Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

School of Computer Science and Statistics

Analysis of Nonverbal Qualities in Dialogue to Improve Turn-taking of Spoken Dialogue Systems

Oliver Kraus

August 19, 2022

A dissertation submitted in partial fulfilment
of the requirements for the degree of
MSc Computer Science - Intelligent Systems

Declaration

I hereby declare that this dissertation is entirely my own work and that it has not been submitted as an exercise for a degree at this or any other university.

I have read and I understand the plagiarism provisions in the General Regulations of the University Calendar for the current year, found at `http://www.tcd.ie/calendar`.

I have also completed the Online Tutorial on avoiding plagiarism 'Ready Steady Write', located at `http://tcd-ie.libguides.com/plagiarism/ready-steady-write`.

Signed: ______________________ Date: ______________________

Abstract

Speech is one of the most important means of communication for humans. With computer systems becoming more and more part of our daily lives, it therefore seems reasonable to develop computer programs that can communicate with human users via speech. These programs are known as spoken dialogue systems. A central part of human dialogue is the organisation of speech into alternating turns. Previous research has shown that humans are highly skilled in anticipating the end of the previous turn to prepare a timely response and keep the conversation fluid. Replicating this behavior is thus also important for enabling natural, fluent and responsive interaction with spoken dialogue systems.

This dissertation puts the focus on the analysis of nonverbal qualities in dialogue to help bridge this gap. In particular, the influence of speaker personality types on nonverbal qualities is examined. Results of this analysis could help to adapt turn-taking models to the personality type of their users to simulate more natural conversation. Furthermore, in order to examine the predictive power of nonverbal qualities on turn-taking, a logistic regression model is presented that makes continuous turn-taking decisions based on nonverbal qualities. To achieve this, a multimodal dataset of spoken English task-based dialogues is utilised. Next to annotations of the turns and other nonverbal qualities such as gaze and laughter, the dataset also contains personality scores for the participants based on the Big Five model. This dissertation finds evidence that the openness trait of the Big Five model, influences the total time spoken in a dialogue, the average time between turns and the amount of times a person gets interrupted during dialogue. There is also evidence that the extroversion factor influences the amount of gaps in speech a person leaves during dialogue. The presented continuous turn-taking model does not outperform a last-known value baseline.

Acknowledgements

I would like to thank Dr. Carl Vogel for his help and guidance.

Contents

List of Figures

List of Tables

1 Introduction

Speech is arguably one of the most important and most commonly used form of communication for humans. Furthermore, it is fast and provides excellent error resilience since potential misunderstandings can be resolved immediately. With intelligent machines and computer systems becoming increasingly part of our daily lives, it seems sensible to develop systems that allow human users to operate them via speech.

Such computer programs are known as spoken dialogue systems and can replace more established user interfaces such as a mouse and keyboard input. Next to being a potentially more natural interface than a mouse and keyboard input, spoken dialogue systems also provide remote control in situations where the user's hands may be occupied. When spoken dialogue systems first emerged, one of the main challenges was understanding the user's speech input. However, with the rise of deep learning techniques, these problems have been widely solved, putting the focus on more engaging, interactive and believable interactions with spoken dialogue systems. Over the last years, especially Intelligent Personal Voice Assistants such as Amazon's Alexa or Google Assistant have become popular. However, research suggests that these systems are mainly used for simple tasks such as asking for the weather forecast. This indicates that modern spoken dialogue systems fail to fully realise the potential of speech as a form of communication.

In order to overcome these one-turn interactions with spoken dialogue systems and provide truly responsive, interactive and natural feeling conversations, it is essential to replicate human-level turn-taking in spoken dialogue systems. Turn-taking describes the organisation of dialogue into alternating turns and is characterised by humans trying to minimise gaps and overlaps between turns. On the one hand, the previous speaker should not be interrupted. On the other hand, gaps between speech should be short to keep the conversation fluid. To achieve these dynamics, humans process a wide range of cues that help them anticipate the ending of the previous speaker's turn.

Analysing these cues can be a key in replicating human-level turn-taking in spoken dialogue systems. In this thesis, the focus is put on the analysis of nonverbal qualities in dialogue that can help a spoken dialogue system to predict turn-taking dynamics. This means that signals such as the distribution of speaker activity, gaze and laughter are analysed instead of the spoken content of the dialogue. These signals have the advantage that they are applicable irrespective of the spoken language, are faster to process and less prone to errors during signal processing since no natural language understanding is required. Therefore, the central question of the thesis is to what extent the processing of nonverbal signals in dialogue can support spoken dialogue systems in turn-taking decisions.

1.1 Thesis Structure

The remaining parts of the thesis are structured as follows. First, chapter 2 provides an overview of the related literature required for understanding the dissertation. Based on this literature review, the research questions are then derived in the summary of chapter 2. After this, chapter 3 gives an overview of the research methods used in the thesis to answer the posed research questions. In chapter 4, the experiments conducted to answer the research questions are described. Next, chapter 5 presents the results of the experiments. Finally, chapter 6 summarises the dissertation and provides an outlook on possible future work.

2 Related Work

In this chapter, literature that is relevant to the thesis is discussed. First, literature about turn-taking in human dialogue is presented. Then, the following sections describe literature about spoken dialogue systems, the Big Five personality traits, nonverbal analysis of dialogue and the role of laughter in dialogue. Finally, the research question is deduced from the literature review.

2.1 Turn Taking in Human Dialogue

Turn-taking in human dialogue describes the organization of dialogue into alternating turns of the participants and is a complex cognitive task. Participants must process what is being said while preparing their own responses in a short amount of time. It is desirable for participants to keep only short gaps between utterances to keep the conversation fluid. On the other hand, participants in a dialogue have to avoid speaking too early to not interrupt the previous speaker. Research suggests that gaps between turns in human dialogue typically last about 0-250 ms [Heldner and Edlund, 2010] [Stivers et al., 2009]. However, the magnitude of this gap is influenced by factors like the nature of the dialogue and cognitive load. For instance, [Trimboli and Walker, 1984] found that gaps in competitive dialogues like a debate are shorter than in friendly dialogues. These dynamics are stable across languages and cultures, which is remarkable since different languages vary significantly in terms of syntax [Weilhammer and Rabold, 2003] [Stivers et al., 2009]. Depending on the language, a verb, which can make up a large part of the meaning of a sentence, can also come very late in a sentence leaving less time to react.

Moreover, there is extensive research that has quantified latency in speech production. For example, [Indefrey and Levelt, 2004] found that the delay between seeing a picture and forming a single word to describe the picture is in the magnitude of 600 ms. Furthermore, [Griffin and Bock, 2000] quantified the delay to form a short sentence at 1200 ms. This means that the latency for producing even a single word is higher than the typical gap in human dialogue. This raises the question of how humans are able to hold conversations with minimal gaps between turns. In the following, we look at the two main theories that aim to explain this phenomenon.

2.1.1 Projection Theory

[Sacks et al., 1974] defined a turn as a linguistic construct that is built of "turn-constructional units" (e.g., a word or a phrase). When a turn-constructional unit is finished (called turn-completion point in the paper), a turn change can occur with a gap between the speakers, an overlap between the speakers, or without a gap and without an overlap between the speakers. From their data analysis, they find that the most frequent way of changing turns is with no gap and no overlap. Following this, they proposed the idea that humans can anticipate the end of a turn, which allows them to prepare a timely response. The paper, however, does not further explain the mechanism behind this, but they assumed that syntax, semantics and intonation allow humans to anticipate the ending of a turn.

2.1.2 Signalling Theory

On the other hand, it was proposed that participants in a dialogue send out signals when they want to end a turn [Duncan, 1972]. The difference with the theory by [Sacks et al., 1974] is that humans react to cues they receive rather than anticipate them. [Duncan, 1972] mentions the following cues that indicate the end of a turn:

- Rising or falling intonation at the end of a phrase
- Drawl on the final syllable or on the stressed syllable of a terminal clause
- Termination of hand gesticulation
- Backchannels such as "mmh" or "you know"
- Falling pitch and loudness combined with a socio-centric sequence
- Completion of a grammatical clause

In addition, [Duncan and Niederehe, 1974] found that the more clues there are, the more likely it is that a turn will be completed.

2.1.3 Discussion

There are several studies that discuss and build upon these two theories. A common argument against the signalling theory is that the cues mentioned occur too late in a turn, so participants in a dialogue cannot react to them in time. [Levinson and Torreira, 2015] estimated that a response would occur approximately after 600-1500 ms if humans were responding to cues. In contrast, [Heldner and Edlund, 2010] claimed that gaps larger than 250 ms could possibly be explained by the response to turn-taking cues. Overall, it seems difficult to reject the theory on the basis of timing alone since short intervals between turns could also be explained by responses to "false alarms".

Meanwhile, the main criticism of the projection theory is that [Sacks et al., 1974] does not specify the mechanisms that lead to predicting the end of a turn. This has been further investigated by the following studies. [De Ruiter et al., 2006] had conducted an experiment in which subjects listened to a conversation and were asked to press a button when they thought a turn was about to end. Participants were very accurate in this task, with reaction times of less than 200 ms, suggesting that they were anticipating the end of a turn rather than reacting to a signal. In a further experiment, subjects were listening to a filtered version of the conversations with no intonation and were still able to predict the endings of the turns with a high degree of accuracy. In a final experiment, participants heard a version of the conversations with intonation but with a filter so that the spoken words could no longer be identified. In this version of the experiment, the participants' accuracy decreased significantly, and the authors concluded that humans rely primarily on syntactic and semantic information to anticipate turn endings. These findings were reproduced and backed by other studies like [Gambi et al., 2015].

In addition to syntactic and semantic information, several studies investigated the impact of gaze and gestures of participants on projections of turns. For instance, [Kawahara et al., 2012] analyzed multiparty conversations in which one person held a scientific presentation to an audience of two. The authors tried to predict the intervention of the audience with a logistic regression classifier based on prosody and gaze features. They found that a combination of both feature sets yielded the best results. In particular, they suggested that the presenter is more likely to gaze at the person in the audience before that person starts speaking. Similarly, [Mutlu et al., 2012] found following dynamics in three-party dialogue. First, their analysis shows that the current speaker is likely to look at the person to whom they yield the turn, which is consistent with the results from [Kawahara et al., 2012]. Furthermore, they noted that the person taking the floor is likely to look at the current speaker near the end of their turn. Lastly, they noted that the current speaker gazes away from the other participants in the conversation to signal that they have no intention of finishing their turn.

Finally, [Riest et al., 2015] [Heldner and Edlund, 2010] conclude that turn-taking in human dialogue is primarily based on projection while signalling serves as a backup.

2.2 Dialogue Systems

Dialogue systems are computer programs that can simulate a conversation with humans either through text or speech. They can be used to provide user support, drive up user engagement, collect user information or help users to execute a certain action. Furthermore, dialogue systems find use in many domains, including healthcare, education, commerce and daily life. Typically they contain a natural language understanding module, a dialogue state tracker and a natural language generation module. Especially the use of deep learning techniques has advanced these technologies significantly in the past years [Motger et al., 2022]. Two research areas relevant to the dissertation are examined in more detail in the following. First, we discuss state of the art in spoken dialogue systems and then look at the prediction of turn endings in spoken dialogue systems.

2.2.1 State of the Art in Spoken Dialogue Systems

Spoken dialogue systems can communicate with users via speech. The main difference to text-based systems is that they include a speech recognition and text-to-speech modules. Spoken dialogue as a means of communication offers many advantages. It is fast, flexible and offers resilience to error handling and constant validation. In addition, the use of spoken dialogue systems provides a hands-free and eyes-free interface that can be useful for people with disabilities or generally in environments where the user's hands and eyes are engaged, such as in a car [Edlund et al., 2008]. However, contemporary dialogue systems fail to fully utilise these advantages. For example, if the user does not follow a predefined behaviour, the dialogue system may respond with unsatisfactory replies. In fact, research shows that spoken dialogue systems are primarily used for simple tasks, such as asking for the weather or directions [Dubiel et al., 2018].

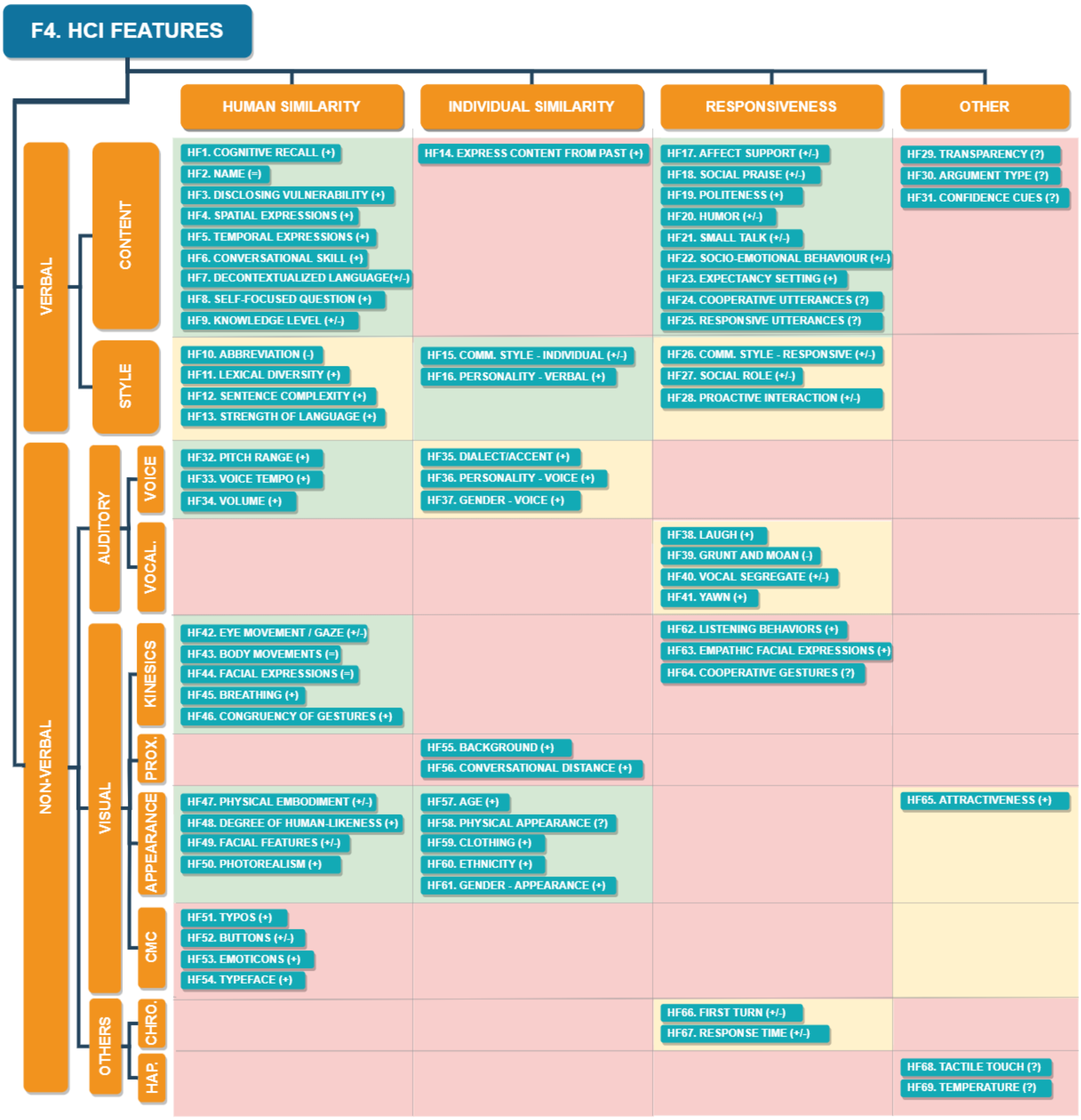
Therefore, it seems reasonable to aim for a more *natural* interaction with humans when developing spoken dialogue systems to realise these systems' full potential. However, this raises the question of what more natural exactly means in this context.

[Edlund et al., 2008] suggested that users perceive dialogue systems metaphorically rather than as conversational partners. For instance, there is the "interface" metaphor and the "android" metaphor. The user's expectations vary depending on the perceived metaphor, and design decisions of the system have to be made accordingly. For example, a system that is presented as an anthropomorphic agent would be expected to have higher communication capabilities than a system that is presented as an interface for a simple task, such as asking about the weather. By clearly communicating the chosen metaphor to the user with design choices, interaction with the dialogue system can be made more natural.

[Dautenhahn et al., 2005] conducted a survey in which participants were asked about their attitudes toward interaction with a hypothetical robot companion and found that people prefer human-like communication in an ideal situation. Achieving human-like communication is an enormous task and requires breakthroughs in multiple research areas, such as knowledge and memory modelling, cognitive AI, and symbolic reasoning [Shum et al., 2018].

In order to focus on a more immediate goal, [Motger et al., 2022] used the term of *user-perceived quality* and argues that this correlates with human-like communication. Figure 2.1 shows an extensive overview of features that impact the user-perceived quality of spoken dialogue systems. In the following, we take a closer look at three open areas of research to improve the user-perceived-quality of exchanges with spoken dialogue systems [Lison and Meena, 2014] [Ward and DeVault, 2017] [de Barcelos Silva et al., 2020].

