the explanation that the models are unstable. The differences in performance between the development and the evaluation phase are therefore more likely to be caused by the configuration of the test set. The database is split so that no samples from the same speaker are divided among the training and test set. The test set therefore has a limited subset of the speakers. In APR, all samples belonging to the same speaker will have the same labels. Therefore, the extent to which a trait is prevalent in the test set is highly dependent on the speakers that are included in the set.

The results of the experiment for APR and natural speech with the REMDE corpus, shown in Table 4.8, indicate that personality mostly could not be classified in this experiment. Although Table 4.7 shows that the best models achieved a statistically significant result for 4 out of 5 traits, their average UAR and standard deviation of the cross validation perform badly. This suggests that the scores that are deemed statically significant can likely be contributed to luck. Taking all the metrics into consideration, the models are not able to get good results for the REMDE dataset. The overall best performing model had an average score below chance during the development phase with a relatively high standard deviation of 0.11, which are signs of a badly performing and unstable model.

Overall, the best results were obtained on the SPC, which is visible in Figure 4.4. It cannot be said with certainty whether this because of APP and natural speech, or because other characteristics of the database make it more suitable for classification. Classifying personality based on unchanging self-assessed personality labels is seen as harder than classifying perceived personality. If the audio recorded does not reflect the scores obtained from the personality test, it will be difficult to identify patterns. The lower results of both APR datasets compared to the APP dataset are in line with this. The experimental setup could also influence the results. The speakers in the REMDE corpus had conversations with a digital avatar in an intimate setting. In contrast, the SPC contains audio from a radio channel. This is a medium where the speakers are aware that their voice is the only way to communicate to the listeners. This could elicit the personality traits of the speaker. Regarding the embeddings and eGeMAPS features, Table 4.9 showed that the embeddings outperformed the eGeMAPS features in most cases. The results of the experiments have shown that the differences in performance were statistically significant. This indicates that the embeddings are suitable for personality classification and can regularly obtain higher scores than the hand-crafted features.

5.2 Limitations

As explained before, using different datasets will impose limitations on the comparisons that can be made. Ideally, a dataset has labels from both the self-assessed personality scores, as personality scores based on the perception of the raters. This

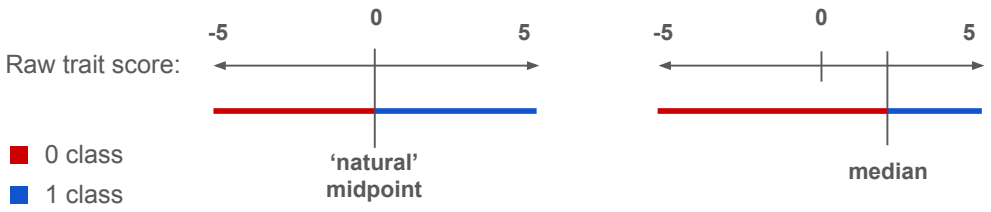


Fig. 5.1: The range in which personality scores will be classified as one of the two classes when the binary threshold is 0, compared to when the binary threshold is set as the median (given the median is not 0). Shifting the binary threshold away from 0 makes this range for one class larger, while the other class has less 'space' in which a score will be classified as such.

would enable direct comparison between APR and APP. For speech styles, an ideal situation would have participants both speak freely and act out scenarios. Because of this limitation, the labelling method and speech style analysis focusses mostly on the results for each dataset specifically. Personality being more difficult to classify in APR datasets than in APP datasets does align with existing research. However, it cannot be said with certainty that this is what caused the results in these experiments. Furthermore, the datasets were recorded in 3 different languages. Language could affect the way personality is prevalent through speech.

A limitation specific to the APR datasets is the distribution of the self-assessed personality scores. Figure 3.3 and 3.4 showed how the scores did not distribute evenly around the midpoint between the highest and lowest possible score for a trait. Because determining binary scores based on this midpoint would produce a highly imbalanced label distribution, the median was chosen as the binary threshold. Although this results in more balanced label distributions, it also alters the range in which scores will be categorized as one of the two classes. This is effect is illustrated in Figure 5.1. It shows the full range of a personality trait from -5 to 5, where 0 is the midpoint. In the left image, the midpoint is chosen as the binary threshold, resulting in an equal range for both classes. The right image shows how these ranges change when the median (in this case 0.2) is chosen as the binary threshold. It becomes clear that the participant must score extremely high on this trait in order to be classified as 'High'. Simultaneously, participants that would score above 0 but below the threshold for this trait, will now be classified as being 'Low' on the personality trait. A speaker scoring 0.15 will now have the same label as a person scoring -3, even though their answers on the self-assessment test might have been very different.

