Bauhaus-Universität Weimar

Faculty of Media

Degree Programme Human-Computer Interaction

Goethe can touch you: Quantifying the Experience of Poetry Recitals by Physiological Means

Master's Thesis

Oliver Singler Matriculation Number 120753

b. 25.03.1997 in Heidelberg

First Referee: Jun.-Prof. Dr. phil. Jan Ehlers

Second Referee: Dr. phil. Magdalena Wolska

Submission date: 20. November 2021

Contents

List of Figures

List of Tables

Abstract

In this thesis, the idea of quantifying the experience of poetry recitals by looking at the physiological reactions of the audience is presented and tested. It is examined if poetry recitals can elicit physiological reactions, similar to other emotional stimuli and if they can be used to derive a person's opinion on a recital. The work is exploratory, therefore the focus is on the feasibility of the idea and on getting aware of its limits.

First, a novel hypothesis is deduced and supported by a literature analysis. To empirically test the hypothesis and to answer the research questions, a study was planned and conducted. For this, four German poems were selected according to a set of objective criteria. The poems were played to the participants in three recital styles. The styles were recited by a computer generated voice, an amateur speaker, and a professional speaker. The physiological reactions, i.e. Heart Rate and Skin Conductance Response, to the different conditions were recorded from a total of 12 participants. In addition to the physiological measures, the participants had to rate the recitals and answer a questionnaire. A new programming framework was developed for the data recording and the analysis. The collected data was analyzed with several methods. The core of the analysis of the physiological data was a trend analysis. The discovered trends suggest a correlation between the subjective opinion about a recital and the physiological reaction.
As the study was small scale, the results are not generalizable. However, they are promising and do support further research on the topic.

Finally, two sets of Machine Learning models were trained on the collected physiological data. One to predict a participant's rating and one to predict the recital style the participant listened to. The models were trained to further test the hypothesis. The results indicate that both kinds of predictions on physiological measures are possible.

The results of the thesis can be applied in several areas. For example, they can be used to improve poetry recitals of Text-To-Speech systems.

1 Introduction

Poetry is part of the German school curricula. The pupils have to recite and analyze many (old) German poems[1]. However, from own experience, the topic is not much appreciated by the pupils. In contrast, a more recent form of poetry recitals, i.e. poetry slams, is well liked amongst the younger generations (Yanofsky et al., 1999). In order to also make older poems more accessible to younger people, a new form of recitals might be leveraged as well - synthesized recitals.

The 134 Amazon Alexa Skills for the keyword "poetry"[2] show an existing interest for synthesized recitals. One of these skills is called "Sophisticated Ape Robotic Poetry Reading". The author was not able to test the skill, but its description sets low expectations: "Alexa reads dispassionately a randomly selected poem [...]" (amazon.com, 2021). For speech assistants to recite poetry passionately instead of dispassionately, they need to be adapted for this genre of text. There has already been some research on this topic. Most of it is on expressive and emotional Text-To-Speech in general, but there is also work on poetry specifically. However, the results leave room for improvement. (cf. chapter 3)

An improved expressiveness is useful, but emotionality might not be the only feature important for pleasant recitals. Other features could include dramaturgical pauses or the rate of speech. To further aid the synthesizing of poetry recitals, several approaches can be taken. For example, it can be looked at how professional reciters are trained and what they focus on in their recitals. Another approach is looking at what the audience likes about recitals. The easiest way would be to survey subjective perceptions via questionnaires. To make this more objective, physiological measures could be additionally taken into account. When correlating physiological reactions to the recitals, the reactions could indicate preferences for certain recital styles and their characteristics. This idea is presented in the present thesis.

Physiological measures could also be used to quantify the quality of recitals, i.e. the stronger the physiological reaction is, the better the recital. However, trying to automatically and objectively rate art, in this case poetry and its recitals, is a very controversial topic[3,4,5]. Already in 1989, the movie "Dead Poets Society" addressed this. In an iconic scene, the teacher requests the pupils to rip out

[1] https://www.bildung-mv.de/export/sites/bildungsserver/downloads/unterricht/rahmenplaene_allgemeinbildende_schulen/Deutsch/RP_DEU_AHR_7-10.pdf
(09.11.2021, 11:00pm)

[2] https://www.amazon.com/s?k=poetry&i=alexa-skills
(21.10.2021, 3:45pm)

[3] https://dumbscientist.com/archives/can-art-be-evaluated-objectively
(24.10.2021, 07:30pm)

[4] https://the-artifice.com/thats-just-like-your-opinion-man-an-argument-that-art-is-objective/
(24.10.2021, 08:00pm)