Figure 2.1: Overview of features that impact the perceived quality of dialogue systems. The heat map shows how many research papers deal with the respective area (green means more and red stands for less). The influence of the feature on the quality is indicated in brackets after the respective feature. A "+" indicates a positive impact, a "-" indicates a negative impact, "=" stands for neutral impact, "+/-" means that it depends on the domain of the dialogue system and "?" means that there is not enough literature for a clear judgement. Source: [Motger et al., 2022]


F4. HCI FEATURES
HUMAN SIMILARITY
INDIVIDUAL SIMILARITY
RESPONSIVENESS
OTHER
VERBAL
CONTENT
STYLE
NON-VERBAL
AUDITORY
VOICE
VOCAL.
VISUAL
KINESICS
PROX.
APPEARANCE
CMC
OTHERS
CHRO.
HAP.
HF1. COGNITIVE RECALL (+)
HF2. NAME (=)
HF3. DISCLOSING VULNERABILITY (+)
HF4. SPATIAL EXPRESSIONS (+)
HF5. TEMPORAL EXPRESSIONS (+)
HF6. CONVERSATIONAL SKILL (+)
HF7. DECONTEXTUALIZED LANGUAGE(+/-)
HF8. SELF-FOCUSED QUESTION (+)
HF9. KNOWLEDGE LEVEL (+/-)
HF14. EXPRESS CONTENT FROM PAST (+)
HF17. AFFECT SUPPORT (+/-)
HF18. SOCIAL PRAISE (+/-)
HF19. POLITENESS (+)
HF20. HUMOR (+/-)
HF21. SMALL TALK (+/-)
HF22. SOCIO-EMOTIONAL BEHAVIOUR (+/-)
HF23. EXPECTANCY SETTING (+)
HF24. COOPERATIVE UTTERANCES (?)
HF25. RESPONSIVE UTTERANCES (?)
HF29. TRANSPARENCY (?)
HF30. ARGUMENT TYPE (?)
HF31. CONFIDENCE CUES (?)
HF10. ABBREVIATION (-)
HF11. LEXICAL DIVERSITY (+)
HF12. SENTENCE COMPLEXITY (+)
HF13. STRENGTH OF LANGUAGE (+)
HF15. COMM. STYLE - INDIVIDUAL (+/-)
HF16. PERSONALITY - VERBAL (+)
HF26. COMM. STYLE - RESPONSIVE (+/-)
HF27. SOCIAL ROLE (+/-)
HF28. PROACTIVE INTERACTION (+/-)
HF32. PITCH RANGE (+)
HF33. VOICE TEMPO (+)
HF34. VOLUME (+)
HF35. DIALECT/ACCENT (+)
HF36. PERSONALITY - VOICE (+)
HF37. GENDER - VOICE (+)
HF38. LAUGH (+)
HF39. GRUNT AND MOAN (-)
HF40. VOCAL SEGREGATE (+/-)
HF41. YAWN (+)
HF42. EYE MOVEMENT / GAZE (+/-)
HF43. BODY MOVEMENTS (=)
HF44. FACIAL EXPRESSIONS (=)
HF45. BREATHING (+)
HF46. CONGRUENCY OF GESTURES (+)
HF62. LISTENING BEHAVIORS (+)
HF63. EMPATHIC FACIAL EXPRESSIONS (+)
HF64. COOPERATIVE GESTURES (?)
HF55. BACKGROUND (+)
HF56. CONVERSATIONAL DISTANCE (+)
HF47. PHYSICAL EMBODIMENT (+/-)
HF48. DEGREE OF HUMAN-LIKENESS (+)
HF49. FACIAL FEATURES (+/-)
HF50. PHOTOREALISM (+)
HF57. AGE (+)
HF58. PHYSICAL APPEARANCE (?)
HF59. CLOTHING (+)
HF60. ETHNICITY (+)
HF61. GENDER - APPEARANCE (+)
HF65. ATTRACTIVENESS (+)
HF51. TYPOS (+)
HF52. BUTTONS (+/-)
HF53. EMOTICONS (+)
HF54. TYPEFACE (+)
HF66. FIRST TURN (+/-)
HF67. RESPONSE TIME (+/-)
HF68. TACTILE TOUCH (?)
HF69. TEMPERATURE (?)

Incrementality

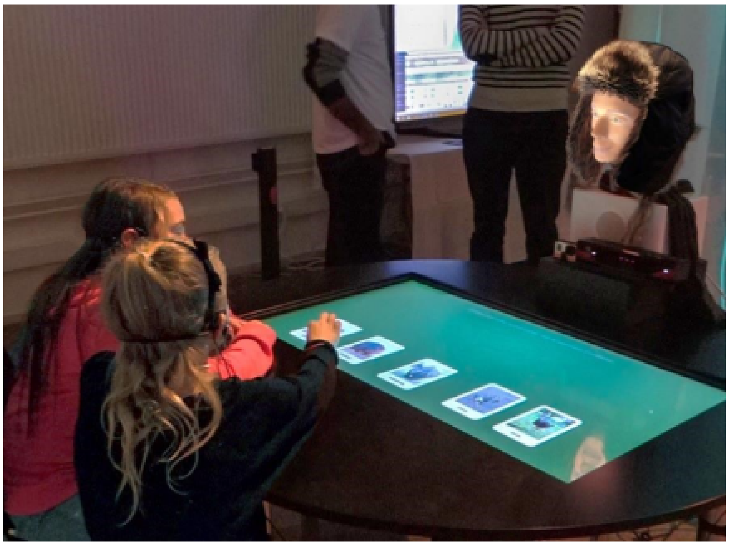
As discussed in section 2.1, humans prepare their response in dialogue while processing what is being said. This leads to minimal response times in human dialogue. The same behaviour is desirable for an exchange with a dialogue system. This means that dialogue systems should not wait until a turn was finished to start generating a response but rather start processing and refining the response incrementally with incomplete information. For example, [Tsai et al., 2019] compared an incremental and a non-incremental version of a movie recommendation dialogue system. They found that people who tested both systems not only perceived the incremental system to be "smoother" (which is a fact, since the delays between turns are shorter) but also rated the quality of the incremental system's recommendations higher, even though the recommendations were the same for both systems. This shows that the incremental system not only made the communication more human-like but also increased the user-perceived quality.

Multimodality

It was shown in section 2.1 that nonverbal information such as gestures or gaze has an impact on dialogues, for example, when humans anticipate the end of a turn. On the one hand, spoken dialogue systems can thus utilise multi-modal information to improve their turn-taking predictions. On the other hand, spoken dialogue systems can control the flow of the dialogue by imitating gaze or gestures. Also, the integration of visual input data allows the dialogue system to react better to the situation it operates in (for example, when referencing an object in the room). An example of the successful implementation of multi-modal information was presented by [Skantze et al., 2015]. In their work, a spoken dialogue system is described that is integrated into a robot head with a human appearance. The system was exhibited at the Swedish National Museum of Science and was able to play a simple collaborative card ordering game with visitors. In addition to the standard systems of a spoken dialogue system, the proposed system tracked the head and hand movement of the players with a Kinect camera.

Furthermore, the robot head was able to imitate human behaviour by turning its head, making eye movements, smiling and raising eyebrows. This enabled the system to both process visual turn-taking cues of the players and emit turn-taking cues to the players. The authors found that the players were much more likely to yield the turn to the spoken dialogue system when the system produced combinations of turn-taking cues, such as directing the gaze to the current speaker, smiling and raising eyebrows. Conversely, their analysis showed that players were much more likely to take the turn when the system signalled that it had finished its turn. Figure 2.2 shows the experimental setup of the system.

Figure 2.2: Experimental setup of the system described in [Skantze et al., 2015].



This example illustrates the positive potential impact of integrating multi-modal data streams on managing a fluent, natural dialogue. However, [Skantze et al., 2015] note that the behaviour of their proposed system is primarily based on hand-crafted policies, which makes it difficult to generalise the system for other settings.

Adaptivity

Communication styles are highly individual and are related to factors such as age and psychological background [Shum et al., 2018]. Therefore, instead of giving the same response to every person, dialogue systems should be highly customisable and adaptable to the situation they are operating in and to the person they are addressing. This could include adaptions to the preferred communication style of a person or the personality of a person. For example, [Wang et al., 2019] collected and analysed a dataset of human-to-human dialogues in which one person tried to convince another person to donate to charity. Furthermore, they collected the demographic background as well as the psychological background (e.g., Big Five personality traits, Schwartz Portrait scores) of the dialogue participants. The annotated dialogues were then examined for correlations between successful persuasion strategies and the participants' demographic and psychological data respectively. For instance, their research suggests that extroverted participants are more easily persuaded using emotional persuasion strategies. Using this information, a spoken dialogue system could be developed that adapts its persuasion strategy based on the demographic data and personality type of its interlocutor.

An attempt at creating an adaptive dialogue system is presented by [Zheng et al., 2019]. The authors analysed a dataset of 20.4 million written dialogues from 8.47 million Chinese speakers on social media. For each user in the dataset, age, gender and location tags were available. The dialogues were then grouped into an unbiased dataset and age, gender and location biased datasets. These four datasets were used to train seq2seq language generation models, which provided single-turn responses to questions that were part of the original dataset. All four models were then manually evaluated regarding appropriateness and fluency. The authors concluded that the biased datasets (which provided age/gender/location-specific responses) were rated significantly better than the unbiased dataset. This indicates that customised dialogue systems increase the user-perceived quality.

However, there seems to be little research on dialogue systems that goes beyond customisation based on demographic data such as gender or age.

2.2.2 Turn Taking Prediction in Spoken Dialogue Systems

In section 2.1 we have discussed the nature of human turn-taking in dialogue. Replicating these dynamics is desirable for a spoken dialogue system to give the user the feeling of natural communication. Following, we will discuss two main approaches.

Silence-Based Turn Taking Predictions

Systems that use this approach divide speech into evenly spaced segments (e.g., 250-500 ms). At the end of each segment, the system checks whether a certain duration of silence has been detected. If the duration exceeds a threshold, the system decides if the current speaker has finished their turn or if they just paused and did not yield their turn. This is achieved by analysing turn-taking cues, such as those listed in section 2.1.2. Several machine learning architectures have been proposed using different sets of turn-taking cues as features. [Ferrer et al., 2002] used a decision tree algorithm with intonation and pitch of the speaker as features. [Kawahara et al., 2012] proposed a logistic regression classifier with gaze features combined with intonation and pitch of the speaker. More recently, [Razavi et al., 2019] presented a naive Bayes classifier with a combination of lexical and acoustic features.

These approaches are based on the signalling theory from section 2.1.2 and assume that turn-taking in human dialogue is purely reactive. As we discussed in section 2.1, this is not accurate. Furthermore, the architectures proposed are non-incremental, which leads to long delays between the turns as described in section 2.2.1. Also, the performance of these models depends on the choice of segment length. When speech is divided into bigger segments, it is more likely that more turn-taking cues can be detected, but this also increases the delay between the end of a turn and the response. On the other hand, if the segments are too short, relevant turn-taking cues may be missed, reducing the accuracy of the model. Lastly, [Heldner and Edlund, 2010] noted that turn changes could often occur with slight overlaps in speech, for example, when a speaker produces a *backchannel*, such as "mhh" or "yeah". Therefore, models that assume a certain duration of silence between turns are struggling to detect these turn changes.

Continuous Turn Taking Predictions

In order to address the shortcomings of silence-based approaches, it was suggested that turn-taking predictions should be made continuously rather than only after an extended period of silence. For this, the speech is divided into much smaller segments (e.g., 50 ms), and after each segment, the system makes a prediction about who is going to speak in the upcoming segments. Figure 2.3 shows the difference between the continuous and the silence-based approach. [Skantze, 2017] used an LSTM model that was trained on human-to-human dialogue to predict speech activity up to 3 s in the future for each dialogue participant. An LSTM (Long short-term memory) is a deep neural network architecture. Their model makes a binary classification (speech or silence) in 50 ms time steps and can thus be used indirectly to predict the end of a turn. Voice activity, pitch and volume of speech were used as input data. Since an LSTM is a deep learning architecture, no manual feature engineering was required. The functionality of this approach is visualised in figure 2.4. Their model achieved better-than-human performance on the original dataset. The authors also applied the model to a corpus of human-machine dialogues, with a significant decrease in accuracy compared to the original dataset of human-human dialogues. This is probably due to the different characteristics of human-machine dialogue compared to human-human dialogue.

Figure 2.3: A comparison of the silence-based (non-incremental) approach (on the left) with the continuous (incremental) approach (on the right). Source: [Skantze, 2017]

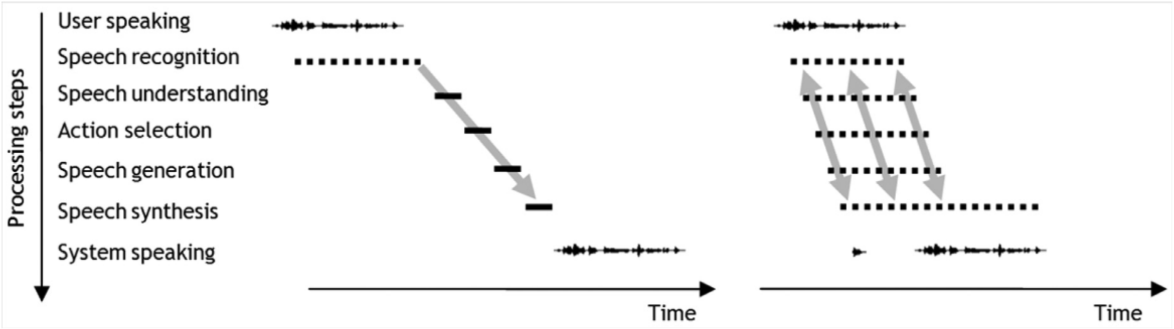
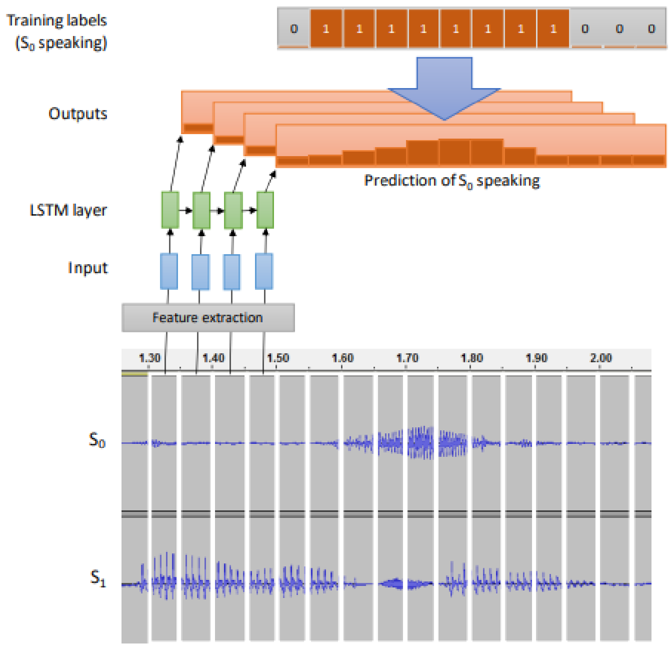


Figure 2.4: A visualisation of the functionality of the continuous approach proposed by [Skantze, 2017]. Source: [Skantze, 2017]



This approach was later implemented in several other publications. For example, [Ward et al., 2018] presented an improved LSTM architecture that outperformed the original model by [Skantze, 2017]. [Maier et al., 2017] also used an LSTM architecture but made use of linguistic features in addition to voice activity, pitch and volume of the speech as originally proposed by [Skantze, 2017]. Also, [Maier et al., 2017] trained their model on a dataset of human-machine dialogue, making it more suitable for use in a spoken dialogue system.

2.3 Big Five Personality Traits

In section 2.2.1 it was shown that a current research challenge lies in adapting dialogue systems to human personality. Therefore, an approach to classify human personality is presented.

Describing human personality is a difficult task since there is a wide variety of nuances that make each person different. [Allport and Odbert, 1936] theorised that people possess a number of traits (such as "optimistic" or "loyal") that can explain how people think and react in different situations. Over time, the Big Five model emerged to combine these traits into a multifactorial model. The model has been proven to be robust across cultures and is well researched. It consists of the five factors extroversion, agreeableness, neuroticism, openness and conscientiousness [Digman, 1990]. First of all, extroversion defines the extent to which people enjoy company, excitement and social interaction. A highly extroverted person also tends to be more emotionally expressive and assertive. Secondly, the agreeableness factor explains a person's tendency toward social harmony, cooperation, and altruism. Neuroticism refers to the emotional instability of a person. Someone with a high level of neuroticism is more likely to experience sudden mood swings, anxiety and is less resistant to stress. Openness describes a person's willingness to have new experiences. A wide range of interests, a high level of creativity and imagination are also associated with a high degree of openness. Lastly, the conscientiousness factor defines to which extent people are disciplined, organised and how well they can control their impulses. Table 2.1 gives an overview of the factors with commonly associated adjectives for high and low levels of the factors, respectively.