The data division with multiple labels could also influence the performance. The datasets are divided into a training set for the development phase, and a test set. This is done so that the binary labels of all traits and gender are relatively well stratified, and so that the speakers are not shared among training and test set. The

data in the test set could, by chance, contain speakers that much stronger represent a specific trait than the data that was used for cross validation. This would cause a model to perform pessimistically during the development phase when compared to the test set. This effect can especially play a role in APR, as all samples from a speaker have the same label.

Finally, there are also limitations intrinsic to the task of classifying self-assessed personality from paralinguistic information. The goal of APR is to classify a person's persistent personality, as assessed by the person himself. The definition of personality, as stated in Chapter 1, defines personality to consist of an enduring set of traits. Short speech segments might simply not carry the information that is needed to identify these enduring traits, as they are not defined on such a timescale. This is a limitation mostly specific to APR, as the task of APP aims to classify the perception of that specific short segment. A second general limitation is the extent to which traits are shown in paralinguistic information. Not all traits are expressed externally in equal ways. This makes certain traits more 'available' in the data [5]. There could be limits to what extent personality traits such as Openness can be classified using externally expressed information such as speech.

Future Work

6

There are multiple extensions to this thesis that would make interesting contributions. One such extension would be to fine-tune the deep learning model. When fine-tuning a pretrained model, you unfreeze the last few layers of the model. You then retrain those layers using your own data. It thereby becomes an end-to-end model implementation, removing the need for the shallow classification models that were used in this thesis. Fine-tuning has been shown to sometimes produce better results as it can effectively capture information that shallow models can not [52].

An improvement to the collection of datasets would be an APP dataset with controlled speech. With such a dataset, all combinations of the labelling methods and speech styles can be investigated. This could reveal additional patterns on the effectiveness of each labelling method and speech style across databases. A database like this can be made by creating APP labels for the Nautilus Speaker Characterization corpus. This would simultaneously enable direct comparisons between APP and APR using controlled speech.

Additionally, APP labels could be created for the REMDE dataset. This would target the same combination as the Speaker Personality Corpus, but would also make direct comparisons between the APR and APP on controlled speech possible. As the models performed the worst on the REMDE dataset, it would be interesting to see if using labels based on perception would improve the scores.

Creating APP labels for the datasets is a time and resource costly operation, as each sample needs to be assessed by external raters. Obtaining APR labels for the APP datasets would require less labelling. However, the speakers in the datasets would have to be found and participate in a personality self-assessment test. For a database such as the SPC, where the audio is extracted from radio, this is not feasible. Creating APP labels for APR datasets would therefore be the best solution in most cases, and would be a valuable contribution to speech-based personality classification research.

The transcriptions of the speech data could be used, so that semantics are included in the speech analysis of the personality traits. This would provide an additional layer of information to the speech. Furthermore, most research on speech personality classification uses binary labels. In some Personality Computing research, different approaches have already been tried, such as using 3 classes [29]. These approaches could be further explored to see if this improves performance.

It would also be valuable to experiment with different training and test set divisions. This could give more clarity on what caused the difference between the cross validation and evaluation averages on the APR datasets. However, personality classification as done in this thesis is a multi-label task with speaker-independent

training and test sets. This limits the number of ways in which a dataset can be divided while still being stratified for all the traits.

Interesting future work on the REMDE dataset would be to combine the different recorded modalities, such as eye tracking and brain activity, with the speech data. This would target the limitation that some traits are expressed mostly internally. Related work making use of other modalities, such as facial expressions [79], has been shown to produce very high results, even with self-assessed personality labels [79].

Conclusion

7

In this thesis, the influence of labelling methods (APR and APP) and speech styles (controlled and natural speech) on personality classification is explored. This is done using paralinguistic information from speech data. Experiments were conducted on three datasets that target three of the combinations. Furthermore, the influence of embeddings versus handcrafted features on personality classification was investigated. In this chapter, I will conclude my findings and answer the research questions.

RQ 1 *How well can speech-based personality classification be performed using datasets with different labelling methods (APR and APP) and speech styles (controlled and natural speech)?*

To answer this question, three sub questions were defined.

SQ 1.1 *For which traits can improvements be made on the challenge baseline and winner using the database for APP and natural speech (SPC)?*

For this experiment, the models were trained on the Speaker Personality Corpus, and compared to the baseline and winner of the Speaker Trait Challenge. The results showed that for Extraversion, Agreeableness, and Neuroticism, significantly higher UAR scores were obtained than both the baseline and the winner of the challenge. Except for Conscientiousness, all traits were best classified with models that were also used for the baseline. As the types of these models were the same, a plausible explanation for the increase in performance is that the features were more effective. Conscientiousness and Openness scored lower than the baseline results. The label distribution of Conscientiousness was considerably different from what was presented in the challenge, as discussed in Section 3.1.3. This could explain the lower results for this trait. While the models trained on Openness achieved worse results than the baseline and the challenge winner, this trait was also the hardest to classify in the challenge baseline and winner. A possible reason for the bad performance of Openness is that the trait is closely related to internally expressed behaviour, making it hard to be perceived. Overall, 3 out of 5 traits scored better than the challenge baseline and winner, and thus these traits could be improved.