[5] https://christopherpjones.medium.com/subjectivity-and-objectivity-in-art-cc41d55c76a5
(24.10.2021, 08:00pm)

the preface “Understanding Poetry” of their textbooks. The chapter concerns “[...] determining [a] poems greatness [...]” (Weir, 1989, 00:20:46 h). The teacher argues that poetry cannot be rated objectively. By extension, it should therefore also not be possible to objectively rate its recitals.
However, the method proposed in this thesis does not claim to provide objective ratings. Rather, it uses the subjective perception of the audience and tries to quantify individual experiences. A generalization is hard to achieve due to high inter-individual differences in physiological reactions. The goal is to examine the idea of *quantifying the experience of poetry recitals by physiological means* (cf. chapter 4). The thesis project constitutes basic, preliminary research. The literature review in chapter 3 indicates an existing academic interest.

The results of the project can be leveraged to help understand the interplay of poetry and its recitals better. They can help identify which features are important for good recitals. These features can be taken into account when adapting Text-To-Speech systems for poetry. Such a customization would be considered successful if the physiological reactions to the modified system are similar to the reaction to human recitals. (cf. chapter 9) The Text-To-Speech synthesis would be improved for a very specific area of application. However, the modification should not only be helpful for reciting poems but also for other emotional genres of text. Such an improved Text-To-Speech system could be used in voice assistants.
Additionally, the results could be used to build Machine Learning models to predict the rating of a recital based on the physiological reaction (cf. chapter 6).

The author decided on the research topic for numerous reasons. First, the subject fits the author’s academic background and his personal interests. He has a bachelor’s degree in computational linguistics, with a minor in psychology. As he wanted to connect both fields more deeply, he picked Human-Computer Interaction as his master’s. A main topic of interest concerns the experience of humans. This includes the more subjective User Experience and the externally measurable physiological experience, as well as their connections. The thesis touches upon both kinds. It lays the foundation to improve the User Experience of voice assistants by drawing from the analysis of physiological reactions.
Furthermore, voice assistants are a very current topic, they get more usable by the day. However, the author uses Apple Siri himself on a daily basis and regularly notices room for improvement.

Today, we live in a short lived world. The global attention span got shorter over the last years (Lorenz-Spreen et al., 2019). The information overflow and shorter attention spans require to hook and catch the audience immediately, as they otherwise might just continue to the next best stimulus. This requires to stimulate their senses in a positive way and therefore calls for adjustments to marketing strategies (ranieriandco.com, 2021). To appeal to the potential audience, (video) advertising frequently utilizes emotions to keep viewers engaged (Teixeira et al., 2012).
Standing out with emotional, good quality recitals presented in a novel way, i.e. synthesized with Text-To-Speech systems, could help **increase the appreciation of poetry as an expressive form of art**.

Since not all people can consume poetry in its conventional form, be it due to dyslexia or limited eyesight, synthesized recitals would also significantly expand the prospective audience. In addition, not every poem has been professionally recorded, and an automatic method for producing good quality recitals would **make lesser-known poems in particular more accessible**.

2 Hypothesis

Before presenting the hypothesis and the research questions of the present study, five key topics need to be defined. The definitions are important for a common understanding and increase the study's reproducibility. Furthermore, the motivation behind the hypothesis is explained.

2.1 Definitions

Poetry

Everyone has a gut feeling what poetry is, but it is hard to quantify (scientifically) (Mill, 1860). Stavenhagen (2019) states that "[...] poetry is a 'condensed' linguistic product that deals with its subject in a small space and with relative brevity." (Stavenhagen, 2019, p. 28). Besides the condensed form of content, the metre is identified as a defining property of poetry. The presence of rhymes is not a necessity but common. Usually, rhetorical devices are used in poetic texts.

Recital

A recital, and more specifically a poetry recital, refers to reading a text, i.e. a poem, aloud for an audience (Cambridge Advanced Learner's Dictionary & Thesaurus, 2021; Merriam-Webster, 2021). In the context of this study, it is further defined that the audience does not need to be present, it can also be imagined. The reciter tries to engage the audience and tries to achieve that the audience experiences the content of the poem. Each reciter adds their own artistic layer to a poem. It does not matter if the recital is performed by an amateur or by a professional.

Text-To-Speech Synthesis

Text-To-Speech (TTS) synthesizers artificially generate human-sounding speech. Written text is entered into a program which converts it into speech. Today, different viable methods can be used for the synthesis. Some are based on short recordings of voice actors, some try to emulate the human vocal tract. (Balyan et al., 2013; Mache et al., 2015; Schröder, 2001)

Physiological Reaction

A physiological reaction is the reaction of the body to a stimulus, accompanied by increasing arousal (Wang et al., 2018). There are many different kinds of physiological reactions. In this thesis, two kinds of reactions are considered: The Heart Rate (HR) and the Skin Conductance Response (SCR).