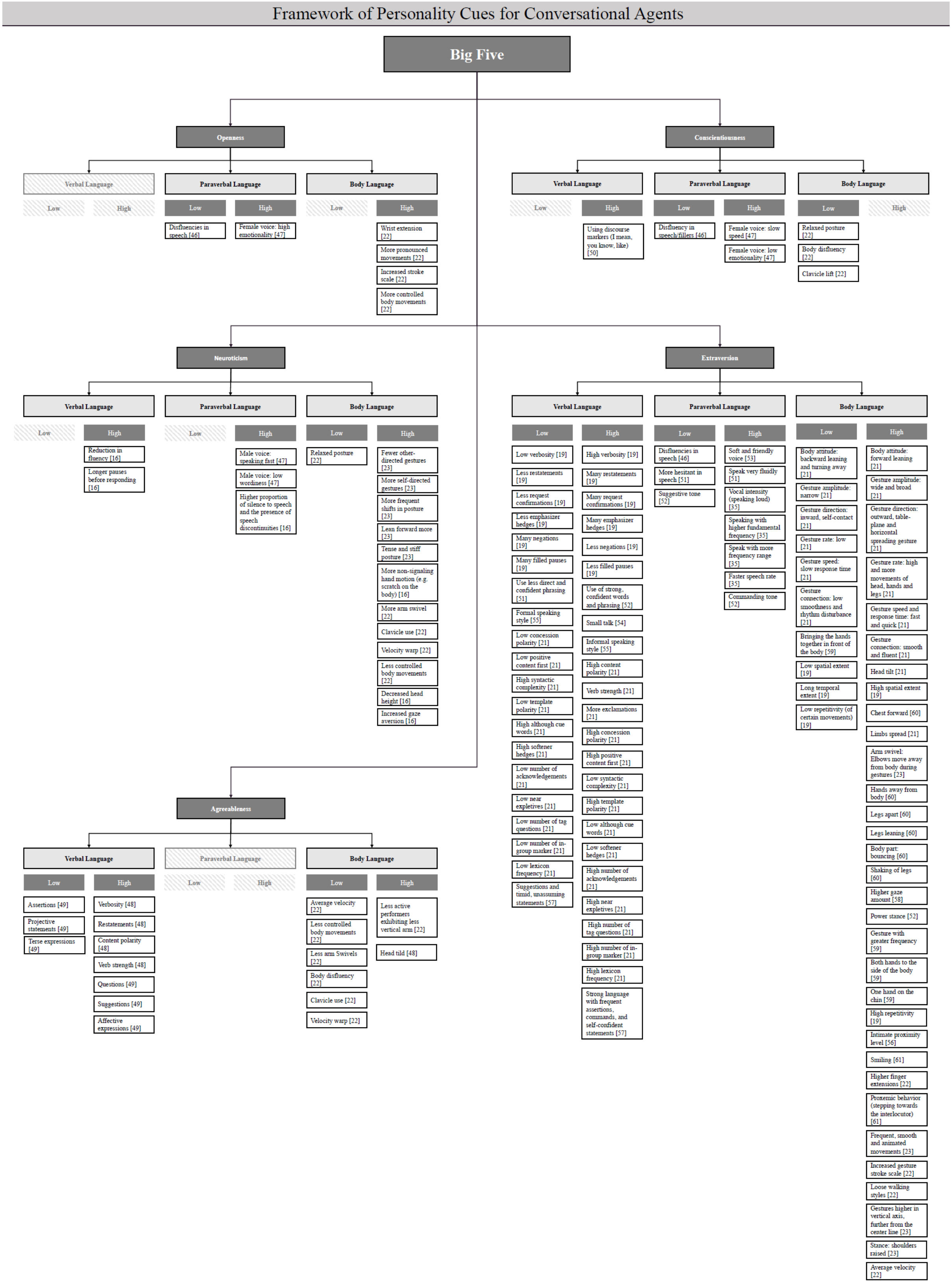
Table 2.1: Big-Five personality factors and commonly associated adjectives. Source [Mairesse and Walker, 2008]

Factor	High	Low
Extroversion	sociable, talkative, optimistic, assertive	quiet, passive, reserved, shy
Agreeableness	altruistic, trustworthy, understanding, sympathetic	unfriendly, selfish, suspicious, uncooperative, malicious
Neuroticism	anxious, neurotic, rude, depressed	calm, self-confident, reliable
Openness	creative, intellectual, imaginative, curious	practical, conservative, ignorant
Conscientiousness	organised, disciplined, hardworking, competent	lazy, unreliable, forgetful, impulsive

As discussed in section 2.2.1, a major challenge in modern dialogue systems is to adapt the systems to different personality types. For this reason, many studies have attempted to identify dialogue characteristics that correlate with the factors in the Big Five model. For example, [Ahmad et al., 2022] has conducted an extensive literature review summarising verbal, paraverbal, and body language cues that were found to correlate with high and low levels of the respective Big Five factors. Figure 2.5 shows a complete overview of the cues that were identified as part of the literature review by [Ahmad et al., 2022].

A total of 148 cues were identified in the paper. With 90 cues, most were identified for the extroversion factor, while only six cues were identified for the openness factor. Furthermore, seven cues were identified for the conscientiousness factor. This shows that a heavy imbalance exists in cues that were identified for the respective Big Five personality factors. Moreover, it is evident that most cues are based on body language and verbal cues, while paraverbal cues have been little explored. Finally, some cues were found to be contradictory as they were reported for both low and high values of the respective factor. In summary, these factors show that this area of research is still under-researched.

Figure 2.5: Overview of verbal, paraverbal and body language cues that were found to correlate with high and low levels of the respective Big Five factors. Source: [Ahmad et al., 2022]

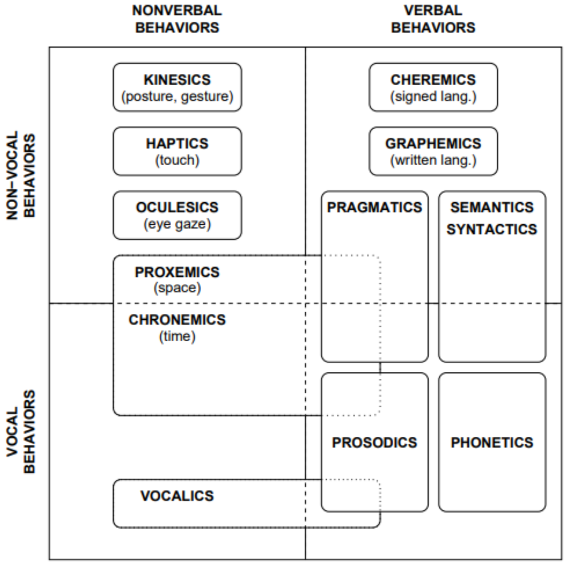

Framework of Personality Cues for Conversational Agents
Big Five
Openness
Verbal Language
Low
High
Paraverbal Language
Low
Disfluencies in speech [46]
High
Female voice: high emotionality [47]
Body Language
Low
High
Wrist extension [22]
More pronounced movements [22]
Increased stroke scale [22]
More controlled body movements [22]
Conscientiousness
Verbal Language
Low
High
Using discourse markers (I mean, you know, like) [50]
Paraverbal Language
Low
Disfluency in speech/fillers [46]
High
Female voice: slow speed [47]
Female voice: low emotionality [47]
Body Language
Low
Relaxed posture [22]
Body disfluency [22]
Clavicle lift [22]
High
Neuroticism
Verbal Language
Low
High
Reduction in fluency [16]
Longer pauses before responding [16]
Paraverbal Language
Low
High
Male voice: speaking fast [47]
Male voice: low wordiness [47]
Higher proportion of silence to speech and the presence of speech discontinuities [16]
Body Language
Low
Relaxed posture [22]
High
Fewer other-directed gestures [23]
More self-directed gestures [23]
More frequent shifts in posture [23]
Lean forward more [23]
Tense and stiff posture [23]
More non-signaling hand motion (e.g. scratch on the body) [16]
More arm swivel [22]
Clavicle use [22]
Velocity warp [22]
Less controlled body movements [22]
Decreased head height [16]
Increased gaze aversion [16]
Extraversion
Verbal Language
Low
Low verbosity [19]
Less restatements [19]
Less request confirmations [19]
Less emphasizer hedges [19]
Many negations [19]
Many filled pauses [19]
Use less direct and confident phrasing [51]
Formal speaking style [55]
Low concession polarity [21]
Low positive content first [21]
High syntactic complexity [21]
Low template polarity [21]
High although cue words [21]
High softener hedges [21]
Low number of acknowledgements [21]
Low near expletives [21]
Low number of tag questions [21]
Low number of in-group marker [21]
Low lexicon frequency [21]
Suggestions and timid, unassuming statements [57]
High
High verbosity [19]
Many restatements [19]
Many request confirmations [19]
Many emphasizer hedges [19]
Less negations [19]
Less filled pauses [19]
Use of strong, confident words and phrasing [52]
Small talk [54]
Informal speaking style [55]
High content polarity [21]
Verb strength [21]
More exclamations [21]
High concession polarity [21]
High positive content first [21]
Low syntactic complexity [21]
High template polarity [21]
Low although cue words [21]
Low softener hedges [21]
High number of acknowledgements [21]
High near expletives [21]
High number of tag questions [21]
High number of in-group marker [21]
High lexicon frequency [21]
Strong language with frequent assertions, commands, and self-confident statements [57]
Paraverbal Language
Low
Disfluencies in speech [46]
More hesitant in speech [51]
Suggestive tone [52]
High
Soft and friendly voice [53]
Speak very fluidly [51]
Vocal intensity (speaking loud) [35]
Speaking with higher fundamental frequency [35]
Speak with more frequency range [35]
Faster speech rate [35]
Commanding tone [52]
Body Language
Low
Body attitude: backward leaning and turning away [21]
Gesture amplitude: narrow [21]
Gesture direction: inward, self-contact [21]
Gesture rate: low [21]
Gesture speed: slow response time [21]
Gesture connection: low smoothness and rhythm disturbance [21]
Bringing the hands together in front of the body [59]
Low spatial extent [19]
Long temporal extent [19]
Low repetitivity (of certain movements) [19]
High
Body attitude: forward leaning [21]
Gesture amplitude: wide and broad [21]
Gesture direction: outward, table-plane and horizontal spreading gesture [21]
Gesture rate: high and more movements of head, hands and legs [21]
Gesture speed and response time: fast and quick [21]
Gesture connection: smooth and fluent [21]
Head tilt [21]
High spatial extent [19]
Chest forward [60]
Limbs spread [21]
Arm swivel: Elbows move away from body during gestures [23]
Hands away from body [60]
Legs apart [60]
Legs leaning [60]
Body part: bouncing [60]
Shaking of legs [60]
Higher gaze amount [58]
Power stance [52]
Gesture with greater frequency [59]
Both hands to the side of the body [59]
One hand on the chin [59]
High repetitivity [19]
Intimate proximity level [56]
Smiling [61]
Higher finger extensions [22]
Proxemic behavior (stepping towards the interlocutor) [61]
Frequent, smooth and animated movements [23]
Increased gesture stroke scale [22]
Loose walking styles [22]
Gestures higher in vertical axis, further from the center line [23]
Stance: shoulders raised [23]
Average velocity [22]
Agreeableness
Verbal Language
Low
Assertions [49]
Projective statements [49]
Terse expressions [49]
High
Verbosity [48]
Restatements [48]
Content polarity [48]
Verb strength [48]
Questions [49]
Suggestions [49]
Affective expressions [49]
Paraverbal Language
Low
High
Body Language
Low
Average velocity [22]
Less controlled body movements [22]
Less arm Swivels [22]
Body disfluency [22]
Clavicle use [22]
Velocity warp [22]
High
Less active performers exhibiting less vertical arm [22]
Head tild [48]

2.4 Nonverbal Analysis of Dialogue

It was shown in section 2.3 that little is known about nonverbal signals that are related to Big Five personality traits. This section, therefore, presents research that deals with the nonverbal analysis of dialogue.

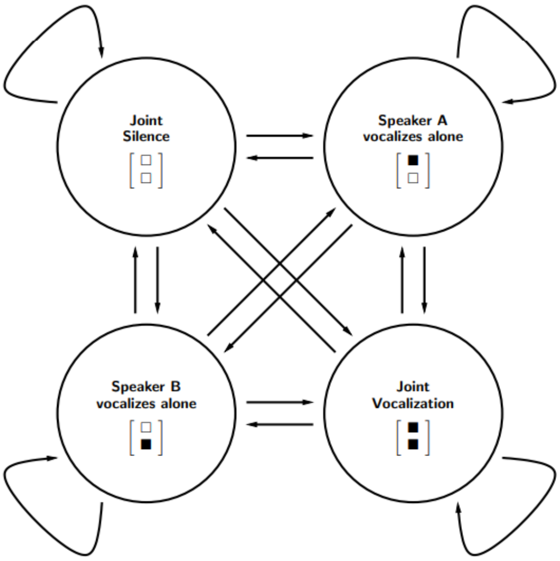
Irrespective of the spoken language, humans are able to infer the "mode" of dialogue (such as an argument or collaborative dialogue) based on information such as the volume and pitch of the voices, gestures, body language and elapsed time between the turns of the speakers. For example, loud talking and short time between turns could indicate an argument, while calm voices and relaxed body language could hint toward a collaborative dialogue. However, understanding the context of an overheard conversation is a challenging task for spoken dialogue systems. The nonverbal analysis of dialogue can help bridge this gap and is thus important for the design of spoken dialogue systems. Figure 2.6 shows a rough division between nonverbal and verbal behaviours in dialogue. In particular, the overlap of vocal behaviour with nonverbal behaviour (*chronemics* and *vocalics*) is of interest for spoken dialogue systems. It can be processed with a speech recognition module (which is usually part of a spoken dialogue system anyway) and does not require additional sensory input such as a camera. Chronemics includes the analysis of the timing of speech activity in dialogue, while the study of vocalics includes nonverbal manipulation of voice such as pitch or volume.

Figure 2.6: A rough classification of verbal and nonverbal behaviours in dialogue. Source: [Laskowski, 2011]



As a means of content-free representation of dialogues, [Jaffe et al., 1967] first came up with the idea of modelling dialogues as Markov models. Dialogue is represented as a set of states with transition probabilities between the states. For example, a two-party dialogue can be represented with the states "person A speaking", "person B speaking", "joint speech", and "joint silence". Figure 2.7 shows an example of such a Markov model. [Laskowski, 2011] further developed this idea and presented a general model for multiparty dialogue.

Figure 2.7: A Markov model for two-party dialogue. Source: [Laskowski, 2011]



2.5 Laughter in Dialogue

Laughter is prevalent in human-to-human dialogue and is oftentimes associated with humorous situations. However, laughter is also an important nonverbal signal that fulfills a variety of social functions such as expressing disbelief, sarcasm or to express sympathy [Mazzocconi et al., 2020].

Depending on these different social functions, literature has attempted to divide laughter into different categories. [Koutsombogera and Vogel, 2022] used a binary classification of *mirthful* and *discourse* laughter to analyse occurrence patterns of laughter in different thematic structures of dialogue. An instance of laughter is classified as mirthful when it happens due to amusement. All other instances of laughter that not happen due to amusement but rather to fulfill a social function are classified as discourse laughter. Meanwhile other works like [Reuderink et al., 2008] have classified laughter based on acoustic features in hearty, amused, satirical and social laughter. In [Szameitat et al., 2009], laughter is classified based on the expressed emotion into joy, tickling, taunting and "schadenfreude" laughter. In summary, no clear classification of laughter has emerged in the literature, and it therefore seems reasonable to choose a classification that best suits the research purpose at hand.

However, it does not seem like laughter has been considered as a feature for turn-taking models in existing literature.

2.6 Summary

The literature review showed that while spoken dialogue systems have made significant progress through the use of deep learning techniques, they are still mainly used for simple tasks such as asking for the weather forecast. This indicates that spoken dialogue systems are perceived only as an interface and not as actual interlocutors. To hold natural feeling dialogues with human users that go beyond just one-turn interactions, it is therefore essential to develop spoken dialogue systems that are capable of replicating human-level turn-taking.

In order to overcome this challenge, the three research areas of incrementality, multimodality and adaptivity have been identified. In particular, developing systems that adapt to the user's personality type remains challenging. While previous research has identified dialogue qualities that correlate with the various factors of the Big Five personality model, there is a strong imbalance in the number of cues identified for the different Big Five factors. Furthermore, most of these cues are based on body language or are verbal. There is only limited research that has identified paraverbal and nonverbal cues that are correlated with the Big Five factors.

Based on this, we pose the following research questions:

Q1 What are nonverbal qualities in dialogue related to Big Five personality factors?

Q2 How well does a continuous turn-taking model perform that makes its predictions based on nonverbal qualities?

In order to answer these questions, the following approach is proposed:

1. A multimodal dataset of spoken English dialogues is analysed. The dataset contains information about the Big Five personality factors of the speakers, as well as dialogue annotations. These annotations are utilised to calculate nonverbal dialogue qualities for each speaker. The qualities are then compared between the speakers in order to find out whether certain personality traits influence dialogue qualities.

2. A logistic regression model is built on basis of the dataset mentioned above. The model continuously predicts speech activity for the dialogues in the dataset based on nonverbal features. The performance and the feature importance scores are then analysed.

3 Methods

This section provides a description of the research methods used in the thesis to answer the research questions that were introduced in section 2.6. First, the used dataset is described. After that, the preprocessing of the dataset is described, followed by an explanation of the used machine learning methods and the used evaluation methods.

3.1 Dataset

For this research, we use the MULTISIMO dataset. The dataset is a multimodal corpus which contains audio and video recordings of human spoken English dialogues. The dialogues are set in a game environment, and each dialogue consists of three participants with one facilitator and two players. The corpus in total contains 23 dialogues. Of these, 18 sessions have been published with an average length of 10 minutes and a total duration of about three hours. The recording of the dialogues took place in the School of Computer Science and Statistics at Trinity College Dublin in 2018. The corpus is publicly available for download [1]. The participants come from a variety of cultural backgrounds, with 16 native English speakers and 33 non-native English speakers. Furthermore, the gender of the participants is balanced. [Koutsombogera and Vogel, 2018]

3.1.1 Structure of the Dialogues

Before recording the dialogues, 100 people were recruited to answer a set of three questions. The task of the two players in each dialogue is to find the three most frequent answers to each question and then rank them according to the frequency of their occurrence. The three questions asked to all players are:

> "Name three instruments that can be found in a symphony orchestra."
>
> "Name three public places where people are likely to catch the flu."
>
> "Name three things that people cut."

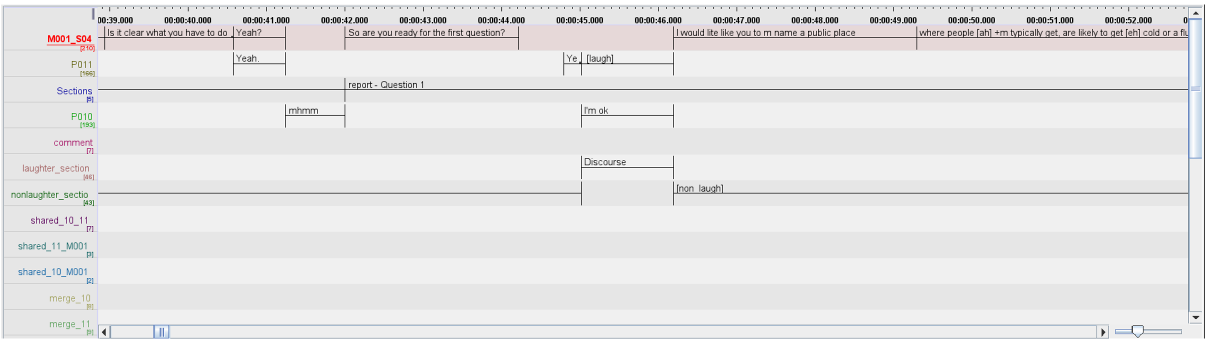
[1] http://multisimo.eu/datasets.html

In each dialogue, the facilitator starts with an introduction in which the game is explained. Then the facilitator asks the players the three questions mentioned above. For each question, the players have then to find the three right answers and order them according to their popularity. The facilitator can intervene to help the players answer the questions or ensure both players have the same amount of speaking time. After the three questions are correctly answered, the facilitator closes the conversation. The design of the game is intended to elicit cooperative behaviour among players toward a common goal.