SQ 1.2 *Which traits can be classified better than chance using the database for APR and controlled speech (NSC)?*

In this experiment, UAR scores significantly above chance were obtained for Extraversion, Agreeableness, Conscientiousness, and Neuroticism. Extraversion and Conscientiousness achieved the best scores during evaluation. Their cross validation scores during the development phase were considerably lower. However, the

standard deviation of the cross validation scores for these traits was also low, which is a sign of stability in the models when being exposed to new data. This suggests that the difference in the scores between the development and evaluation phase was not caused because of unstable models. One explanation is that the APR labels make the prevalence of traits in the test set highly dependent on which speakers are included. In conclusion, 4 out of 5 scores could be classified significantly better than chance, although additional experimentation with different test sets could provide more insights on the validity of the results.

SQ 1.3 *Which traits can be classified better than chance using the database for APR and natural speech (REMDE)?*
In this experiment, all traits except Conscientiousness achieved scores slightly, but significantly, above chance. However, the average UAR during the cross validation phase was often below chance. Additionally, the standard deviations of the cross validation were high, indicating that the models were unstable. It is therefore concluded that the models could not reliably classify the traits better than chance. These results suggest that it is hard to perform APR using audio that is recorded with natural speech.

Overall, the APR datasets performed worse than the APP dataset. This is in line with findings of previous research and could suggest that there are limitations to the extent to which APR can be performed. There were no clear patterns regarding the controlled and natural speech styles. Natural speech with APP achieved the highest scores while natural speech with APR achieved the lowest results. A dataset with the fourth combination, APP and controlled speech, could provide additional insights into the effectiveness of the labelling methods and speech styles. A standardized dataset that includes both labelling methods and both speech styles would enable detailed comparisons between results.

RQ 2 *How do the models perform on embeddings compared to hand-crafted features for speech-based personality classification?*
Table 4.9 showed that for all three experiments, the embeddings achieved better results than the eGeMAPS features on most traits. In total, the embeddings obtained higher scores 18 times, compared to 9 times for the eGeMAPS features. Therefore, the embeddings overall performed better than the hand-crafted features in speech-based personality classification. The pretrained model can successfully be used to extract features for personality classification.

This thesis provides the first research on both APR and APP including a speech style perspective. Two databases have been used for the first time for these tasks. In all experiments, significant UAR scores were achieved for multiple personality traits. The reliability of these scores differs per experiment. The APP dataset obtained better

results than the APR datasets. Further experimentation can clarify to what extent these results generalize. Additionally, embeddings have shown to be a valuable contribution to personality computing by outperforming hand-crafted features. The methodology presented in this thesis can be used as a foundation for future research on labelling methods and speech styles. As part of the methodology, a PyPi package was introduced that makes dividing datasets into multi-label stratified speaker-independent training and test sets considerably easier. The results and limitations can help determine the direction of future Personality Computing research.

Bibliography

[1]Gordon W. Allport and Henry S. Odbert. „Trait-names: A psycho-lexical study“. In: *Psychological Monographs* 47 (1936). Place: US Publisher: Psychological Review Company, pp. i–171 (cit. on p. 1).

[2]G. W. Allport. *Personality: a psychological interpretation*. Personality: a psychological interpretation. Pages: xiv, 588. Oxford, England: Holt, 1937 (cit. on p. 1).

[3]Donald W. Fiske. „Consistency of the factorial structures of personality ratings from different sources“. In: *The Journal of Abnormal and Social Psychology* 44 (1949). Place: US Publisher: American Psychological Association, pp. 329–344 (cit. on p. 1).

[4]Milton Rokeach. *The nature of human values*. The nature of human values. Pages: x, 438. New York, NY, US: Free Press, 1973 (cit. on p. 1).

[5]Aidan G.C. Wright. „Current Directions in Personality Science and the Potential for Advances through Computing“. In: *IEEE Transactions on Affective Computing* 5.3 (July 2014). Conference Name: IEEE Transactions on Affective Computing, pp. 292–296 (cit. on pp. 1, 2, 5, 9, 41, 42, 45).

[6]Benjamin P. Chapman, Brent Roberts, and Paul Duberstein. „Personality and longevity: knowns, unknowns, and implications for public health and personalized medicine“. eng. In: *Journal of Aging Research* 2011 (2011), p. 759170 (cit. on p. 1).