The Heart Rate is the frequency of the heart beat, measured in beats per minute (bpm). It is controlled by the parasympathetic nervous system (Wang et al., 2018).

The Skin Conductance Response is one component of the Electrodermal Activity (EDA). The EDA is quantified by the electric resistance of the skin, which is influenced by the sweat glands. The activity of the sweat glands increases, i.e. they produce more sweat, with increasing arousal. When the activity increases, the electrical resistance decreases and the conductance increases. The conductance is measured in microsiemens (µS). The Skin Conductance Response is the direct response to a stimulus, it is the phasic component. The other component is the Skin Conductance Level (SCL), the background activity or tonic component. (Christopoulos et al., 2019)

Emotions

Many psychological papers on the topic of emotions are not defining the term correctly before using it (Cabanac, 2002). They presume the existence of a common understanding.

Cabanac (2002) proposed their own definition: emotions are cognitive experiences with a high intensity and a very distinct sentiment. The sentiment is an important requirement, it can either be strongly positive or strongly negative. Usually, basic emotions make up finer grained emotions.

In the following, three proposed sets of basic emotions are described. Watson and Morgan (1917) argue that rage, fear, and love can be recognized in infant behavior and are therefore the basic human emotions. Paul Ekman proposed seven basic emotions but later argued that every emotion is a basic emotion (Ekman, 1999). Plutchik (1980) presented four negative emotions: fear, sadness, disgust, and anger and four positive emotions: joy, trust, surprise, and anticipation as their basic emotions. All three of these sets conform to the definition by Cabanac (2002). (cf. Singler, 2019)

In the questionnaire of the present study, Plutchik's basic emotions are used. The author worked with this set before and it is frequently used in emotion related publications (e.g. Zhou et al., 2016; Araque et al., 2018; Kim and Klinger, 2018; Mohammad and Turney, 2013).

2.2 Motivation

Many studies have shown that the human body exhibits physiological reactions to experienced emotions (cf. Wang et al., 2018; Christopoulos et al., 2019; Leite et al., 2019). Besides others, the Heart Rate and the Skin Conductance Response are properties that increase. Both can be easily measured externally in a non-invasive way. The reaction of Heart Rate and Skin Conductance Response to emotions is unspecific, i.e. it is not possible to distinguish between a negative and a positive stimulus. However, stronger emotions result in stronger reactions. (Stemmler, 2004; Rickard, 2004)

Many poems touch emotional topics like love or death. The poets try to elicit emotions in the readers, making poetry an emotional genre of text (Mill, 1860; Wassiliwizky et al., 2017). In general, texts that look like poems are perceived more emotional (Peskin, 2008). Reciters try to capture and share these emotions, which makes recitals an emotional form of art.

If poetry recitals can incite emotions in the audience, poetry recitals can incite measurable physiological reactions. The stronger the experienced emotion is, the stronger the reaction is. Measuring evoked emotions is easier with more emotional poems. Under the assumption that good recitals are able to elicit more emotions, no matter if positive or negative ones, good recitals should produce stronger physiological reactions. Therefore, by measuring the physiological reaction, it can be determined how well a recital is perceived and the reaction can be used as an indicator of quality.

The study, presented in chapter 4, can be seen as basic research, to be later built upon. The results will show if the quality of recitals can be adequately measured by physiological reactions and if future work can build upon the idea. So far, there has been no work on this specific topic (cf. chapter 3). With the results, poetry and its recitals could be understood better. Features of recitals could be identified that are liked or disliked. They could also help to improve the understanding of how emotions could be incorporated into recitals.

2.3 Hypothesis

Based on the preceding motivation, the following hypothesis is defined for the present thesis:

> **The quality of poetry recitals can be quantified, measuring physiological reactions.**

The hypothesis is supported by a literature review (cf. chapter 3) and tested in a lab study (cf. chapter 4). The independent variables in the study are poetry recitals of reciters with different levels of skill (IV1) and different poems to be recited (IV2). The dependent variables are the Heart Rate, the Skin Conductance Response and the rating of the participants. (cf. chapter 4) It is expected that there are significant differences in the physiological experience of poetry recitals. The study and its results are presented in the following chapters.

Additional to the hypothesis, the following research questions will be tried to be answered:

- How are different recital styles perceived?
- Are Heart Rate and Skin Conductance Response adequate measures for the experience of recitals?

3 Related Work

This thesis touches upon a topic with limited previous research. At the time of writing, no scientific work on using physiological reactions to poetry recitals to quantify the experience has been published yet. The thesis can build upon only few previous findings.