3.1.2 Annotations

In addition to the audio and video recordings of the dialogues, the dataset contains the transcriptions of the dialogues. The speech annotations contain the segmentation of turns, timings and spoken content for both the facilitator and the players. Gaps are also annotated when no speaker is active and overlaps when more than one speaker is active. Moreover, instances of laughter are annotated along with the timing of the laughter and whether it was mirthful or discursive. Furthermore, there are gaze annotations of the participants available for two dialogues. These annotations contain timings and the gaze focus (gaze at facilitator, player one, player two, or gaze away) for all participants. The annotations were carried out manually using the Transcriber software. The annotated files can be opened using the ELAN software. The ELAN software is a tool for creating annotations for audio and video recordings and was developed by the Max Planck Institute for Psycholinguistics. The tool can be downloaded free of charge [2]. Figure 3.1 shows an example screenshot of the dialogue annotations opened in the ELAN software.

Figure 3.1: Dialogue annotations in the ELAN software



Is it clear what you have to do

Yeah?

So are you ready for the first question?

I would lite like you to m name a public place

where people [ah] +m typically get, are likely to get [eh] cold or a flu

Yeah.

Ye

[laugh]

report - Question 1

mhmm

I'm ok

Discourse

[non laugh]

3.1.3 Big Five Personality Assessment

All participants completed the Big Five Inventory before the recording of the dialogues. The Big Five Inventory is a test consisting of 44 statements related to a person's personality. For

[2] https://archive.mpi.nl/tla/elan/download

each statement, the participants were asked to indicate whether they agreed or disagreed with it on a scale from one to five. Based on this questionnaire, scores for each Big Five personality factor (extroversion, agreeableness, neuroticism, openness and conscientiousness) were assessed for each participant. Also, the percentile rank of each participant across the five personality factors was calculated. Finally, the percentiles are normed on the overall population of the participants. The MULTISIMO dataset contains an anonymised table with all participants' absolute scores and percentile rank of each personality factor.

3.2 Data Preprocessing

This section describes general preprocessing steps that were taken. Additional preprocessing steps that were only executed for the respective experiments are described in detail in chapter 4.

The dialogues are published as eaf files that can be opened with the ELAN software. From the ELAN software, the annotations of the dialogues are exported to CSV format so they can be processed automatically. The resulting CSV files did not contain column names. Therefore the following columns are defined for each CSV file/dialogue:

- TYPE (Type of the annotation. The different types of annotations are explained below)
- TYPE2 (Dialogue participant that the annotation refers to)
- START (Timestamp of the start of the annotated action in HH:MM:SS format)
- START2 (Timestamp of the start of the annotated action in seconds)
- END (Timestamp of the ending of the annotated action in HH:MM:SS format)
- END2 (Timestamp of the ending of the annotated action in seconds)
- DURATION (Duration of the annotated action in HH:MM:SS format)
- DURATION2 (Duration of the annotated action in seconds)
- ACTION (Content of the annotation)

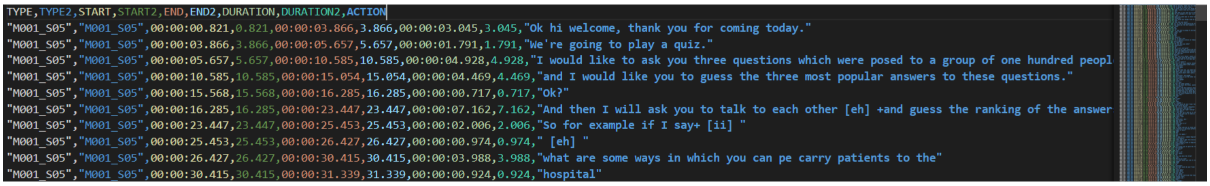
The TYPE column contains different types of annotations. In the following, all types that are relevant to the thesis are listed:

- ID of the facilitator - contains spoken content of the facilitator in string format
- ID of player 1 - contains spoken content of player 1 in string format
- ID of player 2 - contains spoken content of player 2 in string format

- Sections - contains time stamps of starting and end times of the five sections of each dialogue (introduction, question 1, question 2, question 3, closing)
- Turns - contains speech activity of all players (without spoken content) for the whole dialogue in chronological order
- Laughter sections for the facilitator - timestamps for laughter sections of the facilitator and whether it was discourse or mirthful laughter
- Laughter sections for player 1 - timestamps for laughter sections of player 1 and whether it was discourse or mirthful laughter
- Laughter sections for player 2 - timestamps for laughter sections of player 2 and whether it was discourse or mirthful laughter
- Non-laughter sections - timestamps for all sections which do not contain laughter

An example of a preprocessed CSV file is shown in figure 3.2

Figure 3.2: Screenshot of a preprocessed CSV file



```
TYPE,TYPE2,START,START2,END,END2,DURATION,DURATION2,ACTION
"M001_S05","M001_S05",00:00:00.821,0.821,00:00:03.866,3.866,00:00:03.045,3.045,"Ok hi welcome, thank you for coming today."
"M001_S05","M001_S05",00:00:03.866,3.866,00:00:05.657,5.657,00:00:01.791,1.791,"We're going to play a quiz."
"M001_S05","M001_S05",00:00:05.657,5.657,00:00:10.585,10.585,00:00:04.928,4.928,"I would like to ask you three questions which were posed to a group of one hundred peopl
"M001_S05","M001_S05",00:00:10.585,10.585,00:00:15.054,15.054,00:00:04.469,4.469,"and I would like you to guess the three most popular answers to these questions."
"M001_S05","M001_S05",00:00:15.568,15.568,00:00:16.285,16.285,00:00:00.717,0.717,"Ok?"
"M001_S05","M001_S05",00:00:16.285,16.285,00:00:23.447,23.447,00:00:07.162,7.162,"And then I will ask you to talk to each other [eh] +and guess the ranking of the answer
"M001_S05","M001_S05",00:00:23.447,23.447,00:00:25.453,25.453,00:00:02.006,2.006,"So for example if I say+ [ii] "
"M001_S05","M001_S05",00:00:25.453,25.453,00:00:26.427,26.427,00:00:00.974,0.974," [eh] "
"M001_S05","M001_S05",00:00:26.427,26.427,00:00:30.415,30.415,00:00:03.988,3.988,"what are some ways in which you can pe carry patients to the"
"M001_S05","M001_S05",00:00:30.415,30.415,00:00:31.339,31.339,00:00:00.924,0.924,"hospital"
```

Session 3 has a different scheme of annotations, making it difficult to process it along with the rest of the sessions automatically. Therefore, session 3 is removed from the dataset and will not be considered in further analysis.

Some dialogue annotations contained typos that hindered the respective session's automatic processing. These typos were corrected (e.g., session 8 had the value "queation 3" for the "Sections" row).

3.3 Machine Learning Methods

Machine learning is a field of research that combines statistical methods and computer science. Machine learning algorithms can infer behaviour from data patterns without the need to define a procedure for all possibilities, as in traditional programming. These algorithms have been developed for many years, but they became particularly popular over the last decade when increased computing power made it possible to process large data sets.

Machine learning algorithms can be divided into supervised, unsupervised, and reinforcement learning algorithms. In supervised learning, the machine learning model is "trained" on labelled data with the goal of finding a pattern that can predict a given target variable. After the "training" is completed, the model can make predictions on unseen data. Supervised learning algorithms can predict categorical data (classification) and continuous data (regression). An example of supervised learning would be the classification of objects in an image. Unsupervised learning algorithms do not require labelled training data. These types of algorithms do not make predictions for a given target variable but rather try to find hidden patterns in a dataset. An example use case for unsupervised learning would be identifying groups of customers from a dataset that follow a similar behaviour. Reinforcement learning algorithms are the closest to the way humans learn. The algorithm interacts with an environment and is graded based on a cost- and reward function. Over time the agent "learns" the optimal policy to interact with the environment, which yields the desired behaviour. An example of reinforcement learning would be an agent that learns to play a game over time by playing the game millions of times until it finally finds an optimal policy to win the game.

Following, the machine learning methods that are used in the thesis are described in detail. First, logistic regression is described and justification is given as to why it is used. Then, the conceptual approach to feature engineering is described.

3.3.1 Logistic Regression

Logistic regression is a supervised learning method that is used for classification problems. For supervised learning, a dataset of a dependent variable and multiple independent variables is divided into training and test set. The supervised learning algorithm is first "trained" on the training set to find a pattern in the data that can predict the dependent variable. The model is then validated on the test set. The details of the "training" of a logistic regression model are described in the following.

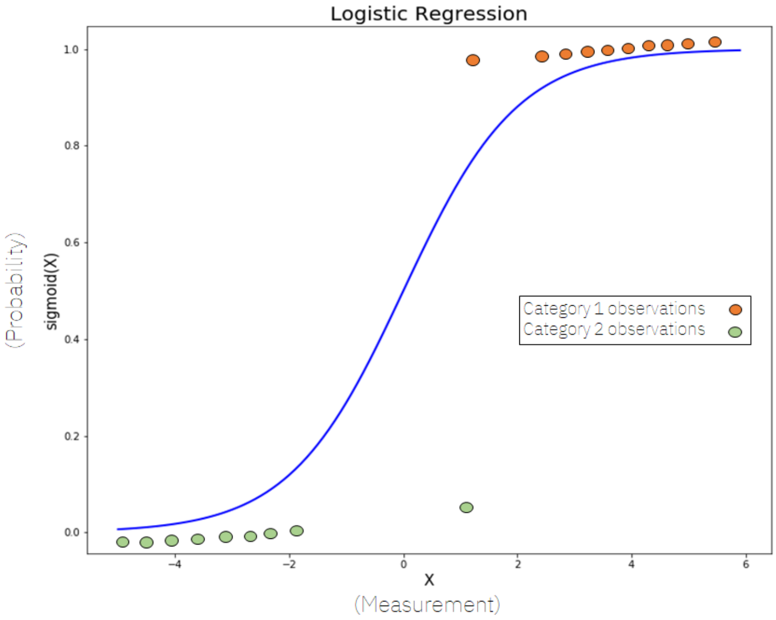
A target function maps a vector of input values (called features) to a predicted output. Sometimes the target function is also called "hypotheses" in literature. However, in this thesis, the term "target function" will be used to not confuse it with the terms introduced in section 3.4.1. The target function $h_\theta(x)$ for logistic regression is defined as:

$$h_\theta(x) = \frac{1}{1 + e^{-\theta^T x}} \tag{1}$$

where x is the set of features for a sample and θ is the set of weights for the features.

We can see that the target function maps a set of input features to a value between 0 and 1. The output of the target function can thus be interpreted as the probability that a sample is part of class A or class B. The sample is then classified as class A or B based on a defined probability threshold. For example, suppose the threshold is 0.5. In that case, every sample below that gets classified as A, and everything above that gets classified as B. Figure 3.3 illustrates this with an example.

Figure 3.3: Sigmoid function for logistic regression. Samples with a high output value of the target function are classified into category 1 (red), while samples with a low output value for the target function are classified into category 2 (green)
Source: `https://amalaj7.medium.com/logistic-regression-eb29032511079`



In order to find the set of weights θ that divide classes A and B optimally, the *gradient descent* algorithm is used. The gradient descent algorithm utilises a *cost function* that measures the difference between the predicted output and actual values. According to the output of the cost function, the weights θ are adjusted. The gradient descent algorithm then repeats this procedure until the model *converges* which means that the output of the cost function does not become any smaller. The final set of weights θ indicates how much each feature impacted the prediction. When a feature has a weight close to zero, it means they hardly impact the final prediction. Conversely, a feature with a high (absolute) final weight has a high impact on the final prediction of the model. For logistic regression, the cost function is defined as:

$$cost(h_\theta(x), y) = \begin{cases} \log(h_\theta(x)), & \text{if } y = 1 \\ -\log(1 - h_\theta(x)), & \text{if } y = 0 \end{cases} \tag{2}$$

where h_θ is the target function, y is the predicted output, and x is the vector of input features.

When using large feature sets, it can happen that the model *overfits*. This means that the model exactly fits the training data but does not generalise to unseen data, which negatively impacts the quality of the predictions. In order to avoid overfitting when using large feature sets, a so-called *regularisation* term can be added to the cost function.

The standard logistic regression algorithm only works for binary classes but can be modified to work with multiple classes. If we consider a classification problem with three classes, A, B and C, logistic regression is first executed for class, and the two remaining classes, B and C, are grouped together. Then, logistic regression is executed again for class B and classes A and C are grouped together. This procedure is done until all classes have been covered and all samples have been classified.

The literature review has shown that state-of-the-art systems use LSTM networks (see section 2.2) to make turn-taking predictions continuously. However, LSTM networks require the tuning of many hyperparameters (such as choice of activation functions, selection of layers, number of neurons in the layers, etc.) and operate as a "black box", making it difficult to understand how the predictions were obtained.

This work focuses on understanding the impact of specific turn-taking cues rather than optimising performance. Logistic regression is therefore used in this work for two main reasons. Firstly, logistic regression is comparatively easy to use since only one hyper-parameter (regularisation) has to be tuned. Secondly, logistic regression is more straightforward to interpret than more complicated neural network models because the coefficients indicate which features impact the model's predictions. Understanding the effects of each feature can yield interesting information.

For this work, the *LogisticRegression* model from the *sklearn* library is used.

3.3.2 Feature Engineering

An important part of building supervised machine learning models is the selection of the right features on which the predictions are based. This section, therefore, describes the conceptual approach to feature engineering. We start by building an initial set of features for the model. Next, the recursive feature elimination (RFE) algorithm from the *sklearn* library is used to select the optimal set of features. The algorithm iteratively trains the machine learning model. It then utilises the importance score of the features (such as the feature weights in the logistic regression model) to prune the feature set. After each iteration, the n features with the lowest importance score are pruned until a final set of m features is reached, where n and m are integers that are passed to the RFE function.

3.4 Evaluation Methods

This section presents the methods required for the evaluation part of the thesis. First statistical hypotheses testing is described, followed by a description of the Wilcoxon Rank Sum Test. Then, performance measurements for classifier algorithms are presented. Precision, Recall and F1 score are applicable for unbalanced datasets, whereas accuracy and error rates are better suited for balanced datasets.

3.4.1 Hypotheses Testing

When performing a statistical test, a null hypothesis and alternative hypotheses are established. For example, the null hypothesis might be that a data set has an underlying normal distribution. The alternative hypothesis would be that the dataset does not follow a normal distribution. Together with a statistical test, a p-value is calculated, which indicates how likely it is that an observation occurred under the assumption that the null hypothesis is correct. The lower the p-value, the more likely it is that the null hypothesis is false. When the p-value is low enough, the null hypotheses can be rejected, and the alternative hypotheses can be accepted instead.

That means a result can be considered statistically significant if the p-value is low enough, so it is improbable that the null hypothesis was falsely rejected. A common threshold for statistical significance is $p \leq 0.05$ while a result with $p \leq 0.01$ is considered highly significant.

3.4.2 Wilcoxon Rank Sum Test

The *Wilcoxon Rank Sum Test* (also called Mann-Whitney U Test) is a non-parametric test that determines whether the distributions of two groups have the same underlying shape (this can also be interpreted as a check for a significant difference between the median of the two groups).

To apply the Wilcoxon Rank Sum Test following assumptions have to be fulfilled:

1. The dependent variable has to be continuous or ordinal
2. The independent variable has to be categorical with two groups (dichotomous)
3. The two groups of the independent variable must not be related (independence of observations). An example of a related variable would be when a participant in a study belongs to both examined groups.

Since the test is non-parametric, no underlying distribution of the examined data is assumed. For this reason, the Wilcoxon Rank Sum Test is often used as an alternative to the t-Test, which requires a normal distribution of the dependent variable but also has more statistical power.

The null hypotheses and the alternative hypothesis can be formulated as:

H_0 The two groups have the same underlying distribution. This means we expect to see the same values in the dependent variable for both groups.

H_1 The two groups do not have the same underlying distribution. This means we expect to see different values in the dependent variable for both groups.

In the Wilcoxon Rank Sum Test, the ranks of both groups are summed up to R_1 and R_2. The statistic U of the test is then calculated by:

$$U_1 = n_1 * n_2 + \frac{n_1 * (n_1 + 1)}{2} - R_1 \tag{3}$$

$$U_2 = n_1 * n_2 + \frac{n_2 * (n_2 + 1)}{2} - R_2 \tag{4}$$

$$U = min(U_1, U_2) \tag{5}$$

where n_1 and n_2 are the respective sizes of the groups being compared.

For this work, we use the *mannwhitneyu* function from the *scipy stats* package to calculate the Wilcoxon Rank Sum Test.

3.4.3 Precision

Precision is a measurement used to evaluate a classifier's performance. It is defined as the ratio of true positive (TP) classifications to all classifications made (true positive + false positive classifications). This means Precision measures the percentage of correctly classified values out of all relevant values. Maximising Precision is therefore essential when one is concerned about minimising the number of false positives.

$$Precision = \frac{TP}{TP + FP} \tag{6}$$

where TP is the number of values that were correctly classified as positive and FP is the number of values that were incorrectly classified as positive.

3.4.4 Recall

Recall is another measurement used to evaluate a classifier's performance. It is defined as the ratio of true positive classifications to all actual positive values in the dataset. This means Recall measures the percentage of relevant results that were correctly classified. Maximising Recall is therefore essential when one is concerned about minimising the number of false negatives.

$$Recall = \frac{TP}{TP + FN} \tag{7}$$

where TP is the number of values that were correctly classified as positive and FN is the number of values that were incorrectly classified as negative.