[7]Ian J. Deary, Alexander Weiss, and G. David Batty. „Intelligence and Personality as Predictors of Illness and Death: How Researchers in Differential Psychology and Chronic Disease Epidemiology Are Collaborating to Understand and Address Health Inequalities“. eng. In: *Psychological Science in the Public Interest: A Journal of the American Psychological Society* 11.2 (Aug. 2010), pp. 53–79 (cit. on pp. 1, 2).

[8]Roman Kotov, Wakiza Gamez, Frank Schmidt, and David Watson. „Linking "big" personality traits to anxiety, depressive, and substance use disorders: a meta-analysis“. eng. In: *Psychological Bulletin* 136.5 (Sept. 2010), pp. 768–821 (cit. on pp. 1, 2).

[9]D Samuel and T Widiger. „A meta-analytic review of the relationships between the five-factor model and DSM-IV-TR personality disorders: A facet level analysis“. en. In: *Clinical Psychology Review* 28.8 (Dec. 2008), pp. 1326–1342 (cit. on p. 1).

[10]Robert McCrae and Paul Costa. „The Structure of Interpersonal Traits: Wiggins's Circumplex and the Five-Factor Model“. In: *Journal of personality and social psychology* 56 (May 1989), pp. 586–95 (cit. on p. 1).

[11]Daniel J. Ozer and Verónica Benet-Martínez. „Personality and the prediction of consequential outcomes“. In: *Annual Review of Psychology* 57 (2006), pp. 401–421 (cit. on p. 1).

[12]Alessandro Vinciarelli and Gelareh Mohammadi. „A Survey of Personality Computing“. In: *IEEE Transactions on Affective Computing* 5.3 (July 2014). Conference Name: IEEE Transactions on Affective Computing, pp. 273–291 (cit. on pp. 1, 2).

[13]Robert P; Tett, Douglas N ; Jackson, and Mitchell Rothstein. „Personality Measures as Predictors of Job Performance: A Meta-Analytic Review“. In: *Personnel Psychology; Winter* 44 (1991) (cit. on pp. 1, 5).

[14]Michal Kosinski, David Stillwell, and Thore Graepel. „Private traits and attributes are predictable from digital records of human behavior“. In: *Proceedings of the National Academy of Sciences of the United States of America* 110.15 (Apr. 2013), pp. 5802–5805 (cit. on pp. 1, 5).

[15]Arthur E. Poropat. „A Meta-Analysis of the Five-Factor Model of Personality and Academic Performance“. In: *Psychological Bulletin* 135.2 (Mar. 2009), pp. 322–338 (cit. on pp. 1, 5).

[16]Kathryn E. Flynn and Maureen A. Smith. „Personality and Health Care Decision-Making Style“. In: *The Journals of Gerontology: Series B* 62.5 (Sept. 2007), P261–P267 (cit. on pp. 1, 2).

[17]Lawrence Egeren. „A Cybernetic Model of Global Personality Traits“. In: *Personality and social psychology review : an official journal of the Society for Personality and Social Psychology, Inc* 13 (May 2009), pp. 92–108 (cit. on p. 1).

[18]Caroline Davis, Karen Patte, Stacey Tweed, and Claire Curtis. „Personality traits associated with decision-making deficits“. en. In: *Personality and Individual Differences* 42.2 (Jan. 2007), pp. 279–290 (cit. on p. 1).

[19]Alessandro Bessi. „Personality traits and echo chambers on facebook“. en. In: *Computers in Human Behavior* 65 (Dec. 2016), pp. 319–324 (cit. on p. 1).

[20]David Funder. „Personality“. In: *Annu. Rev. Psychol.* 52 (2001), pp. 197–221 (cit. on p. 1).

[21]Raymond M. Bergner. „What is personality? Two myths and a definition“. en. In: *New Ideas in Psychology* 57 (Apr. 2020), p. 100759 (cit. on pp. 1, 2).

[22]Jess Feist and Gregory J. Feist. *Theories of Personality*. en. Google-Books-ID: IYLsAAAAMAAJ. McGraw-Hill, 2006 (cit. on p. 1).

[23]Randy Larsen and David Buss. *Personality Psychology: Domains of Knowledge about Human Nature*. Jan. 2005 (cit. on p. 1).

[24]Walter Mischel, Yuichi Shoda, and Ozlem Ayduk. *Introduction to Personality: Toward an Integrative Science of the Person*. en. Google-Books-ID: 0bYWDgAAQBAJ. John Wiley & Sons, Sept. 2007 (cit. on p. 1).

[25]Le Vy Phan and John F. Rauthmann. „Personality computing: New frontiers in personality assessment“. In: *Social and Personality Psychology Compass* 15.7 (July 2021). Publisher: John Wiley and Sons Inc (cit. on p. 1).

[26]Oliver P John and Sanjay Srivastava. „The Big-Five Trait Taxonomy: History, Measurement, and Theoretical Perspectives“. In: (1999) (cit. on p. 1).