However, there has been a lot of research on physiological reactions in general, as well as in related fields. Google Scholar lists over 3.5 million results for the keyword "physiological reaction"[1]. As shown in chapter 2, the underlying mechanisms of the presented topic can be explained with research in these related fields. Work in four key areas of interest is presented below.

3.1 Perception of Emotions

The human perception of emotions is an ongoing research topic. In related studies, the brain activity is often monitored with functional magnetic resonance imaging (fMRI).

In such a fMRI-study, Gandour et al. (2003) researched how intonation and emotions are perceived in Chinese by natives and non-natives. The participants had to judge both in sentences with neutral semantics. The authors discovered that intonation and emotion are, in some parts, processed independently.
A fMRI study can provide deep insights into the emotional processes of the brain. Due to the associated costs and logistics, such an approach would only be feasible in future research.

Koelsch et al. (2006) likewise conducted a perceptional fMRI-study. Their goal was to determine which areas of the brain are activated when listening to pleasant and unpleasant music, resp. auditory stimuli. The music was thought to induce emotions. The results show that the pleasant and unpleasant music was indeed perceived as pleasant and unpleasant and evoked related emotions.
In future work, the reaction to music and to poetry could be compared.

Other studies leverage different physiological measures, like the Heart Rate and the Skin Conductance Response. Wang et al. (2018) investigated the relation between Heart Rate and Skin Conductance Response, and Pupil Dilation. The participants had to recognize emotions in presented faces while the physiological measures were recorded. The authors found that Pupil Dilation is a good indicator of emotional arousal but is best used together with other indicators.

[1] `https://scholar.google.com/scholar?q=physiological%reaction` (06.11.2021, 7:15pm)

In the present study, the use of the Pupil Dilation as an indicator would not be appropriate (cf. chapter 4). The authors dealt with short, single stimuli while this study works with longer, continuous stimuli. Still, the authors confirmed that Heart Rate and Skin Conductance Response are indicators of emotional arousal.

Leite et al. (2019) let participants explore architectural safe and unsafe virtual environments while recording their physiological arousal. As dependent variables, the Skin Conductance Level and the Heart Rate were measured amongst others. Additionally, the data was correlated with self-reports. The authors were able to confirm that unsafe environments trigger negative emotions and stated that the Skin Conductance Level was a better indicator than the Heart Rate.
In comparison to the present study, Leite et al. (2019) took a different approach to Electrodermal Activity. First, they used the Skin Conductance Level instead of the Skin Conductance Response. Furthermore, they calculated changes for a sliding window instead of doing a trend analysis. In future research, both approaches could be compared.

3.2 Poetry and Emotions

Very little work has been published on the relation of poetry and emotions.

Wassiliwizky et al. (2017) have done a similar study to Koelsch et al. (2006) but with poetry. The authors examined whether poems can trigger emotional responses. In two phases, they played poetry recitals and measured the reactions of the participants. They decided to use recorded recitals, as they wanted to rule out the reading abilities of the participants as a confounding factor. In the first phase, they looked at physiological data (facial expressions, Heart Rate, Skin Conductance Response, and Piloerection). During the second phase, they used fMRI to examine the brain activity. In both phases, the participants had to self-report when they experienced chills. The authors' main interests were the occurrence of chills and Piloerection, i.e. goosebumps, and their correlation to the other measures. The results got compared to the results from similar research on music. The authors found many similarities and some key differences. Finally, the authors looked for text features that correlate with the occurrence of chills. They identified a poem's formal structure as one feature: the occurrence of chills increase at the end of lines, stanzas, and poems. The other identified feature is social address. Furthermore, the authors found that negative emotions were also enjoyed by the participants.
Wassiliwizky et al. (2017) only used professional recitals and did not compare the reactions to different recital styles. Their work showed that poetry recitals are capable of eliciting emotions in the audience, as 77% of the participants had an emotional reaction, i.e. felt chills, while listening to the recitals. This finding supports the reasoning behind the in chapter 2 presented hypothesis. Furthermore, the authors reasoned that a professional can replicate the formal structure of a poem better than non-professionals. If this is correct, differences between the styles should show in the present study. Another key difference is that Wassiliwizky et al. (2017) looked for specific events, i.e. the occurrence of chills, while the present study looks at the overall reaction to an entire recital.

3.3 Text-To-Speech and Emotions

Several authors already published their work on expressive and emotional Text-To-Speech. Expressive Text-To-Speech is not actively worked on in the present study, as it is only an area of application. Still, the research provides insights into features of speech that are considered making it more emotional.