3.4.5 F1 Score

It is not possible to maximise both Recall and Precision at the same time. This leads to a trade-off where, depending on the application, it must be decided whether minimising false negatives or minimising false positives is more important. When neither Recall nor Precision is clearly more important, one can measure the performance of a binary classifier with the F1-score, which is defined as the harmonic mean of Recall and Precision.

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{8}$$

3.4.6 Accuracy and Error Rate

When the examined dataset is balanced, one can use the accuracy and error rate measurement. Accuracy describes the percentage of all correctly classified elements in all elements of the data set. In contrast, the error rate describes the percentage of all negatively classified elements in all elements of the data set.

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \tag{9}$$

$$ErrorRate = \frac{FP + FN}{TP + FN + FP + TN} \tag{10}$$

where TP are the values that were correctly classified, TN are the values that were correctly classified as negative, FP are the values that were falsely classified as positive, and FN are the values that were falsely classified as negative.

4 Evaluation

This chapter describes the experiments conducted to answer the research questions introduced in section 2.6. Furthermore, all necessary implementation steps to conduct the experiments are explained. The research methods required for this were formally introduced in chapter 3. Finally, the results of the conducted experiments can be found in chapter 5.

First, the experiments to analyse the influence of big five personality traits on dialogue qualities are described. Then, the implementation of the continuous turn-taking model that is based on nonverbal features is described.

4.1 Influence of Big Five Personality Traits on Dialogue Qualities

This experiment aims to analyse whether differences in Big Five personality factors influence nonverbal qualities in dialogue. The results could help to better understand individuals' nonverbal communication styles, which can be useful for the design of turn-taking models in spoken dialogue systems. This experiment therefore addresses research question **Q1**. In order to analyse the effect of Big Five personality traits on nonverbal qualities in dialogue, the following approach is taken.

For all qualities, only the population of players is considered. Facilitators are not examined together with the players since their role in the dialogues is fundamentally different from the role of the players. Furthermore, the facilitators are not examined separately since only three of them exist. This means the overall population of facilitators is too small to draw any conclusions. All players are grouped into either a "Low" or a "High" group for all Big Five personality factors. Since the overall population of players is limited to 34, binary classification is chosen so that the group sizes remain large enough to make statistically relevant claims. Players with a percentile score of ≥ 50 are grouped into the respective "High" group of the factor, while players with a percentile score of < 50 are grouped into the "Low" group of the respective factor. These percentile scores are provided as a part of the MULTISIMO dataset (see section 3.1). The distributions of players for all groups are shown in table 4.1. All "High" groups contain 18 players, while the "Low" groups contain 16 players. This is due to the fact that the percentile values of the players were normed to the total population of participants when the MULTISIMO dataset was created. This means that the "High" and "Low" groups of personality factors should not be understood as absolute values, but rather in relation to the other participants in the experiment (e.g. a player with "High" extroversion is *more* extroverted than the average player in the experiment).

Table 4.1: Distribution of players for the "High" and "Low" groups of the respective Big Five personality factor

Factor	Number of Players in "High" Group	Number of Players in "Low" Group
Extroversion	18	16
Agreeableness	18	16
Neuroticism	18	16
Openness	18	16
Conscientiousness	18	16

All examined dialogue qualities are calculated for all players in the next step by looping over all CSV files in the MULTISIMO dataset. The specific calculations for the dialogue qualities are described below. With the calculated qualities, a dataframe with the following columns is constructed:

- PLAYER - contains the ID of the player
- EXT - contains 1 if the player is in the "High" extroversion group, 0 otherwise
- AGR - contains 1 if the player is in the "High" agreeableness group, 0 otherwise
- CON - contains 1 if the player is in the "High" conscientiousness group, 0 otherwise
- NEU - contains 1 if the player is in the "High" neuroticism group, 0 otherwise
- OPE - contains 1 if the player is in the "High" openness group, 0 otherwise
- Columns for all examined dialogue qualities - contain calculated results of the respective dialogue quality

For all examined qualities and personality traits, the original dataframe is then split into two dataframes. The first dataframe contains all players who are part of the "High" group of the examined personality factor and their respective dialogue quality values. The second dataframe contains all players who are part of the "Low" group of the examined personality factor and their respective dialogue quality values. A Wilcoxon Rank Sum test is then conducted to determine if there is a statistically significant difference in dialogue qualities between the group with low/high scores on each Big Five personality factor. Furthermore, the mean and standard deviation of the respective dialogue quality is calculated for all groups.

In section 3.4.2 all assumptions were described which have to be fulfilled for using the Wilcoxon rank sum test. In the following, it is verified whether these assumptions are all met.

1. The dependent variable (the respective dialogue quality) is continuous
2. The independent variable (the respective personality trait) is dichotomous because of the classification in Low/High
3. The third assumption is the independence of observations. This is given because no player can be in the "Low" group as well as in the "High" group at the same time. Furthermore, the classification of a player is not influenced by any other player. Therefore, the independence of observations is given.

All assumptions are fulfilled, and the Wilcoxon Rank Sum test can be used. In the following, the calculated dialogue qualities are described in detail.

Relative Speech Time

This quality is intended to examine whether players with a certain personality factor speak significantly more or less during a dialogue. The relative speech time of a player is calculated by dividing the sum of the duration of all turns of a player by the total duration of the dialogue. The sum of the duration of a player's turns is calculated by filtering the preprocessed CSV file (as described in section 3.2) for the ID of the respective player (including speech activity with overlap). Then, the sum of the DURATION2 column is calculated. The total duration of the dialogue is calculated by filtering for the max END2 value in the CSV file. A significant difference in personality factors could help dialogue systems predict the speech activity of a person.

$$\text{Relative Speech Time} = \frac{\text{Sum of the duration of all turns of a player}}{\text{Total duration of the dialogue}} \tag{1}$$

Average Time Between Turns

This quality is intended to examine how much time players with certain personality factors between their speech activities. This quality is calculated by subtracting the END2 value of a turn from the START2 value of that player's next turn. A significant difference in personality factors could help dialogue systems predict how likely a person's speech activity is at a given time.

$$\text{Average Time Between Turns} = \frac{\text{Sum of difference between turns of a player}}{\text{Total turns of a player}} \tag{2}$$

Average Turn Duration

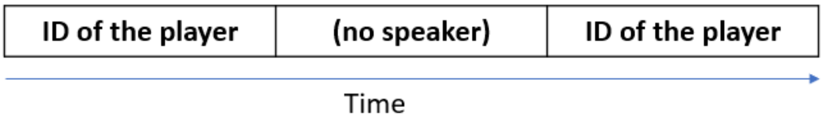
This quality is intended to examine whether players with a certain personality factor tend to take longer/shorter turns. The average turn duration of a player is calculated by filtering the dataset for the ID of the respective player and then taking the mean of the DURATION2 column. A significant difference in personality factors could help dialogue systems predict when a person will finish their turn.

$$\text{Average Turn Duration} = \frac{\text{Sum of the duration of all turns of a player}}{\text{Total turns of a player}} \tag{3}$$

Number of Pauses

This quality is intended to examine whether players with a certain personality factor take more pauses during a dialogue. [Sacks et al., 1974] define a pause as a pattern where a duration of silence happens *within* the turn of a speaker. This can happen when a speaker stops speaking for a while but does not yield the turn or when a speaker stops speaking and no one else is willing to take the turn, so the speaker continues their turn after a while. The pattern searched for in the dataset to calculate this quality is illustrated in figure 4.1. To avoid overstating this metric by players who talk more/make more turns, the player's number of pauses is divided by that player's total number of turns. A significant difference in personality factors could help dialogue systems predict whether a speaker intends to yield their turn or not after a duration of silence.

Figure 4.1: Pattern by which the total number of pauses of a player in the dataset is calculated

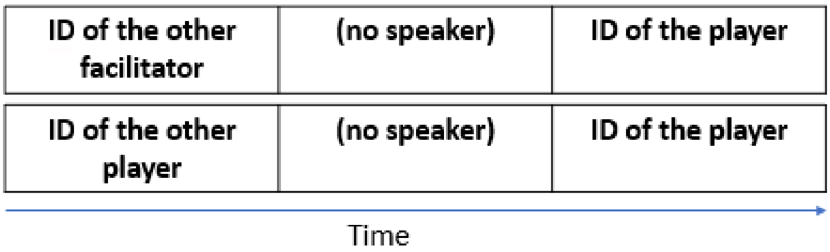


$$\text{Number of Pauses} = \frac{\text{Number of pauses taken by a player}}{\text{Total turns of a player}} \tag{4}$$

Number of Left Gaps

This quality is intended to examine whether players with a certain personality factor leave more gaps when taking their turn. [Sacks et al., 1974] define a gap as a pattern in which a duration of silence occurs between turns by two different speakers. The pattern searched for in the dataset to calculate this quality is illustrated in figure 4.2. To avoid overstating this metric by players who talk more/make more turns, the player's number of gaps is divided by that player's total number of turns. A significant difference in personality factors could help to better understand the turn-taking style of people with these personality factors.

Figure 4.2: The two patterns by which the total number of gaps left by a player in the dataset is calculated



ID of the other facilitator	(no speaker)	ID of the player
ID of the other player	(no speaker)	ID of the player

$$\text{Number of Left Gaps} = \frac{\text{Total gaps left by a player}}{\text{Total turns of a player}} \tag{5}$$

Average Pause Duration

This quality is intended to examine whether players with a certain personality factor tend to take longer/shorter pauses. The same definition of a pause as above applies. This quality is calculated by taking the average of the DURATION2 column for the pauses that were identified above. A significant difference between personality traits could be useful information for a spoken dialogue system to decide when a person will start speaking again after they take a pause.

$$\text{Average Pause Duration} = \frac{\text{Sum of duration of pauses of a player}}{\text{Total number of pauses by a player}} \tag{6}$$

Average Gap Duration

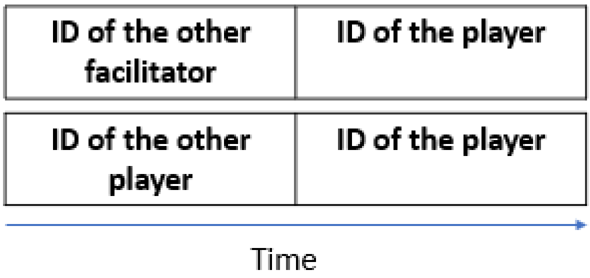
This quality is intended to examine whether players with a certain personality factor tend to leave longer/shorter gaps. The same definition of a gap as above applies. This quality is calculated by taking the average of the DURATION2 column for the gaps that were identified above. A significant difference between personality traits could be helpful information for a spoken dialogue system to predict the speech activity of a person after a gap.

$$\text{Average Gap Duration} = \frac{\text{Sum of duration of gaps left by a player}}{\text{Total number of gaps left by a player}} \tag{7}$$

Turns Taken with no Gap and no Overlap

This quality is intended to examine whether players with a certain personality factor tend to take turns with no gap and no overlap rather than leaving a gap. The pattern searched for in the dataset to calculate this quality is illustrated in figure 4.3. A significant difference in personality factors could help to better understand the turn-taking style of people with these personality factors.

Figure 4.3: The two patterns by which the total number of gaps left by a player in the dataset is calculated

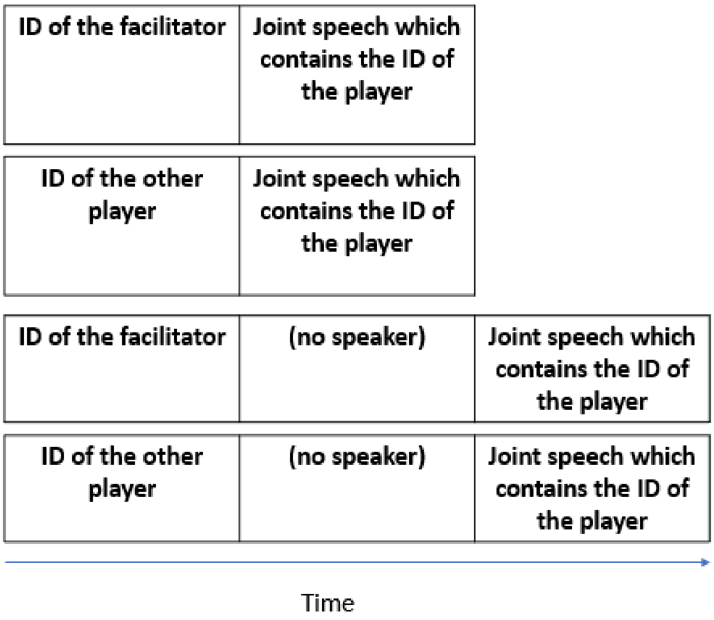


$$\text{Turns with No Gap No Overlap} = \frac{\text{Turns taken by a player with no gap and no overlap}}{\text{Total number of turns by a player}} \qquad (8)$$

Caused Interruptions

This quality is intended to examine whether players with a certain personality factor tend to cause more/fewer interruptions in dialogue. A caused interruption is defined as a pattern in which a player engages in joint speech after the turn of a different player. The patterns searched for in the dataset to calculate this quality are illustrated in figure 4.4. To avoid overstating this metric for players with longer dialogues, the number of interruptions caused is divided by the total length of the dialogue in question. A significant difference in personality factors could help to better understand the turn-taking style of people with these personality factors.

Figure 4.4: The four patterns in the dataset which are counted as a "caused interruption"

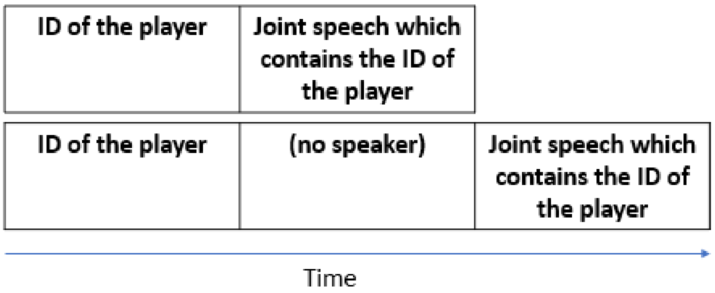

ID of the facilitator
Joint speech which contains the ID of the player
ID of the other player
Joint speech which contains the ID of the player
ID of the facilitator
(no speaker)
Joint speech which contains the ID of the player
ID of the other player
(no speaker)
Joint speech which contains the ID of the player
Time

$$\text{Caused Interruptions} = \frac{\text{Interruptions by a player}}{\text{Total duration of the dialogue}} \tag{9}$$

Interruptions

This quality is intended to examine whether players with a certain personality factor tend to get interrupted more/less. An interruption is defined as a pattern where a player has a solo turn followed by a turn with joint speech (containing the ID of the player). Furthermore, an interruption is defined as a pattern in which joint speech (containing the ID of the player) follows a turn of the player and a duration of silence. This indicates that the player was intending to take a pause, but an interlocutor did not understand that the turn of the player was not finished. The pattern searched for in the dataset to calculate this quality is illustrated in figure 4.5. The player examined can be understood as the role of the user of a spoken dialogue system. A significant difference in personality factors could therefore help to better understand the turn-taking style of individuals in order to lessen the amount of interruptions caused by a spoken dialogue system.

Figure 4.5: The two patterns in the dataset which are counted as an instance of "getting interrupted"



ID of the player	Joint speech which contains the ID of the player	
ID of the player	(no speaker)	Joint speech which contains the ID of the player

$$\text{Interruptions} = \frac{\text{Instances in which a player gets interrupted}}{\text{Total duration of the dialogue}} \tag{10}$$

Total Laughter Duration

This quality is intended to examine whether players with a certain personality factor tend to laugh more during dialogue than other players. This quality is calculated by summing up the DURATION2 column for the laughter sections of the respective player in the dataset. To avoid overstating this metric by players with longer dialogues, this metric is divided by the total duration of the dialogue.

$$\text{Total Laughter Duration} = \frac{\text{Sum of the duration of all laughter of a player}}{\text{Total duration of the dialogue}} \tag{11}$$

Average Laughter Duration

This quality is intended to examine whether players with a certain personality factor tend to engage in longer/shorter periods of laughter. This quality is calculated by taking the average of the DURATION2 column for the laughter sections of the respective player. Players with no laughter are not considered in this quality. A significant difference in personality factors could help to better predict the end of speech activity for a person.

$$\text{Average Laughter Duration} = \frac{\text{Total duration of laughter of a player}}{\text{Total instances of laughter of a player}} \tag{12}$$

Ratio of Discourse to Mirthful Laughter

This quality is intended to examine whether players with a certain personality factor tend to engage more in discourse or mirthful laughter. Players with no laughter are not considered in this quality. This quality is calculated by counting both the instances of "mirthful" and "discourse" laughter by examining the ACTION column in the dataset. A significant difference in personality factors could help to better understand the communication style for a person.

$$\text{Discourse to Mirthful Laughter} = \frac{\text{Instances of discourse laughter of a player}}{\text{Instances of mirthful laughter of a player}} \tag{13}$$

Ratio of Solo to Shared Laughter

This quality is intended to examine whether players with a certain personality factor tend to engage more in solo than shared laughter. Players with no laughter are not considered in this quality. This quality is calculated by counting both the instances of "solo" and "shared" laughter by examining the ACTION column in the dataset. A significant difference in personality factors could help to better understand the communication style for a person.