[27]Robert R Mccrae and Paul T Costa. „Personality Trait Structure as a Human Universal“. In: *Am Psychol.* (1997), pp. 509–516 (cit. on p. 1).

[28]Björn Schuller, Stefan Steidl, Anton Batliner, et al. „The INTERSPEECH 2012 speaker trait challenge“. In: *13th Annual Conference of the International Speech Communication Association 2012, INTERSPEECH 2012*. Vol. 1. 2012, pp. 254–257 (cit. on pp. 2–5, 9, 11, 13, 15, 16, 20, 25, 27, 28, 33, 41).

[29]Guozhen An, Sarah Ita Levitan, Rivka Levitan, et al. „Automatically classifying self-rated personality scores from speech“. In: *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*. Vol. 08-12-September-2016. ISSN: 19909772. International Speech and Communication Association, 2016, pp. 1412–1416 (cit. on pp. 2, 5, 10, 11, 14, 20, 28, 41, 47).

[30]Alessandro Vinciarelli and Gelareh Mohammadi. „More Personality in Personality Computing“. In: *IEEE Transactions on Affective Computing* 5.3 (July 2014). Conference Name: IEEE Transactions on Affective Computing, pp. 297–300 (cit. on pp. 2, 5, 9, 11).

[31]Björn Schuller, Stefan Steidl, Anton Batliner, et al. „A Survey on perceived speaker traits: Personality, likability, pathology, and the first challenge“. In: *Computer Speech and Language* 29.1 (2015). Publisher: Academic Press, pp. 100–131 (cit. on pp. 2, 5, 9, 10, 20).

[32]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“. In: *Journal of Research in Personality* 41.1 (Feb. 2007), pp. 203–212 (cit. on pp. 2, 13).

[33]Atieh Poushneh. „Humanizing voice assistant: The impact of voice assistant personality on consumers' attitudes and behaviors“. en. In: *Journal of Retailing and Consumer Services* 58 (Jan. 2021), p. 102283 (cit. on p. 2).

[34]James Uleman, S Saribay, and Celia Gonzalez. „Spontaneous Inferences, Implicit Impressions, and Implicit Theories“. In: *Annual review of psychology* 59 (Feb. 2008), pp. 329–60 (cit. on p. 2).

[35]Alessandro Nai and Jürgen Maier. „Perceived personality and campaign style of Hillary Clinton and Donald Trump“. en. In: *Personality and Individual Differences* 121 (Jan. 2018), pp. 80–83 (cit. on p. 2).

[36]Scott O. Lilienfeld, Irwin D. Waldman, Kristin Landfield, et al. „Fearless dominance and the U.S. presidency: Implications of psychopathic personality traits for successful and unsuccessful political leadership.“ en. In: *Journal of Personality and Social Psychology* 103.3 (2012), pp. 489–505 (cit. on p. 2).

[37]Tim Polzehl, Katrin Schoenenberg, Sebastian Moller, et al. „On Speaker-Independent Personality Perception and Prediction from Speech“. en. In: (2012) (cit. on p. 3).

[38]Laura Fernández Gallardo and Benjamin Weiss. „The Nautilus Speaker Characterization Corpus: Speech Recordings and Labels of Speaker Characteristics and Voice Descriptions“. In: (2017) (cit. on pp. 3, 4, 20, 21).

[39]*IEMOCAP- Home* (cit. on p. 3).

[40]A Batliner, S Steidl, and E Noth. „Releasing a thoroughly annotated and processed spontaneous emotional database: the FAU Aibo Emotion Corpus“. en. In: (2008) (cit. on p. 3).

[41]Felix Burkhardt, Astrid Paeschke, M. Rolfes, Walter Sendlmeier, and Benjamin Weiss. „A database of German emotional speech“. In: vol. 5. Sept. 2005, pp. 1517–1520 (cit. on p. 3).

[42]T. Vogt and E. Andre. „Comparing Feature Sets for Acted and Spontaneous Speech in View of Automatic Emotion Recognition“. en. In: *2005 IEEE International Conference on Multimedia and Expo*. Amsterdam, The Netherlands: IEEE, 2005, pp. 474–477 (cit. on pp. 3–5, 12).

[43]A. Batliner, K. Fischer, R. Huber, J. Spilker, and E. Nöth. „How to find trouble in communication“. en. In: *Speech Communication* 40.1-2 (Apr. 2003), pp. 117–143 (cit. on pp. 3–5, 12).

[44]Houwei Cao, Ragini Verma, and Ani Nenkova. „Speaker-sensitive emotion recognition via ranking: Studies on acted and spontaneous speech“. en. In: *Computer Speech & Language* 29.1 (Jan. 2015), pp. 186–202 (cit. on pp. 3–5, 12).