Most of the research has been on the general topic, but there has also been work for specific genres. For example, Declerck (2017) developed annotations to support the synthesizing of folktales in form of a multilogue. Eisenreich et al. (2014) presented a fairy tale Text-To-Speech system which uses an ontology.

Rebordão et al. (2009) automatically adapted the prosody of generated speech with the help of the text. Their goal was to convey the subtleties of human speech better and making synthesized speech sound more natural. For example, they adapted the pitch and the rate of speech. In the common Text-To-Speech systems that the authors tested beforehand, the participants were not able to identify the intended emotions correctly. The authors used an emotion classification system on their input sentences and modeled the prosody accordingly. However, it was not explained how they came up with the modeling rules. The improved system performed better than the baseline but significantly worse than human speech.
Nass et al. (2001) analyzed how differently recorded human speech and generated speech elicit emotions in listeners. The authors varied the emotions conveyed by the text's semantics and by the speech, and afterwards queried the participants on the heard emotions. The synthesizer's parameters stem from the authors' previous research. The findings show that Text-To-Speech can convey emotions similar to human speech.
Based on questionnaires, Cohn et al. (2020) surveyed differences in the ability to identify changes in the emotionality of human and synthesized speech. The participants were able to identify the gradual changes in both conditions. However, the changes were perceived stronger for the human voice.
In a review, Schröder (2001) discussed several different approaches to making Text-To-Speech more emotional. They found that different features apply for different methods of synthesis. Of the different possible synthesis methods, the most natural results were achieved with unit selection. The suggested features include the fundamental frequency, the rate of speech, and the loudness. Some of the authors of the reviewed papers performed a corpus analysis to determine these features.

Schröder et al. (2007) gathered requirements for a general emotion markup language. Goal for this language was to support the annotation of emotions, as well as their recognition and generation. The project resulted in the development of EmotionML, which was later used by Charfuelan and Steiner (2013) for expressive speech synthesis. Charfuelan and Steiner (2013) paired the text synthesizer MARY TTS and EmotionML to generate expressive speech. They used only one voice for several narration styles instead of different voices, as previously done. For the training of the voice, they used sentences from audio books. The input sentences were automatically labeled with emotions based on their acoustics. The participants were each played the synthesized sentences with different emotions and had to decide which sounded the most similar to the original,

i.e. expressed the same emotion. Extreme emotions were easier to match. The authors developed this for voices based on unit selection and based on a Hidden Markov Model (HMM) and presented positive results.
Eyben et al. (2012) worked on a similar task using HMM-based voices with unsupervised clustering. Their method performed better than the baseline system.

All these papers show that Text-To-Speech systems can produce expressive speech. They indicate an existing interest in research on the topic. However, they also show that more research is needed for further improvements. The present study's results will show if the features from the related work correspond with the participants' assessments.

3.4 Text-To-Speech and Poetry

Synthesized poetry recitals are still a niche topic. However, the research indicates that one can adapt Text-To-Speech systems for poems and achieve promising results.

The project "The Muse of Poetry" by Arellano et al. (2014) synthesizes poetry recitals and animates a virtual, synchronized face. The authors call this "[...] a unique poetic experience." (Arellano et al., 2014, p. 383). As preprocessing, they analyzed the affectivity of the poems, but did not include this information in their speech synthesis. The emotions were solely induced via the face animations. Testers deemed the voice to be too robotic.
The authors had an unique approach, as they enhanced the auditory stimuli with visual stimuli. Unfortunately, the authors did not use the voice to express emotions. An expressive voice might have further improved the results and could have made the voice sound less robotic. For further research, it would be interesting to look at the different (physiological) experiences of the separate sensory channels and their combination.

Delmonte (2013) developed SPARSAR, an analysis tool for the style of English poems. Besides this paper, there are further publications on SPARSAR (e.g. Delmonte and Prati, 2014). The tool builds on the "American Poetry Style Analyzer" (Kaplan, 2007). It analyses many text features and computes different kinds of indices for them. The indices can be used for the comparison of poems or as pre-processing steps for Text-To-Speech synthesis (Delmonte, 2019). However, emotions or physiological reactions are not considered in SPARSAR.

4 Methods

In the study for this thesis, the physiological reactions to different recital styles of poetry were recorded and compared. For this, the reactions to four poems, in three styles each, were examined. As the three styles, speakers with different levels of skill were used: a Text-To-Speech voice, an amateur speaker, and a professional speaker.
Afterwards, it was analyzed if there is any measurable difference in the physiological reactions while listening to the different recitals. Due to the small population size, the reactions were only compared on an intra-poem level. The expectation was that the participants react most, i.e. with the most arousal, to the professional style. This was supported by the presumption that the professional style should be better at eliciting emotional reactions than an untrained or virtual one. The amateur style should evoke a stronger reaction than the Text-To-Speech style. In the following, style and speaker are used synonymous.