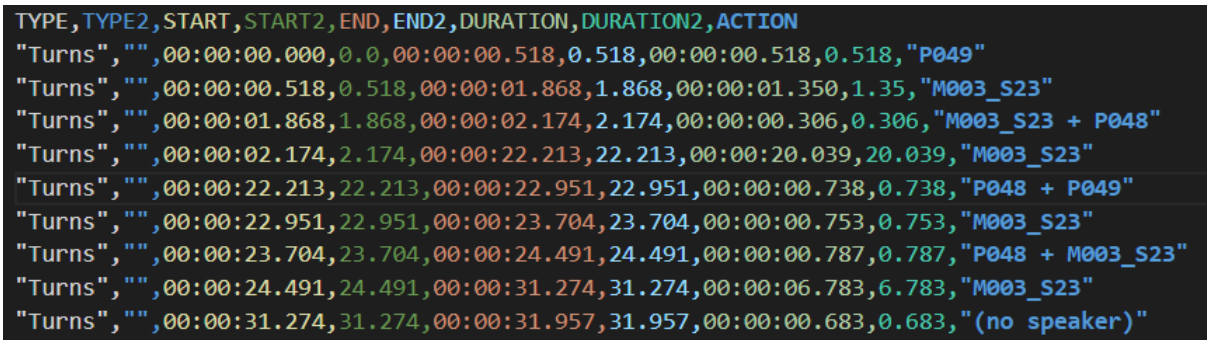
$$\text{Solo to Shared Laughter} = \frac{\text{Instances of solo laughter of a player}}{\text{Instances of shared laughter of a player}} \tag{14}$$

4.2 Turn-Taking Model Based on Nonverbal Features

This experiment aims to analyse how a continuous turn-taking model performs that utilises nonverbal qualities in dialogue as features. As it was discussed in 2.2.2, continuous turn-taking models are state of the art. A logistic regression model is built that predicts future speech activity in the dialogues of the MULTISIMO dataset. For this, the task of predicting future speech activity is modelled as a time series prediction problem. The task of predicting speech activity can be understood as making turn taking decisions, since by predicting the next speaker for every time segment the model predicts whether a speaker is holding or yielding their turn. This experiment addresses research question **Q3**. In order to build the logistic regression model for the turn-taking predictions following approach is proposed.

First, the dialogue CSV files are filtered for the rows which contain the "Turns" annotations (all rows which have the value "Turns" in the TYPE column). These annotations contain the segmentation and timing of the participants' turns. The ACTION column contains the ID of the participant who is currently speaking. IDs starting with a "P0" stand for players, IDs starting with a "M0" stand for the facilitators and "(no speaker)" indicates a segment of silence. Furthermore, turns which contain a "+" (such as "P048 + P049") indicate that there was an overlap between speakers. ACTION is the target value which is meant to be predicted by the logistic regression model. Figure 4.6 shows the structure of the "Turns" annotations in a CSV file. The START2 and END2 columns indicate the start and end time of the respective turn.

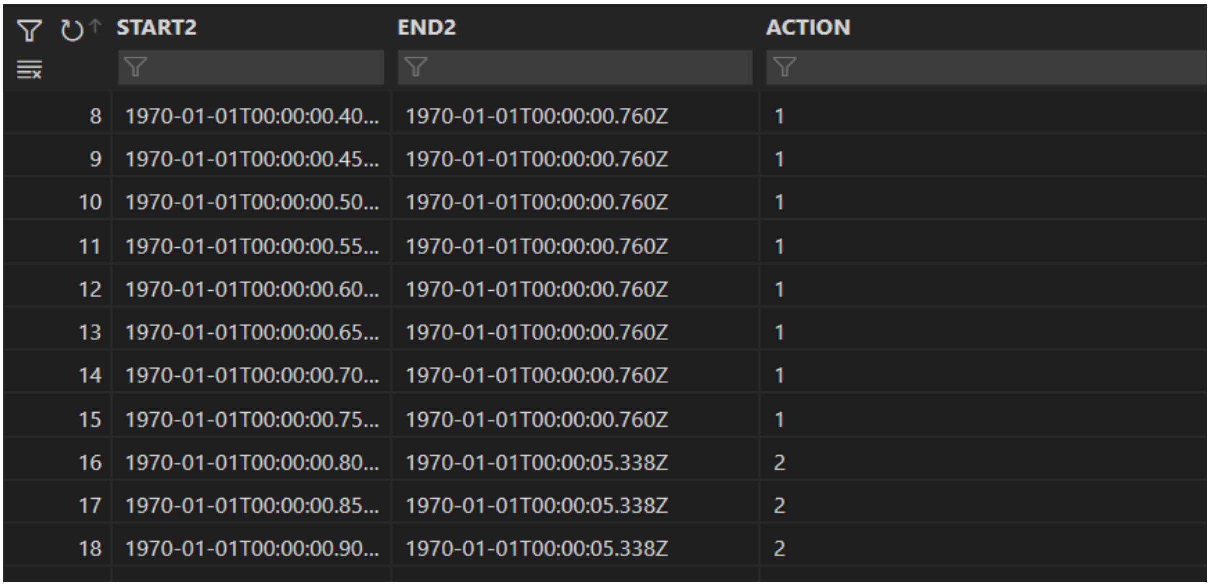
Figure 4.6: Structure of the "Turns" annotations in a CSV file



```
TYPE,TYPE2,START,START2,END,END2,DURATION,DURATION2,ACTION
"Turns","",00:00:00.000,0.0,00:00:00.518,0.518,00:00:00.518,0.518,"P049"
"Turns","",00:00:00.518,0.518,00:00:01.868,1.868,00:00:01.350,1.35,"M003_S23"
"Turns","",00:00:01.868,1.868,00:00:02.174,2.174,00:00:00.306,0.306,"M003_S23 + P048"
"Turns","",00:00:02.174,2.174,00:00:22.213,22.213,00:00:20.039,20.039,"M003_S23"
"Turns","",00:00:22.213,22.213,00:00:22.951,22.951,00:00:00.738,0.738,"P048 + P049"
"Turns","",00:00:22.951,22.951,00:00:23.704,23.704,00:00:00.753,0.753,"M003_S23"
"Turns","",00:00:23.704,23.704,00:00:24.491,24.491,00:00:00.787,0.787,"P048 + M003_S23"
"Turns","",00:00:24.491,24.491,00:00:31.274,31.274,00:00:06.783,6.783,"M003_S23"
"Turns","",00:00:31.274,31.274,00:00:31.957,31.957,00:00:00.683,0.683,"(no speaker)"
```

The START2 and END2 columns are converted to *datetime* format and the index of the CSV file is set to START2. In the next step, the index of the CSV file is re-sampled to a frequency of 50 ms. The resulting NaN values in the dataframe are filled using the forward fill method. This means, the last valid value is propagated forward. On average, the files have a row count of 11205 after the re-sampling. Lastly, the IDs of the participants are replaced with numerical values to make them processable for a machine learning model. Overlaps in speech between players is coded with the same numerical value regardless of which players overlap in speech. Lastly, the TYPE, TYPE2, START, END, DURATION and DURATION2 columns are dropped since they have no further use for the machine learning model. Figure 4.7 shows a preprocessed CSV file after the steps mentioned above have been applied.

Figure 4.7: Preprocessed CSV file for use in the machine learning model

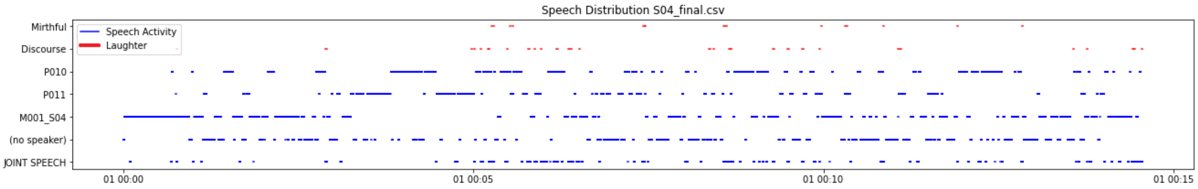


	START2	END2	ACTION
8	1970-01-01T00:00:00.40...	1970-01-01T00:00:00.760Z	1
9	1970-01-01T00:00:00.45...	1970-01-01T00:00:00.760Z	1
10	1970-01-01T00:00:00.50...	1970-01-01T00:00:00.760Z	1
11	1970-01-01T00:00:00.55...	1970-01-01T00:00:00.760Z	1
12	1970-01-01T00:00:00.60...	1970-01-01T00:00:00.760Z	1
13	1970-01-01T00:00:00.65...	1970-01-01T00:00:00.760Z	1
14	1970-01-01T00:00:00.70...	1970-01-01T00:00:00.760Z	1
15	1970-01-01T00:00:00.75...	1970-01-01T00:00:00.760Z	1
16	1970-01-01T00:00:00.80...	1970-01-01T00:00:05.338Z	2
17	1970-01-01T00:00:00.85...	1970-01-01T00:00:05.338Z	2
18	1970-01-01T00:00:00.90...	1970-01-01T00:00:05.338Z	2

In the next step, features are constructed for the logistic regression model. First, *lag-features* are created for the ACTION column. These features consider the past value of the target variable (in this case the ACTION column). The target value from n steps ago is examined to predict future values of the target variable. In this case, lag-features from n=1 to n=20 are created (sample rate is 50 ms, so for example the lag-feature with n=20 considers the target value from one second ago). Lag-features have the limitation that they can't predict sudden changes in the target value. Predictions that are purely based on these features tend to lag behind the actual value. Therefore, other features have to be created. Next, the "consecutive" feature is created which counts how many time steps ago the target value last changed its value. This feature is intended to capture how long the current turn is lasting. Lastly, *rolling-features* are created for each unique value in the ACTION column and for time windows of 0.3 seconds, 0.5 seconds, 1 second, 3 seconds and 5 seconds. These features, consider the rolling count for the respective value in the ACTION column in the respective time frame. For example, the rolling feature for value 1 (player 1 speaking) in a time frame of 3 seconds counts, how many target values in the past 3 seconds were of value "1". In total, 47 feature columns were created after these preprocessing steps. Since values up to five seconds from the past are considered as features, the first five seconds of each dialogue have feature columns with "NaN" values and are therefore dropped. The features are then iteratively pruned using the RFE algorithm described in section 3.3.2. The performance of the final model is then reported.

The same preprocessing steps as above are applied for the laughter and gaze annotations (these are only available for two sessions) to construct features for laughter and gaze. For the laughter features, the annotations from the laughter sections and the non-laughter sections are concatenated to one dataframe. All unique actions (mirthful laughter, discourse laughter, no laughter) are coded as numerical values. In the next step, the features are constructed for the laughter annotations in the same way as for the speech distribution. The same procedure applies for the gaze annotations. For each participant, the gaze annotations are encoded in numerical values. On this basis, the gaze features are created as described above for the speech distributions. Figure 4.8 shows a visualisation of speech activity and laughter in one of the dialogues after the preprocessing steps have been completed. This activity should be predicted by the constructed machine learning model.

Figure 4.8: Visualisation of speech activity in a dialogue



5 Results

This chapter presents and discusses the results of the experiments that were introduced in chapter 4. First, the results of the experiments described in section 4.1 are presented, followed by the results of the experiments described in section 4.2.

5.1 Influence of Big Five Personality Traits on Dialogue Qualities

For each quality, the mean and standard deviation are reported for the total population of players as well as for the groups filtered by the low/high value of the respective personality trait. In addition, the two-sided Wilcoxon Rank Sum test results are given for each comparison of the Low/High groups for each personality factor. Results with a p-value $<$ 0.01 are considered highly significant (denoted by a ***), results with a p-value $<$ 0.05 are considered significant (denoted by a **) and results with a p-value $<$ 0.1 are considered to be approaching significance (denoted by a *). The sample size for each inspected quality is 34. Lastly, the results are interpreted for each quality.

Relative Speech Time

Table 5.1 shows that the openness factor has a statically significant influence on the relative speech time quality. Table 5.2 shows that players with a higher openness tend to speak more than players with a low openness which seems intuitive according to the definition of openness in the Big Five model. Furthermore, it is interesting to note that the extroversion factor does not seem to make a difference. It is also interesting to note that the overall mean for the relative speech time is 0.331, which suggests that all participants (including the facilitator) in the dialogue contribute equally to the conversation on average.

Table 5.1: Results of the two-sided Wilcoxon Rank Sum test comparing the low/high groups for each personality factor for relative speech time.

	EXT	AGR	NEU	CON	OPE
p-value	0.9038	0.9862	0.5789	0.2926	0.0113**
U	148.0	145.0	159.0	113.0	216.0

Table 5.2: Means and standard deviations grouped by high/low values of the respective personality factor for relative speech time

		EXT		AGR	
	Overall	High	Low	High	Low
Mean	0.331	0.332	0.330	0.334	0.328
Std. Dev.	0.079	0.074	0.087	0.086	0.075

	NEU		CON		OPE	
	High	Low	High	Low	High	Low
Mean	0.338	0.323	0.319	0.345	0.361	0.293
Std. Dev.	0.085	0.073	0.081	0.077	0.082	0.059

Average Time Between Turns

Table 5.3 shows that the difference in the average time between turns is approaching significance for the openness factor. Table 5.4 shows that players with a high openness score tend to take turns in shorter succession than players with a low openness score. This result is in line with the description of the openness trait in the Big Five model. Again, the extroversion trait does not seem to have an impact on this quality.

Table 5.3: Results of the two-sided Wilcoxon Rank Sum test comparing the low/high groups for each personality factor for average time between turns.

	EXT	AGR	NEU	CON	OPE
p-value	0.9312	0.5691	0.8896	0.4581	0.0564*
U	141.0	127.0	138.0	156.0	87.0

Table 5.4: Means and standard deviations grouped by high/low values of the respective personality factor for average time between turns

		EXT		AGR	
	Overall	High	Low	High	Low
Mean	2.143	2.117	2.173	2.017	2.255
Std. Dev.	0.821	0.821	0.847	0.696	0.924

	NEU		CON		OPE	
	High	Low	High	Low	High	Low
Mean	2.128	2.162	2.224	2.053	1.920	2.426
Std. Dev.	0.846	0.817	0.800	0.861	0.816	0.762

Average Turn Duration

The data does not support an influence of any personality factor on the average turn duration of a player. Table 5.6 shows that the means are relatively similar for all examined groups. Furthermore, it is interesting to note that the average turn duration for the players is only one second which seems very short. This can be explained by the game setting in which the dialogues were conducted.

Table 5.5: Results of the two-sided Wilcoxon Rank Sum test comparing the low/high groups for each personality factor for average turn duration.

	EXT	AGR	NEU	CON	OPE
p-value	0.8766	0.1840	0.4878	0.3427	0.2670
U	139.0	105.0	163.0	116.0	175.0

Table 5.6: Means and standard deviations grouped by high/low values of the respective personality factor for average turn duration

		EXT		AGR	
	Overall	High	Low	High	Low
Mean	1.013	1.010	1.015	0.974	1.046
Std. Dev.	0.161	0.186	0.132	0.114	0.190

	NEU		CON		OPE	
	High	Low	High	Low	High	Low
Mean	1.029	0.992	0.998	1.029	1.030	0.990
Std. Dev.	0.168	0.155	0.162	0.163	0.141	0.186

Number of Pauses

No significance can be observed for the number of pauses taken. This is mainly due to the fact that the standard deviation is very high for all groups. This could hint that the way the metric was calculated is not appropriate for the dataset. In order to examine this quality in more detail, a narrower definition of a "pause" might be required in terms of timing. Also, it might be possible that the annotations of the dataset are not accurate enough to capture very short pauses during speech.

Table 5.7: Results of the two-sided Wilcoxon Rank Sum test comparing the low/high groups for each personality factor for the number of pauses.

	EXT	AGR	NEU	CON	OPE
p-value	0.7731	0.6655	0.4663	0.3954	1.0000
U	144.5	123.5	164.0	111.0	133.5

Table 5.8: Means and standard deviations grouped by high/low values of the respective personality factor for number of pauses

		EXT		AGR	
	Overall	High	Low	High	Low
Mean	0.052	0.055	0.048	0.046	0.057
Std. Dev.	0.038	0.040	0.036	0.031	0.043

	NEU		CON		OPE	
	High	Low	High	Low	High	Low
Mean	0.058	0.044	0.043	0.062	0.049	0.056
Std. Dev.	0.047	0.021	0.030	0.044	0.031	0.046

Average Pause Duration

No significance can be observed for the average pause duration. [Heldner and Edlund, 2010] report an average pause duration of 730 ms for English dialogue. This is notably shorter than the average pause duration that was found in this experiment. Table 5.10 shows that for all players, the average pause duration is 1363 ms. First, this could confirm the assumption made before that the annotations are not accurate enough to capture such short periods of silence. Also, this result could be attributed to the game mode in which the dialogues were recorded. Since the players are taking a quiz, they just might need time to think for appropriate answers, rather than taking a pause *during* speech. For future experiments, it might therefore be helpful to define *pauses* differently for this dataset.

Table 5.9: Results of the two-sided Wilcoxon Rank Sum test comparing the low/high groups for each personality factor for average pause duration.

	EXT	AGR	NEU	CON	OPE
p-value	0.8359	0.2407	0.5554	0.2477	0.6149
U	150.5	178.5	125.0	110.0	127.5

Table 5.10: Means and standard deviations grouped by high/low values of the respective personality factor for average pause duration

		EXT		AGR	
	Overall	High	Low	High	Low
Mean	1.324	1.464	1.166	1.456	1.207
Std. Dev.	0.638	0.812	0.312	0.737	0.528

	NEU		CON		OPE	
	High	Low	High	Low	High	Low
Mean	1.159	1.533	1.239	1.419	1.381	1.252
Std. Dev.	0.331	0.857	0.587	0.696	0.806	0.337

Number of Gaps

Table 5.11 shows that the extroversion factor has an influence on the number of gaps a player leaves. Table 5.12 shows that players with a high extroversion factor leave more gaps than players with a low extroversion factor. This result is a bit surprising since it seems more intuitive that more introverted players leave more gaps.

Table 5.11: Results of the two-sided Wilcoxon Rank Sum test comparing the low/high groups for each personality factor for the number of gaps.

	EXT	AGR	NEU	CON	OPE
p-value	0.0338**	0.1086	0.1550	0.7430	0.3858
U	206.0	97.0	101.0	154.0	117.0

Table 5.12: Means and standard deviations grouped by high/low values of the respective personality factor for number of gaps

		EXT		AGR	
	Overall	High	Low	High	Low
Mean	0.084	0.096	0.071	0.074	0.092
Std. Dev.	0.035	0.033	0.034	0.032	0.036

	NEU		CON		OPE	
	High	Low	High	Low	High	Low
Mean	0.076	0.094	0.085	0.082	0.078	0.091
Std. Dev.	0.033	0.036	0.039	0.031	0.034	0.036

Average Gap Duration

This quality shows no significant difference between the groups. [Heldner and Edlund, 2010] reports an average gap duration of 400-600 ms in English dialogues which is notably shorter than the average gap duration observed in the dataset. This could be due to similar reasons as for the duration of the pauses.