[45]Farah Chenchah and Zied Lachiri. „Speech Emotion Recognition in Acted and Spontaneous Context“. en. In: *Procedia Computer Science*. The 6th international conference on Intelligent Human Computer Interaction, IHCI 2014 39 (Jan. 2014), pp. 139–145 (cit. on p. 3).

[46]Leimin Tian, Johanna D. Moore, and Catherine Lai. „Emotion recognition in spontaneous and acted dialogues“. en. In: *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*. Xi'an, China: IEEE, Sept. 2015, pp. 698–704 (cit. on p. 3).

[47]Johannes Wagner, Thurid Vogt, and Elisabeth André. „A Systematic Comparison of Different HMM Designs for Emotion Recognition from Acted and Spontaneous Speech“. en. In: *Affective Computing and Intelligent Interaction*. Ed. by Ana C. R. Paiva, Rui Prada, and Rosalind W. Picard. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2007, pp. 114–125 (cit. on p. 3).

[48]Daniel Küstner, Raquel Tato, Thomas Kemp, and Beate Meffert. „Towards Real Life Applications in Emotion Recognition“. en. In: *Affective Dialogue Systems*. Ed. by Elisabeth André, Laila Dybkjær, Wolfgang Minker, and Paul Heisterkamp. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2004, pp. 25–35 (cit. on pp. 3–5, 12).

[49]A Batliner, K Fischer, R Huber, J Spilker, and E Noth. „Desperately Seeking Emotions Or: Actors, Wizards, and Human Beings“. en. In: (2000) (cit. on pp. 4, 5).

[50]Alberto del. Bimbo, Shih-Fu Chang, ACM Special Interest Group on Multimedia., and Association for Computing Machinery. „openSMILE – The Munich Versatile and Fast Open-Source Audio Feature Extractor“. In: (2010). Publisher: Association for Computing Machinery ISBN: 9781605589336, p. 1849 (cit. on pp. 5, 21).

[51]Florian Eyben, Klaus R. Scherer, Bjorn W. Schuller, et al. „The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing“. In: *IEEE Transactions on Affective Computing* 7.2 (Apr. 2016). Publisher: Institute of Electrical and Electronics Engineers Inc., pp. 190–202 (cit. on pp. 5, 21, 41).

[52]Joel Shor, Aren Jansen, Ronnie Maor, et al. „Towards Learning a Universal Non-Semantic Representation of Speech“. In: *Interspeech 2020*. arXiv:2002.12764 [cs, eess, stat]. Oct. 2020, pp. 140–144 (cit. on pp. 5, 22, 47).

[53]Joel Shor and Subhashini Venugopalan. „TRILLsson: Distilled Universal Paralinguistic Speech Representations“. In: *Interspeech 2022*. arXiv:2203.00236 [cs, eess]. Sept. 2022, pp. 356–360 (cit. on pp. 5, 21, 22, 41).

[54]Jacob Peplinski, Joel Shor, Sachin Joglekar, Jake Garrison, and Shwetak Patel. „FRILL: A Non-Semantic Speech Embedding for Mobile Devices“. In: *Interspeech 2021*. arXiv:2011.04609 [cs, eess]. Aug. 2021, pp. 1204–1208 (cit. on pp. 5, 22).

[55]Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. „wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations“. In: *Advances in Neural Information Processing Systems*. Vol. 33. Curran Associates, Inc., 2020, pp. 12449–12460 (cit. on pp. 5, 22).

[56]Houwei Cao, David G. Cooper, Michael K. Keutmann, et al. „CREMA-D: Crowd-sourced Emotional Multimodal Actors Dataset“. In: *IEEE transactions on affective computing* 5.4 (2014), pp. 377–390 (cit. on pp. 5, 22).

[57]Philip Jackson and Sana ul haq. *Surrey Audio-Visual Expressed Emotion (SAVEE) database*. Apr. 2011 (cit. on p. 5).

[58]Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. „VoxCeleb: a large-scale speaker identification dataset“. In: *Interspeech 2017*. arXiv:1706.08612 [cs]. Aug. 2017, pp. 2616–2620 (cit. on pp. 5, 22).

[59]Ken MacLean. *VoxForge*. 2018 (cit. on pp. 5, 22).

[60]Pete Warden. *Speech Commands: A Dataset for Limited-Vocabulary Speech Recognition*. arXiv:1804.03209 [cs]. Apr. 2018 (cit. on pp. 5, 22).

[61]Sweta Karlekar, Tong Niu, and Mohit Bansal. *Detecting Linguistic Characteristics of Alzheimer's Dementia by Interpreting Neural Models*. arXiv:1804.06440 [cs]. Apr. 2018 (cit. on p. 5).