Additional psychological measures were collected with a questionnaire. By linking the physiological reaction to psychological measures, it was looked for markers that indicate the liking or disliking of a recital.

The design of the experiment follows the current scientific standards. The main study was prefaced with a pilot study, which influenced the final design. Changes introduced after the pilot are mentioned where applicable. Two participants took part in the pilot. Their demographics were similar to the participants in the main study (cf. chapter 5): One male and one female participant, both 28 years old. One was a student, the other a research associate. In the pilot study, everything but the Skin Conductance measurement was set up properly. Additionally, an interview on the design was conducted at the end.

4.1 Design

The study was set-up as a simple within-subject design. All participants had to listen to all stimuli in different orders. However, no comparison between poems was done in the following analysis. The design and the study's small size did not allow for more complex comparisons. Therefore, the poems were only analyzed individually on an intra-poem level. In the one-factor within-subject design, four different poems in three recital styles were presented. The styles included an artificial voice, an amateur speaker, and a professional speaker. For consistency, all three were male. For the artificial style, a Text-To-Speech synthesizer was used. The amateur style recordings were provided by an acquaintance, the professional style recordings were acquired online. The three styles were chosen because they represent varying levels of skill and different experiences by the audience were to be expected.

The four recited poems had an objective difference in one dimension, the valence (positive vs. negative). Even though no inter-poem comparisons were performed later, this was included to reduce a potential impact of the general emotion of the texts. To ensure that the poems have a similar complexity, the text comprehension index "Hohenheimer Verständlichkeitsindex" (HIX)[1] had been calculated for each poem but was ultimately disregarded (cf. section 4.2). To further reduce comprehension as a confounding variable, only native speakers were considered as participants. Since there was a large pool of native German speakers available, all poems were in German. To account for potential habituation on an intra-poem level, poems of roughly the same length were selected.

The two biological measures were Heart Rate (HR) and Skin Conductance Response (SCR), which were later linked to psychological measures, gathered via a questionnaire. Heart Rate and Skin Conductance Response are well proven measures for emotions and bodily arousal (Wang et al., 2018). Additionally, the unprocessed sensor data of the Heart Rate measuring device ("Raw Mouse data") was recorded and evaluated. At first, it was considered to include pupil size changes as well. However, the pupil size is only a good measure for changes in the millisecond range (Burley and van Goozen, 2020). For the present study, this resolution would have been too fine-grained.

To counteract serial order effects, the order of the recital styles was balanced, resulting in six unique orders. The order of the poems within one style was randomized for each participant. Balancing them as well would have required a larger cohort. In between the emotional stimuli, neutral stimuli were played, and breaks to relax and counteract the halo effect were added. As the neutral stimuli, 30 second long, monotone readings of the university's data protection policy were used. To prevent the participants from getting used to the neutral stimuli, progressing snippets were used. To account for habituation effects, local baselines were calculated. They were calculated on the neutral stimulus before each emotional stimulus. Between recital styles, there were 30 seconds long silent breaks to relax and refocus.

Directly after each prompt, the participants had to rate the recital on a scale between 1 ("I did not like the recital") and 10 ("I liked the recital very much"). This was done to record an immediate, impromptu reaction. Important was that the participants rated the recital and not the poem's content. To not distract too much and to prevent the anticipation of questions, a more comprehensive questionnaire with further questions followed after all prompts had been heard (cf. section 4.1.1). After the pilot study, the rating scale was expanded from between 1 and 5 to between 1 and 10, to allow for more fine grained ratings.

The recorded Skin Conductance changes, Raw Mouse data, and Heart Rate changes were later normalized with the corresponding local baselines. On this data, the parameters amplitude, rise time, and half recovery time were computed and compared between the different recital styles within a poem. The goal was to correlate the parameters to experienced pleasure and displeasure. Furthermore, the data was linked with the information from the questionnaires.

Before the data collection, a short demographics questionnaire had to be filled out by the participants. The questionnaire asked for age, gender, and primary occupation. Furthermore, participants had to come up with a pseudonym which was referenced in the recorded data and the final questionnaire. (cf. Appendix)

[1] https://klartext.uni-hohenheim.de/hix

4.1.1 Questionnaire

The final questionnaire consisted of quantitative and qualitative questions. The questions asked about the participants' experience and how the different poems and styles were perceived and liked. Some questions related to each other and were used as an indirect measure of validity. The original, German version of the questionnaire can be found in the Appendix. As a reminder, the participants received the texts of the poems while filling out the questionnaire. The answers were later correlated to the recorded physiological data. The goal was to find a connection between the psychological and the physiological experience.