Table 5.13: Results of the two-sided Wilcoxon Rank Sum test comparing the low/high groups for each personality factor for te average duration of gaps

	EXT	AGR	NEU	CON	OPE
p-value	0.8766	0.3605	0.2980	0.9312	0.7549
U	149.0	117.0	173.0	141.0	133.0

Table 5.14: Means and standard deviations grouped by high/low values of the respective personality factor for average gap duration

		EXT		AGR	
	Overall	High	Low	High	Low
Mean	1.363	1.404	1.316	1.267	1.448
Std. Dev.	0.556	0.616	0.495	0.558	0.556

	NEU		CON		OPE	
	High	Low	High	Low	High	Low
Mean	1.387	1.333	1.364	1.361	1.330	1.405
Std. Dev.	0.386	0.732	0.591	0.534	0.564	0.562

Turns Taken with no Gap and no Overlap

This quality shows no significant difference between the groups. Furthermore, the means are very similar for each group. This means that the data does not support any evidence that players with a certain personality type prefer taking turns with no gap and no overlap more than other players.

Table 5.15: Results of the two-sided Wilcoxon Rank Sum test comparing the low/high groups for each personality factor for turns taken with no gap and no overlap

	EXT	AGR	NEU	CON	OPE
p-value	0.6413	0.7957	0.5438	1.0000	0.5908
U	130.0	152.0	160.5	144.0	158.5

Table 5.16: Means and standard deviations grouped by high/low values of the respective personality factor for amount of turns taken with no gap and no overlap

		EXT		AGR	
	Overall	High	Low	High	Low
Mean	0.166	0.163	0.171	0.169	0.165
Std. Dev.	0.041	0.042	0.040	0.035	0.046

	NEU		CON		OPE	
	High	Low	High	Low	High	Low
Mean	0.172	0.159	0.164	0.169	0.167	0.166
Std. Dev.	0.046	0.033	0.037	0.046	0.036	0.047

Caused Interruptions

This quality shows no significant difference between the groups. The mean for the "high" agreeableness group is higher than for the "low" agreeableness group, which seems counter-intuitive according to the definition of the Big Five traits. Also, the mean for caused interruptions is higher for the players with a high openness compared to players with low openness. This result seems to be in line with the previous results that players with high openness have more speech activity and leave less time between their turns. Therefore, it becomes more likely that they interrupt other participants. However, 5.17 shows no statistical significance, so the results should not be over-interpreted.

Table 5.17: Results of the two-sided Wilcoxon Rank Sum test comparing the low/high groups for each personality factor for caused interruptions

	EXT	AGR	NEU	CON	OPE
p-value	0.8225	0.6290	0.4249	0.8766	0.4351
U	151.0	158.5	119.0	139.0	165.5

Table 5.18: Means and standard deviations grouped by high/low values of the respective personality factor for average amount of caused interruptions

		EXT		AGR	
	Overall	High	Low	High	Low
Mean	0.126	0.127	0.125	0.133	0.120
Std. Dev.	0.042	0.048	0.036	0.034	0.049

	NEU		CON		OPE	
	High	Low	High	Low	High	Low
Mean	0.124	0.129	0.127	0.125	0.131	0.120
Std. Dev.	0.044	0.042	0.042	0.043	0.039	0.047

Interruptions

Table 5.19 shows a significant difference in interruptions for the openness factor. Table 5.22 shows that players with high openness are more likely to be interrupted than players with low openness. This result is in line with the previous result that people with a high openness generally speak more throughout the dialogue. Also, according to the documentation of the MULTISIMO dataset, it is the job of the facilitator to ensure both players contribute equally to the dialogue. By speaking more, players with high openness might be more likely to force a reaction by the facilitator.

Table 5.19: Results of the two-sided Wilcoxon Rank Sum test comparing the low/high groups for each personality factor for interruptions

	EXT	AGR	NEU	CON	OPE
p-value	0.9038	0.5459	0.3144	0.9038	0.0201**
U	148.0	162.0	113.0	140.0	210.0

Table 5.20: Means and standard deviations grouped by high/low values of the respective personality factor for average amount of times a player got interrupted

		EXT		AGR	
	Overall	High	Low	High	Low
Mean	0.046	0.046	0.046	0.047	0.045
Std. Dev.	0.018	0.018	0.019	0.017	0.020

	NEU		CON		OPE	
	High	Low	High	Low	High	Low
Mean	0.044	0.049	0.045	0.047	0.053	0.037
Std. Dev.	0.020	0.017	0.015	0.022	0.019	0.014

Total Laughter Duration

This quality shows no significant difference between the groups. Also, the means are very high for all groups. Therefore, this data does not provide any evidence that the total duration of laughter is influenced by the personality of the player.

Table 5.21: Results of the two-sided Wilcoxon Rank Sum test comparing the low/high groups for each personality factor for total laughter duration

	EXT	AGR	NEU	CON	OPE
p-value	0.2751	0.7339	0.9043	0.7819	0.249
U	130.5	118.0	120.5	112.5	89.5

Table 5.22: Means and standard deviations grouped by high/low values of the respective personality factor for total laughter duration

		EXT		AGR	
	Overall	High	Low	High	Low
Mean	0.033	0.034	0.031	0.031	0.034
Std. Dev.	0.021	0.020	0.022	0.019	0.023

	NEU		CON		OPE	
	High	Low	High	Low	High	Low
Mean	0.032	0.033	0.033	0.032	0.030	0.036
Std. Dev.	0.020	0.022	0.021	0.020	0.019	0.022

Average Laughter Duration

Similarly to the total laughter duration, this quality does not show a significant difference between the groups. This also supports the assumption that the personality type of the players does not influence laughter.

Table 5.23: Results of the two-sided Wilcoxon Rank Sum test comparing the low/high groups for each personality factor for average laughter duration

	EXT	AGR	NEU	CON	OPE
p-value	0.6591	0.2634	0.8822	0.4927	0.8709
U	106.0	78.0	112.0	95.0	92.0

Table 5.24: Means and standard deviations grouped by high/low values of the respective personality factor for average laughter duration

		EXT		AGR	
	Overall	High	Low	High	Low
Mean	0.928	0.921	0.937	0.906	0.949
Std. Dev.	0.191	0.170	0.219	0.195	0.192

	NEU		CON		OPE	
	High	Low	High	Low	High	Low
Mean	0.972	0.872	0.900	0.961	0.928	0.929
Std. Dev.	0.180	0.198	0.176	0.209	0.174	0.218

Ratio of Discourse to Mirthful Laughter

This quality shows no significant difference between the groups. The standard deviations are very high for all groups. By removing outliers, it might be possible to construct a statistical significance for the openness factor.

Table 5.25: Results of the two-sided Wilcoxon Rank Sum test comparing the low/high groups for each personality factor for the ratio discourse/mirthful laughter

	EXT	AGR	NEU	CON	OPE
p-value	0.8798	0.4495	0.2236	0.7479	0.1019
U	132.0	107.0	158.5	118.5	169.5

Table 5.26: Means and standard deviations grouped by high/low values of the respective personality factor for the ratio discourse/mirthful laughter

		EXT		AGR	
	Overall	High	Low	High	Low
Mean	0.945	0.982	0.903	0.845	1.033
Std. Dev.	0.989	1.002	1.007	0.886	1.090

	NEU		CON		OPE	
	High	Low	High	Low	High	Low
Mean	1.104	0.740	0.932	0.960	1.082	0.769
Std. Dev.	1.106	0.806	0.994	1.017	1.052	0.907

Ratio of Solo to Shared Laughter

This quality shows no significant difference between the groups. The standard deviations are very high for all groups. By removing outliers it might be possible to construct a statistical significance for the openness factor and the extroversion factor. The mean of the "High" extroversion group is considerably lower than for the "Low" group, suggesting that highly extroverted players rather engage in solo than shared laughter. This seems reasonable since these individuals are characterised as outgoing and very sociable according to the Big Five model. However, the high standard deviation does not allow any statistically relevant statements.

Table 5.27: Results of the two-sided Wilcoxon Rank Sum test comparing the low/high groups for each personality factor for the ratio solo/shared laughter

	EXT	AGR	NEU	CON	OPE
p-value	0.3630	0.9697	0.7030	0.4599	0.1422
U	103.0	126.0	115.5	107.5	87.0

Table 5.28: Means and standard deviations grouped by high/low values of the respective personality factor for the ratio solo/shared laughter

		EXT		AGR	
	Overall	High	Low	High	Low
Mean	1.557	1.296	1.852	1.514	1.594
Std. Dev.	1.057	0.511	1.414	0.901	1.204

	NEU		CON		OPE	
	High	Low	High	Low	High	Low
Mean	1.619	1.477	1.435	1.694	1.532	1.589
Std. Dev.	1.320	0.610	0.883	1.243	1.338	0.566

5.1.1 Summary

The results suggest an influence of the openness trait on the relative speech time, the average time between turns and the interruptions quality. Furthermore, the results suggest an influence of the extroversion trait on the number of left gaps in dialogue. Conversely, these results might hint that these qualities are good predictors for the respective personality trait. No significant difference could be found for the remaining qualities. It is interesting to note that the openness trait shows the most influence on the calculated qualities. In section 2.3 it was discussed that a previously conducted literature review by [Ahmad et al., 2022] showed the least found dialogue cues for the openness factor. Furthermore, some qualities showed very high standard deviations (e.g., the qualities related to laughter). For future experiments, it might be useful to think of an automated method to remove outliers in the data. For example, it is common to remove values outside the range of three standard deviations for data that follows a normal distribution. Lastly, the definitions of a pause and a gap might have to be reconsidered since their average duration was significantly different from the values reported in the literature.

5.2 Turn-Taking Model Based on Nonverbal Features

This section presents the results from the experiment described in section 4.2 and compares them against a last-known value baseline. First, the logistic regression model is trained to predict general speech activity for all values from the ACTION column (silence, facilitator speaking, player 1 speaking, player 2 speaking, joint speech). After that, the logistic regression model is trained to predict only the speech activity of one participant (the facilitator) in the dialogue as a binary classification problem. For all models, an 80-20 train-test split is applied which is a commonly accepted standard value.

The last known value is used as a baseline predictor. This means that the last known value is projected n steps into the future to make a prediction. For example, if the last known value is "player 1 speaking", the baseline would predict that in n steps into the future, player 1 is still speaking.

5.2.1 Speech Activity Prediction

First, the performance of the model is examined with only features from the speech distribution. Accuracy is used as a measure of performance since speech activity is about balanced, as was shown in the previous experiment. Table 5.29 shows the accuracy of the model for 250 ms in the future, table 5.30 shows the accuracy for 1 s in the future and table 5.31 shows the accuracy for 2 s in the future. All three tables show that the performance of the logistic regression model is very close to the baseline. Overall, the accuracy decreases the more time steps are predicted in the future. This is expected since it becomes increasingly harder to predict values in the future. For session 23, the results of the logistic regression model are consistently worse than the baseline. In a few cases (e.g., S14 and S22 for 1 s in the future), the accuracy of the logistic regression model is slightly better than the baseline, but not significantly. Adding the laughter features has a slightly positive effect. The models used an "l1" regularisation with a C parameter of 100. Adjusting the C parameter of the model to control regularisation does not seem to improve the accuracy of the model. For feature selection, the RFE algorithm removes 10 features per iteration, and the accuracy of the model is compared for 10, 20, 30, 40 and 50 total features. The best performing model is reported respectively. Irrespective of the number of steps predicted in the future, the models largely used the lag features (50 ms - 150 ms) and rolling count features for 0.3 seconds and 0.5 seconds to make their predictions. For the predictions 1 s and 2 s into the future, the models started using the laughter features, while for the 250 ms models, the laughter features were pruned. Features that considered values from over a second ago were pruned in every trained model.

Table 5.29: Accuracy of the logistic regression model predicting speech activity 250 ms in the future. The accuracy for Log. Reg Model shows the performance based on speech distribution feature. The following row shows the performance with added laughter features.

	S02	S04	S05	S07	S08	S9	S10	S11	S13	S14	S17	S18	S19	S20	S21	S22	S23
Baseline	0.3841	0.5564	0.5022	0.4924	0.4821	0.4252	0.5316	0.4894	0.4686	0.3824	0.3811	0.4333	0.3885	0.3446	0.6052	0.2734	0.5170
Log. Reg. Model	0.3565	0.5454	0.4872	0.4903	0.3823	0.3354	0.5460	0.4717	0.4477	0.4183	0.3724	0.3712	0.3321	0.3405	0.6024	0.3009	0.4274
Log. Reg. + Laughter	0.3666	0.5560	0.4973	0.5010	0.4013	0.3544	0.5571	0.4852	0.4566	0.4212	0.3820	0.3834	0.3500	0.3519	0.6135	0.3111	0.4313

Table 5.30: Accuracy of the logistic regression model predicting speech activity 1 s in the future. The accuracy for Log. Reg Model shows the performance based on speech distribution feature. The following row shows the performance with added laughter features.

	S02	S04	S05	S07	S08	S9	S10	S11	S13	S14	S17	S18	S19	S20	S21	S22	S23
Baseline	0.3841	0.5564	0.5022	0.4924	0.4821	0.4252	0.5316	0.4894	0.4686	0.3824	0.3811	0.4333	0.3885	0.3446	0.6052	0.2734	0.5170
Log. Reg. Model	0.3565	0.5454	0.4872	0.4903	0.3823	0.3354	0.5460	0.4717	0.4477	0.4183	0.3724	0.3712	0.3321	0.3405	0.6024	0.3009	0.4274
Log. Reg. + Laughter	0.3666	0.5560	0.4973	0.5010	0.4013	0.3544	0.5571	0.4852	0.4566	0.4212	0.3820	0.3834	0.3500	0.3519	0.6135	0.3111	0.4313

Table 5.31: Accuracy of the logistic regression model predicting speech activity 2 s in the future. The accuracy for Log. Reg Model shows the performance based on speech distribution feature. The following row shows the performance with added laughter features.

	S02	S04	S05	S07	S08	S9	S10	S11	S13	S14	S17	S18	S19	S20	S21	S22	S23
Baseline	0.3372	0.4370	0.3412	0.3881	0.3848	0.2849	0.4115	0.3375	0.3888	0.4010	0.3509	0.3293	0.3216	0.2671	0.4718	0.2845	0.3774
Log. Reg. Model	0.2769	0.4391	0.2820	0.4158	0.3706	0.2625	0.4522	0.3770	0.3582	0.3999	0.3634	0.2888	0.2500	0.2761	0.4668	0.2580	0.2990
Log. Reg. + Laughter	0.2881	0.4423	0.2991	0.4236	0.3821	0.2735	0.4610	0.3826	0.3599	0.3981	0.3664	0.3012	0.2653	0.2769	0.4752	0.2634	0.3050

Figure 5.1 and 5.2 show the predicted values plotted against the actual values on the example of session 2. The figures suggest that the predictions lag behind the actual values by a constant factor with a few exceptions in which the predictions are simply false. Incidents in which the model predicts the false value rather than lagging behind the actual value are increased for the model that predicts 1 s into the future. In order to verify the assumption that the predictions lag behind the actual value by a constant factor, the predictions are shifted backwards by n steps, where n is the number of time steps originally predicted into the future. Table 5.32 shows the accuracy of the first five sessions with the shifted predictions. Indeed, the accuracy was significantly increased, confirming that the predictions lag behind the actual values with a constant offset. This means that the logistic regression model mainly propagates the last known value, which is why the accuracy observed above is so similar to the baseline.

Figure 5.1: Actual speech activity against predicted values for 250 ms in the future (on the example of session 2)

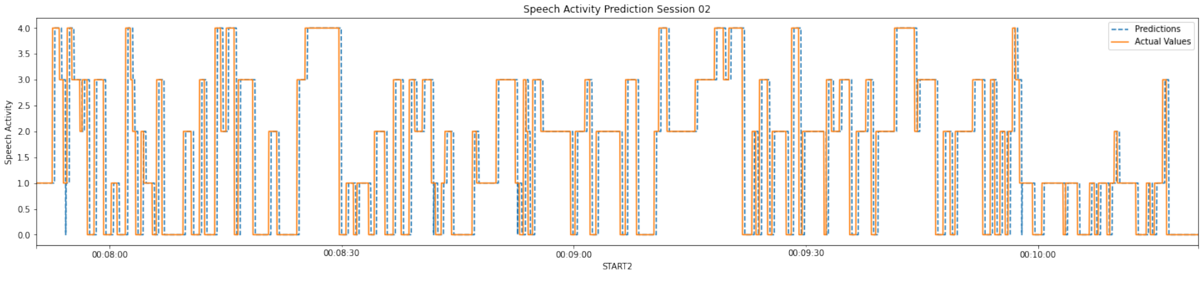


Figure 5.2: Actual speech activity against predicted values for 1 s in the future (on the example of session 2)

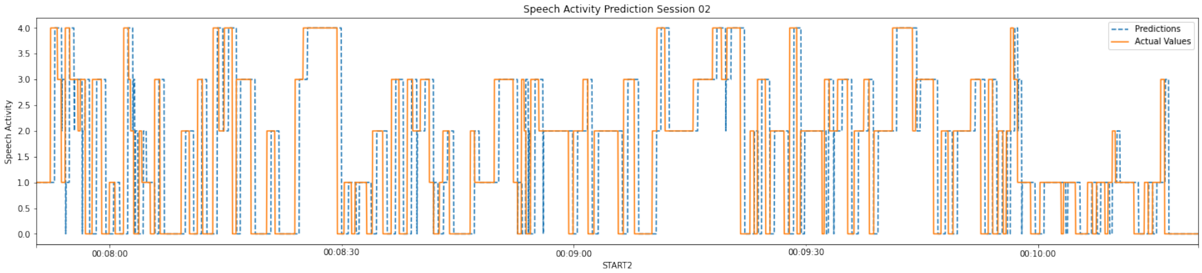


Table 5.32: Accuracy of the logistic regression model predicting speech activity 250 ms in the future after shifting the predictions

	S02	S04	S07	S08	S09
Baseline	0.7427	0.8238	0.8165	0.7783	0.7917
Log. Reg. Model	0.9418	0.9626	0.9605	0.9560	0.9577

In the following, gaze features are added to the model in an attempt to improve the predictive power of the logistic regression model. Table 5.33 shows the results of the model with added gaze features. The gaze annotations were only available for two sessions in the dataset (session 2 and session 18). For 250 ms, the performance is barely affected since the gaze features are pruned during optimisation by the RFE algorithm. For 1 s and 2 s, the model starts using the constructed gaze features. However, the performance is only minimally improved compared to the model that only relies on the distribution of speech activity as features. This behaviour is similar to the laughter features mentioned above. This result is surprising since existing literature has shown that gaze is a predictor for turn-taking in human dialogue (see section 2.1.3). Firstly, the poor performance of the gaze features could be explained by the insufficient accuracy of the annotations. Eye movements are rapid, and because the gaze annotations were manually generated, it seems likely that the annotators have missed subtle changes in gaze. Another reason might be that the existing lag and rolling features can not fully capture the complexity, and more complex features have to be constructed.