[62]Alexei V. Ivanov and Xin Chen. „Modulation spectrum analysis for speaker personality trait recognition“. In: *13th Annual Conference of the International Speech Communication Association 2012, INTERSPEECH 2012*. Vol. 1. 2012, pp. 278–281 (cit. on pp. 9, 20, 28).

[63]Kartik Audhkhasi, Angeliki Metallinou, Ming Li, and Shrikanth Narayanan. „Speaker Personality Classification Using Systems Based on Acoustic-Lexical Cues and an Optimal Tree-Structured Bayesian Network“. In: vol. 1. Sept. 2012 (cit. on p. 9).

[64]Yazid Attabi and Pierre Dumouchel. „Anchor models and WCCN normalization for speaker trait classification“. In: *13th Annual Conference of the International Speech Communication Association 2012, INTERSPEECH 2012* 1 (Jan. 2012), pp. 522–525 (cit. on p. 9).

[65]Claude Montacié and Marie José Caraty. „Pitch and intonation contribution to speakers' traits classification“. In: *13th Annual Conference of the International Speech Communication Association 2012, INTERSPEECH 2012*. Vol. 1. 2012, pp. 526–529 (cit. on pp. 9, 20).

[66]Gopala Krishna Anumanchipalli, Hugo Meinedo, Miguel Bugalho, et al. „Text-dependent pathological voice detection“. In: *13th Annual Conference of the International Speech Communication Association 2012, INTERSPEECH 2012*. Vol. 1. 2012, pp. 530–533 (cit. on p. 9).

[67]Jouni Pohjalainen, Serdar Kadioglu, and Okko Räsänen. „Feature Selection for Speaker Traits“. In: (2012) (cit. on p. 9).

[68]Johannes Wagner, Florian Lingenfelser, and Elisabeth Andre. „A Frame Pruning Approach for Paralinguistic Recognition Tasks“. In: *13th Annual Conference of the International Speech Communication Association 2012, INTERSPEECH 2012* 1 (Jan. 2012) (cit. on p. 9).

[69]Clément Chastagnol and Laurence Devillers. „Personality traits detection using a parallelized modified SFFS algorithm“. In: *13th Annual Conference of the International Speech Communication Association 2012, INTERSPEECH 2012* 1 (Jan. 2012) (cit. on p. 9).

[70]Harm Buisman and E. Postma. „The log-Gabor method: speech classification using spectrogram image analysis“. In: 2012 (cit. on p. 9).

[71]D. Wu. „Genetic algorithm based feature selection for speaker trait classification“. In: 1 (Jan. 2012), pp. 294–297 (cit. on p. 9).

[72]Michelle Sanchez, Aaron Lawson, Dimitra Vergyri, and Harry Bratt. „Multi-System Fusion of Extended Context Prosodic and Cepstral Features for Paralinguistic Speaker Trait Classification“. In: *13th Annual Conference of the International Speech Communication Association 2012, INTERSPEECH 2012* 1 (Jan. 2012) (cit. on p. 9).

[73]Yoav Freund and Robert E Schapire. „A Short Introduction to Boosting“. en. In: () (cit. on p. 9).

[74]Marc-André Carbonneau, Eric Granger, Yazid Attabi, and Ghyslain Gagnon. „Feature Learning from Spectrograms for Assessment of Personality Traits“. In: *IEEE Transactions on Affective Computing* 11.1 (Jan. 2020). Conference Name: IEEE Transactions on Affective Computing, pp. 25–31 (cit. on p. 9).

[75]Effat Jalaeian Zaferani, Mohammad Teshnehlab, and Mansour Vali. „Automatic Personality Traits Perception Using Asymmetric Auto-Encoder“. In: *IEEE Access* 9 (2021). Conference Name: IEEE Access, pp. 68595–68608 (cit. on p. 9).

[76]Firoj Alam and Giuseppe Riccardi. „Fusion of acoustic, linguistic and psycholinguistic features for Speaker Personality Traits recognition“. In: *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. ISSN: 2379-190X. May 2014, pp. 955–959 (cit. on p. 9).

[77]Francois Mairesse, Marilyn Walker, Matthias Mehl, and Roger Moore. „Using Linguistic Cues for the Automatic Recognition of Personality in Conversation and Text“. In: *J. Artif. Intell. Res. (JAIR)* 30 (Sept. 2007), pp. 457–500 (cit. on p. 10).

[78]Gloria Zen, Bruno Lepri, Elisa Ricci, and Oswald Lanz. „Space speaks: towards socially and personality aware visual surveillance“. In: (Oct. 2010) (cit. on p. 10).

[79]Hung-Yue Suen, Kuo-En Hung, and Chien-Liang Lin. „TensorFlow-Based Automatic Personality Recognition Used in Asynchronous Video Interviews“. In: *IEEE Access* 7 (2019). Conference Name: IEEE Access, pp. 61018–61023 (cit. on pp. 10, 48).