The questionnaire was divided into several different sets of similar questions.

Ratings

The first set of questions begun with an overview of the recitals' ratings from the participant. The sheet got filled out by the examiner during the data recording. The participants had the opportunity to reevaluate their ratings after hearing all stimuli, leading to more comparable results while still incorporating the first impression.

Descriptions

The second set of questions asked the participants to describe the speakers from the three styles as they would describe them to someone who had not listened to them and to highlight the differences. Furthermore, the participants had to state what they liked and what they disliked about the speakers, aiming to find features and characteristics that the participants appreciate.

Preferences

The next set started with asking the participants from which of the speakers they would most likely listen to another recital to. This question validated the participants' answers to the open questions, as well as their impromptu ratings. Additionally, they had to order the speakers by their preferences and match the best fitting speaker to each poem. The target was to find a pattern of which features fit best with which poem.

Emotions

The following set begun with the task to associate each poem with one of Plutchik's eight basic emotions[2] (Plutchik, 1980), to examine if the main emotion has an influence on the liking of a recital. Furthermore, the participants had to choose which speaker conveyed the emotions the best per poem and indicate how the emotions became apparent. The question was added to help find preferred features.

Content

In the subsequent set, the participants were asked to sort the poems by perceived emotionality, as well as by personal preferences. These questions were about the poems' contents, not about the recitals, trying to determine its influence on a recital's rating.

[2] anger, anticipation, disgust, fear, joy, sadness, surprise, trust

Bodily Changes

The next set of questions asked about bodily changes during the listening period and about changes in attention. Both were aimed to find out whether the participants noticed any changes themselves and if the data needed to be cleaned from external influences. The participants also had to state if they used a technique for concentrating, e.g. visualizing the content. This might have influenced how concentrated the participants were, which in turn could influence the quality of the data. It was followed by the question if they did know any poem before the study. This question targeted detecting potential reduced or heightened physiological reactions, if the participant already knew the poems beforehand.

Comments

To incorporate the participants' views on the topic at hand, they were asked if they think that recitals can be objectively judged and if so, which features could be used for this. Finally, the last question provided the opportunity for the participants to add an open ended comment. For example, for commenting on the study, its design or for adding their own insights and ideas on the topic.

Besides adapting the rating scale, the possibility to adjust the rating was added to the questionnaire after the pilot study. Overall, some wording of the questions was adapted to be better comprehensible. For example, "Stimme" (voice) was changed to "Sprecher" (speaker) after the pilot study to be more precise.

4.2 Stimuli

For the study, four poems had to be selected. Besides gathering these emotional stimuli, the neutral stimuli had to be generated. The following criteria were established for the poem selection.

As the first criterion for selection, the poems had to be available on the website `deutschelyrik.de`. On this website, Fritz Stavenhagen, a professional German voice actor, publishes his poetry recitals. The website was chosen as the source for the professional recitals due to its big selection of high quality recitals. By the owner's account, over 1,600 recitals are available with their respective texts, all recorded by himself. More information on the speaker can be found later on.

Initially it was planned to use Goethe's ballad "Erlkönig" in a later phase of this thesis. The usage was decided upon before the planning of the study phase. To have stimuli of comparable length, the poems needed to be a maximum of about 25% shorter or longer than "Erlkönig". As the professional recital of "Erlkönig" took 2:08 minutes, the other recitals, when recited by the professional, should take between 1.5 minutes and 2.5 minutes.

As mentioned previously, the goal was to include two mainly positive and two mainly negative valenced poems, i.e. two happy and two sad poems. To limit the number of recitals for which to determine the valence, the poems had to either belong to the category "ballad" (presumed negative valence) or the category "humor" (presumed positive valence). The two categories were assigned by Fritz Stavenhagen on deutschelyrik.de.

A total of five ballads, excluding "Erlkönig", and 14 humorous poems met these requirements. For the final selection of the stimuli, the emotional valence had to be identified. To determine a poem's valence, a script utilized the NRC Word-Emotion Association Lexicon (EmoLex) (Mohammad and Turney, 2013) and calculated the novel measure "EmoScore".

EmoLex is an extensive dictionary, associating words with Plutchik's eight basic emotions. The dictionary marks associations in a binary fashion: either a word is associated with one or more emotions, or it is not. Therefore, the dictionary does not indicate the strength of such an association. Associations with positive and negative valence are indicated in a similar way. To gather the data, the authors leveraged crowdsourcing. The dictionary was originally assembled for English, but with the help of Google Translate it was extended to 105 languages, disregarding if associations would change. (Mohammad and Turney, 2013; Mohammad, 2021). There are 14,182 unique English words in the dictionary. However, not all of these words have an unique translation to German. The number of unique German words is 11,200.