Table 5.33: Performance of the logistic regression model predicting speech activity with added gaze features.

	250 ms		1 s		2 s	
	S02	S18	S02	S18	S02	S18
Baseline	0.7427	0.8020	0.3841	0.4333	0.3372	0.3293
Log. Reg. Model	0.7376	0.8032	0.3565	0.3712	0.2769	0.2888
Log Reg. + Gaze	0.7373	0.8030	0.3677	0.3833	0.2890	0.2943

5.2.2 Facilitator Activity Prediction

This experiment aims to analyse whether the speech activity of one participant can be predicted continuously. In contrast to the previous section, this is a binary classification. The predicted target value is the speech activity of the facilitator (with joint speech). This might be an easier task since the model only has to predict the turns of the facilitator. Therefore, for this experiment, the target value is unbalanced, and performance is examined with precision and recall.

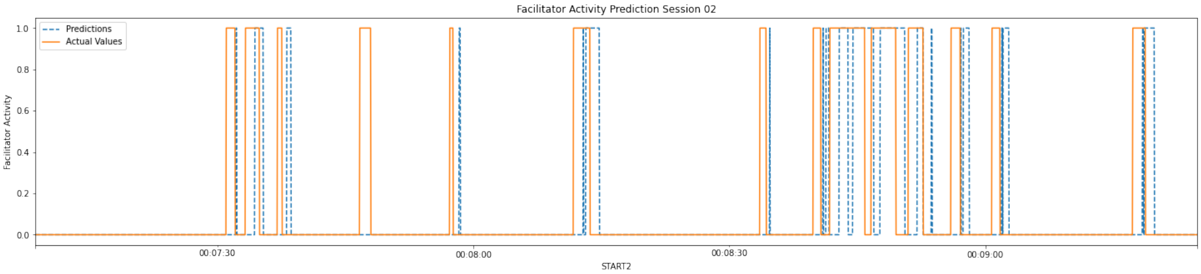
Table 5.34 shows that the logistic regression model behaves similarly to the previous section. The performance of the respective model is very close to the last-known value baseline.

Figure 5.3 shows the plot of the predicted target value against the actual value for 2 s into the future. The figure shows that the same problem exists as in the previous experiment. The predictions lag behind the actual value with a constant factor. The model can not outperform the last-known value baseline.

Table 5.34: <Precision/Recall> of the logistic regression model predicting facilitator activity 250 ms in the future. Only the first four dialogues are reported since the pattern is the same as in the previous experiment.

	S02	S04	S05	S07
Baseline	0.7786/0.7878	0.8384/0.8384	0.8423/0.8493	0.8442/0.8442
Log. Reg. Model	0.7586/0.7505	0.8387/0.8324	0.8443/0.8431	0.8487/0.8510
Log. Reg. + Laughter	0.7551/0.7878	0.8390/0.8347	0.8421/0.8446	0.8473/0.8465

Figure 5.3: Actual facilitator activity against predicted values for 2 s in the future (on the example of session 2)



5.2.3 Summary

The constructed logistic regression models can not outperform a last-known value baseline. In some cases, the logistic regression models perform slightly better than the baseline, but not significantly. Adding gaze and laughter features does not affect the predictions for 250 ms in the future. Models that predict 1 s and 2 s in the future start utilising the additional features. However, the observed improvement in performance is minimal. It was also shown that the predictions lag behind the actual values by a constant factor. The predictive power of the constructed models is therefore limited. To improve performance, trying a different model architecture (such as LSTM) might be helpful. Constructing hand-crafted features that can predict speech activity does not seem feasible.

6 Conclusion

This dissertation presented an analysis of nonverbal signals with the aim of improving turn-taking in spoken dialogue systems. Based on the literature review, the two research objectives of analysing the influence of Big Five personality factors on nonverbal dialogue qualities and the construction of a continuous turn-taking model were deduced. The analysis of the influence of Big Five personality factors on nonverbal dialogue qualities provided evidence that the openness factor influences the total time spoken in dialogue. Furthermore, there is evidence that the openness factor influences the average time between turns of the speaker and the number of times the speaker gets interrupted. There is also evidence that the extroversion factor influences the number of left gaps by a speaker in dialogue. However, these qualities were calculated for a very specific dialogue scenario. It, therefore, remains an open question to which extent these qualities can be useful for improving turn-taking predictions. The results should rather be understood as an encouragement to conduct similar research on bigger datasets with a broader range of dialogue topics. No influence of personality factors was found for the other dialogue qualities. This might be due to the quality of the annotations or the inadequacy of the dataset for this specific task. It is also possible that the way the qualities were calculated did not fully capture the complexity.

For the second research question, a logistic regression model was proposed that predicted turn-taking decisions based on nonverbal features. However, the model was not able to outperform a last-known value baseline. In particular, the addition of gaze and laughter features did not improve the predictive power of the model.

6.1 Future Work

This dissertation leaves opportunities for future research. First of all, the qualities that were shown to be influenced by the personality factors of the players should be validated on other datasets. Conversely, it might be interesting to test on other datasets if these qualities are a good predictor for personality types. Also, it can be interesting to explore the interaction between personality types in the MULTISIMO corpus. For example, it may be investigated whether there are differences in dialogue qualities when two very extroverted players interact with each other or when two introverted players interact with each other. For the presented turn-taking model, it might be interesting to see whether a model based on an LSTM architecture can improve the performance (since related research in the field has successfully used this type of model before). Lastly, it remains a challenge to include the qualities that were found to be influenced by personality type in an actual turn-taking model.

Bibliography

[1] Quim Motger, Xavier Franch, and Jordi Marco. Software-based dialogue systems: Survey, taxonomy and challenges. *ACM Comput. Surv.*, mar 2022. ISSN 0360-0300. doi: 10.1145/3527450. URL https://doi.org/10.1145/3527450.

[2] Gabriel Skantze, Martin Johansson, and Jonas Beskow. Exploring turn-taking cues in multi-party human-robot discussions about objects. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, ICMI '15, page 67–74, New York, NY, USA, 2015. Association for Computing Machinery. ISBN 9781450339124. doi: 10.1145/2818346.2820749. URL https://doi.org/10.1145/2818346.2820749.

[3] Gabriel Skantze. Towards a general, continuous model of turn-taking in spoken dialogue using LSTM recurrent neural networks. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 220–230, Saarbrücken, Germany, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-5527. URL https://aclanthology.org/W17-5527.

[4] Rangina Ahmad, Dominik Siemon, Ulrich Gnewuch, and Susanne Robra-Bissantz. A framework of personality cues for conversational agents. In *Proceedings of the 55th Hawaii International Conference on System Sciences, January 3-7, 2022. Ed.: T. Bui*, pages 4286–4295, 2022. ISBN 978-0-9981331-5-7.

[5] K Laskowski. Predicting, detecting and explaining the occurrence of vocal activity in multi-party conversation (doctoral dissertation). *Language Technologies, Institute School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA*, 2011.

[6] François Mairesse and Marilyn A Walker. A personality-based framework for utterance generation in dialogue applications. In *AAAI Spring Symposium: Emotion, Personality, and Social Behavior*, pages 80–87, 2008.

[7] Mattias Heldner and Jens Edlund. Pauses, gaps and overlaps in conversations. *Journal of Phonetics*, 38(4):555–568, 2010. ISSN 0095-4470. doi: https://doi.org/10.1016/j.wocn.2010.08.002. URL https://www.sciencedirect.com/science/article/pii/S0095447010000628.

[8] Tanya Stivers, N. J. Enfield, Penelope Brown, Christina Englert, Makoto Hayashi, Trine Heinemann, Gertie Hoymann, Federico Rossano, Jan Peter de Ruiter, Kyung-Eun Yoon, and Stephen C. Levinson. Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences*, 106(26):10587–10592, 2009. doi: 10.1073/pnas.0903616106. URL https://www.pnas.org/doi/abs/10.1073/pnas.0903616106.

[9] Carmelina Trimboli and Michael B Walker. Switching pauses in cooperative and competitive conversations. *Journal of Experimental Social Psychology*, 20(4):297–311, 1984. ISSN 0022-1031. doi: https://doi.org/10.1016/0022-1031(84)90027-1. URL https://www.sciencedirect.com/science/article/pii/0022103184900271.

[10] Karl Weilhammer and Susen Rabold. Durational aspects in turn taking. In *Proceedings of the International Conference of Phonetic Sciences*, pages 2145–2148, 2003.

[11] Peter Indefrey and Willem JM Levelt. The spatial and temporal signatures of word production components. *Cognition*, 92(1-2):101–144, 2004.

[12] Zenzi M. Griffin and Kathryn Bock. What the eyes say about speaking. *Psychological Science*, 11(4):274–279, 2000. doi: 10.1111/1467-9280.00255. URL https://doi.org/10.1111/1467-9280.00255. PMID: 11273384.

[13] Harvey Sacks, Emanuel A. Schegloff, and Gail Jefferson. A simplest systematics for the organization of turn-taking for conversation. *Language*, 50(4):696–735, 1974. ISSN 00978507, 15350665. URL http://www.jstor.org/stable/412243.

[14] Starkey Duncan. Some signals and rules for taking speaking turns in conversations. *Journal of personality and social psychology*, 23(2):283–292, 1972. URL https://doi.org/10.1037/h0033031.

[15] Starkey Duncan and George Niederehe. On signalling that it's your turn to speak. *Journal of experimental social psychology*, 10(3):234–247, 1974. URL https://doi.org/10.1016/0022-1031(74)90070-5.

[16] Stephen C. Levinson and Francisco Torreira. Timing in turn-taking and its implications for processing models of language. *Frontiers in Psychology*, 6, 2015. ISSN 1664-1078. doi: 10.3389/fpsyg.2015.00731. URL https://www.frontiersin.org/articles/10.3389/fpsyg.2015.00731.

[17] Jan-Peter De Ruiter, Holger Mitterer, and Nick J Enfield. Projecting the end of a speaker's turn: A cognitive cornerstone of conversation. *Language*, 82(3):515–535, 2006.

[18] Chiara Gambi, Torsten Jachmann, and Maria Staudte. The role of prosody and gaze in turn-end anticipation. In *Proceedings of the 37th Annual Conference of the Cognitive Science Society*, pages 764–769. Cognitive Science Society Austin, TX, 2015.

[19] Tatsuya Kawahara, Takuma Iwatate, and Katsuya Takanashi. Prediction of turn-taking by combining prosodic and eye-gaze information in poster conversations. In *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[20] Bilge Mutlu, Takayuki Kanda, Jodi Forlizzi, Jessica Hodgins, and Hiroshi Ishiguro. Conversational gaze mechanisms for humanlike robots. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 1(2):1–33, 2012.

[21] Carina Riest, Annett B Jorschick, and Jan P de Ruiter. Anticipation in turn-taking: mechanisms and information sources. *Frontiers in psychology*, 6:62–75, 2015. URL `https://doi.org/10.3389/fpsyg.2015.00089`.

[22] Jens Edlund, Joakim Gustafson, Mattias Heldner, and Anna Hjalmarsson. Towards human-like spoken dialogue systems. *Speech Communication*, 50(8–9):630–645, August 2008. ISSN 0167-6393. doi: 10.1016/j.specom.2008.04.002. URL `https://doi.org/10.1016/j.specom.2008.04.002`.

[23] Mateusz Dubiel, Martin Halvey, and Leif Azzopardi. A survey investigating usage of virtual personal assistants. *arXiv preprint arXiv:1807.04606*, 2018. URL `https://doi.org/10.48550/arXiv.1807.04606`.

[24] K. Dautenhahn, S. Woods, C. Kaouri, M.L. Walters, Kheng Lee Koay, and I. Werry. What is a robot companion - friend, assistant or butler? In *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1192–1197, 2005. doi: 10.1109/IROS.2005.1545189.

[25] Heung-Yeung Shum, Xiao-dong He, and Di Li. From eliza to xiaoice: challenges and opportunities with social chatbots. *Frontiers of Information Technology & Electronic Engineering*, 19(1):10–26, 2018. URL `https://doi.org/10.1631/FITEE.1700826`.

[26] Pierre Lison and Raveesh Meena. Spoken dialogue systems: The new frontier in human-computer interaction. *XRDS*, 21(1):46–51, October 2014. ISSN 1528-4972. doi: 10.1145/2659891. URL `https://doi.org/10.1145/2659891`.

[27] Nigel G Ward and David DeVault. Challenges in building highly-interactive dialog systems. *Ai Magazine*, 37(4):7–18, 2017. URL `https://doi.org/10.1609/aimag.v37i4.2687`.

[28] Allan de Barcelos Silva, Marcio Miguel Gomes, Cristiano André da Costa, Rodrigo da Rosa Righi, Jorge Luis Victoria Barbosa, Gustavo Pessin, Geert De Doncker, and

Gustavo Federizzi. Intelligent personal assistants: A systematic literature review. *Expert Systems with Applications*, 147(C), June 2020. ISSN 0957-4174. doi: 10.1016/j.eswa.2020.113193. URL https://doi.org/10.1016/j.eswa.2020.113193.

[29] Vivian Tsai, Timo Baumann, Florian Pecune, and Justine Cassell. Faster responses are better responses: Introducing incrementality into sociable virtual personal assistants. In *9th International Workshop on Spoken Dialogue System Technology*, pages 111–118. Springer, 2019.

[30] Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. Persuasion for good: Towards a personalized persuasive dialogue system for social good. *arXiv preprint arXiv:1906.06725*, 2019. URL https://doi.org/10.48550/arXiv.1906.06725.

[31] Yinhe Zheng, Guanyi Chen, Minlie Huang, Song Liu, and Xuan Zhu. Personalized dialogue generation with diversified traits. *arXiv preprint arXiv:1901.09672*, 2019. URL https://doi.org/10.48550/arXiv.1901.09672.

[32] Luciana Ferrer, Elizabeth Shriberg, and Andreas Stolcke. Is the speaker done yet? faster and more accurate end-of-utterance detection using prosody. In *Seventh international conference on spoken language processing*, September 2002. doi: 10.21437/ICSLP.2002-565.

[33] Seyedeh Zahra Razavi, Benjamin Kane, and Lenhart K Schubert. Investigating linguistic and semantic features for turn-taking prediction in open-domain human-computer conversation. In *INTERSPEECH*, pages 4140–4144, 2019. doi: 10.21437/Interspeech.2019-3152.

[34] Nigel G Ward, Diego Aguirre, Gerardo Cervantes, and Olac Fuentes. Turn-taking predictions across languages and genres using an lstm recurrent neural network. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 831–837. IEEE, 2018. doi: 10.1109/SLT.2018.8639673.

[35] Angelika Maier, Julian Hough, David Schlangen, et al. Towards deep end-of-turn prediction for situated spoken dialogue systems. *Proceedings Interspeech 2017*, pages 1676–1680, August 2017. doi: 10.21437/Interspeech.2017-1593.

[36] Gordon W Allport and Henry S Odbert. Trait-names: A psycho-lexical study. *Psychological monographs*, 47(1):1–171, 1936. doi: https://doi.org/10.1037/h0093360.

[37] John M Digman. Personality structure: Emergence of the five-factor model. *Annual Review of Psychology*, 41(1):417–440, 1990. doi: 10.1146/annurev.ps.41.020190.002221. URL `https://doi.org/10.1146/annurev.ps.41.020190.002221`.

[38] Joseph Jaffe, Stanley Feldstein, and Louis Cassotta. Markovian models of dialogic time patterns. *Nature*, 216(5110):93–94, 1967. doi: https://doi.org/10.1038/216093a0.

[39] Chiara Mazzocconi, Ye Tian, and Jonathan Ginzburg. What's your laughter doing there? a taxonomy of the pragmatic functions of laughter. *IEEE Transactions on Affective Computing*, 2020. doi: 10.1109/TAFFC.2020.2994533.

[40] Maria Koutsombogera and Carl Vogel. Understanding laughter in dialog. *Cognitive Computation*, 14:1405–1420, 2022. doi: https://doi.org/10.1007/s12559-022-10013-7.

[41] Boris Reuderink, Mannes Poel, Khiet Truong, Ronald Poppe, and Maja Pantic. Decision-level fusion for audio-visual laughter detection. In *Machine Learning for Multimodal Interaction*, pages 137–148, Berlin, Heidelberg, 2008. Springer Berlin Heidelberg.

[42] Diana P Szameitat, Kai Alter, André J Szameitat, Chris J Darwin, Dirk Wildgruber, Susanne Dietrich, and Annette Sterr. Differentiation of emotions in laughter at the behavioral level. *Emotion*, 9(3):397–405, 2009. doi: https://doi.org/10.1037/a0015692.

[43] Maria Koutsombogera and Carl Vogel. Modeling collaborative multimodal behavior in group dialogues: The multisimo corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France, may 2018. European Language Resources Association (ELRA). ISBN 979-10-95546-00-9.