[80]Ligia Maria Batrinca, Nadia Mana, Bruno Lepri, Fabio Pianesi, and Nicu Sebe. „Please, tell me about yourself: automatic personality assessment using short self-presentations“. en. In: *Proceedings of the 13th international conference on multimodal interfaces*. Alicante Spain: ACM, Nov. 2011, pp. 255–262 (cit. on p. 10).

[81]Alexei Ivanov, Giuseppe Riccardi, Adam Sporka, and Jakub Franc. „Recognition of Personality Traits from Human Spoken Conversations.“ In: Aug. 2011, pp. 1549–1552 (cit. on p. 11).

[82]Robert E. Schapire and Yoram Singer. „BoosTexter: A Boosting-based System for Text Categorization“. en. In: *Machine Learning* 39.2 (May 2000), pp. 135–168 (cit. on p. 11).

[83]Maria Koutsombogera, Parth Sarthy, and Carl Vogel. „Acoustic Features in Dialogue Dominate Accurate Personality Trait Classification“. In: *2020 IEEE International Conference on Human-Machine Systems (ICHMS)*. Sept. 2020, pp. 1–3 (cit. on p. 13).

[84]Mohsen Fallahnezhad, Mansour Vali, and Mehdi Khalili. „Automatic Personality Recognition from reading text speech“. In: *2017 Iranian Conference on Electrical Engineering (ICEE)*. May 2017, pp. 18–23 (cit. on p. 13).

[85]Shashank Jaiswal, Siyang Song, and Michel Valstar. „Automatic prediction of Depression and Anxiety from behaviour and personality attributes“. In: *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*. ISSN: 2156-8111. Sept. 2019, pp. 1–7 (cit. on p. 13).

[86]Gelareh Mohammadi and Alessandro Vinciarelli. „Automatic Personality Perception: Prediction of Trait Attribution Based on Prosodic Features“. In: *IEEE Transactions on Affective Computing* 3.3 (July 2012). Conference Name: IEEE Transactions on Affective Computing, pp. 273–284 (cit. on p. 13).

[87]Paul Costa and R. Mccrae. „Neo PI-R professional manual“. In: *Psychological Assessment Resources* 396 (Jan. 1992) (cit. on p. 13).

[88]Eddy Brixen. „Spectral degradation of speech captured by miniature microphones mounted on persons' heads and chests“. In: May 1996 (cit. on p. 18).

[89]Tim Sainburg, Marvin Thielk, and Timothy Q. Gentner. „Finding, visualizing, and quantifying latent structure across diverse animal vocal repertoires“. en. In: *PLOS Computational Biology* 16.10 (Oct. 2020). Ed. by Frédéric E. Theunissen, e1008228 (cit. on p. 19).

[90]Laura Fernández Gallardo. „Nautilus Speaker Characterization (NSC) Corpus“. In: (2017) (cit. on p. 20).

[91]Björn Schuller, Stefan Steidl, Anton Batliner, et al. „The INTERSPEECH 2016 Computational Paralinguistics Challenge: Deception, Sincerity & Native Language“. In: 2016 (cit. on p. 22).

[92]Leonardo Pepino, Pablo Riera, and Luciana Ferrer. *Emotion Recognition from Speech Using Wav2vec 2.0 Embeddings*. arXiv:2104.03502 [cs, eess]. Apr. 2021 (cit. on p. 22).

[93]Joel Shor, Aren Jansen, Wei Han, Daniel Park, and Yu Zhang. „Universal Paralinguistic Speech Representations Using Self-Supervised Conformers“. In: *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. arXiv:2110.04621 [cs, eess]. May 2022, pp. 3169–3173 (cit. on p. 22).

[94]Konstantinos Sechidis, Grigorios Tsoumakas, and Ioannis Vlahavas. „On the Stratification of Multi-Label Data“. In: (2011). ISBN: 16105158060.98 (cit. on p. 22).

[95]John Shawe-Taylor and Nello Cristianini. „Kernel Methods for Pattern Analysis“. en. In: () (cit. on p. 26).

[96]Zhongliang Li, Rachid Outbib, Stefan Giurgea, et al. „Online implementation of SVM based fault diagnosis strategy for PEMFC systems“. In: Feb. 2015 (cit. on p. 26).

[97]Leo Breiman. „Random Forests“. en. In: *Machine Learning* 45.1 (Oct. 2001), pp. 5–32 (cit. on p. 27).

[98]T. Cover and P. Hart. „Nearest neighbor pattern classification“. In: *IEEE Transactions on Information Theory* 13.1 (Jan. 1967). Conference Name: IEEE Transactions on Information Theory, pp. 21–27 (cit. on p. 28).