For each potential stimulus, a Python script calculated the probability of a word belonging to an emotion. As stated above, some German words occurred multiple times with different associations. These occurrences were merged using a logical OR-function for the association indicators. Words without a translation ("NO TRANSLATION") were removed. All remaining words were lowercased.

Before calculating the probabilities, the poems had to be pre-processed. Using NLTK (Bird et al., 2009) and its "RegexpTokenizer", all punctuation was removed and the texts were tokenized. Like the words in the dictionary, the poems were lowercased. Afterwards the words were lemmatized, using the Python package HanTa (Wartena, 2019).

After the pre-processing steps, the number of associated words per emotion were counted per poem and divided by the number of words in the respective poem. The same was done for the valence. Additionally, the probabilities of a word being associated with any positive emotion[3] or any negative emotion [4] were calculated respectively. The probability for a negative emotion was subtracted from the probability for a positive emotion. The resulting novel score is called "EmoScore". A higher EmoScore is expected for more positive poems, a lower EmoScore is expected for more negative poems. The EmoScore was implemented as an addition to the valence. It was expected that the EmoScore and the probability for positive valence subtracted by the probability for negative valence ("Valence" in table 4.1) should be similar. In most cases they had a similar tendency. However, these two metrics do not always give the expected results. For example, Goethe's "Erlkönig" received an EmoScore of 0.07, indicating a very positive poem. Actually, 0.07 is one of the highest scores out of the examined poems. Contrary to the these results, the poem tells a very sad and negative story and a low score would be expected. This limitation arises due to the approach of counting associated words and not taking the content of the poems into account.

[3] anticipation, joy, surprise, and trust
[4] anger, disgust, fear, and sadness

Since the EmoScore does not always perform as expected, an additional subjective assessment of the poems was performed. The EmoScore only served as a first indicator which poems to analyze more in-depth. The Python script generated a csv-file with the calculated probabilities. The results for the four chosen poems can be found in table 4.1.

After the subjective analysis, it was decided on the following four poems: As the negative valenced poems, “Die Grenadiere” by Heinrich Heine and “Der Fischer” by Johann Wolfang von Goethe were chosen. Similarly to “Erlkönig”, “Der Fischer” has a high EmoScore whilst being a negatively valenced poem. As the positive valenced poems, “Entschuldigungsbrief” by Joachim Ringelnatz and “Wahre Liebe” by Kurt Tucholsky were chosen.

To ensure that the selected poems are of similar comprehensibility, the “Hohenheimer Verständlichkeitsindex” (HIX) was calculated for the four poems with the Tool TextLab (Kercher, 2013). However, all poems achieved unexpectedly good scores, indicating that the index would need extensive tuning for adequate results. As this would be out of scope for this thesis, the index was not considered in the selection process. Instead, the comprehensibility was subjectively judged and deemed similar.

Additionally to the emotionality, some more information on the poems has been gathered. Rhyme scheme, metre, and cadence can be found in table 4.1. The poems were written between the late 18th century and the early 20th century. Both ballads are older than the humorous poems.

With the poems decided upon, the recitals had to be collected, recorded or generated. The final recitals of all three styles had the same structure: The speaker read the poem’s title and started reciting the text. The author was always omitted. This structure followed the recordings by the professional speaker. To remove gender as a potential confounding variable, all three voice actors had the same gender.

For the professional style, Fritz Stavenhagen, owner and voice actor of the website `deutschelyrik.de`, kindly provided his recordings. Fritz Stavenhagen is a professional voice actor with many years of experience. His work includes a wide field, from audio plays to commercials. Besides doing voice acting, he also worked as

Table 4.1: Valence results of the chosen final poems with additional information. Valence is the probability that a word in the poem has negative valence subtracted from the probability that it has positive valence. EmoScore is the probability that a word is associated with the emotion anticipation, joy, surprise, or trust subtracted from the probability that it is associated with the emotion anger, disgust, fear, or sadness. In both cases, positive values are associated with positive valence and negative values with negative valence. The additional information is the rhyme scheme, the metre, and the cadence of the poem.

Poem	Valence	EmoScore	Rhyme Scheme	Metre	Cadence
Entschuldigungsbrief	0.06	0.05	Alternate/Coupled	Iambic	strong
Der Fischer	0.05	0.06	Alternate	Iambic	strong
Die Grenadiere	-0.06	0.00	Alternate	Iambic	weak
Wahre Liebe	0.05	0.08	Alternate/Coupled	Iambic	strong

director, singer, and actor. His academic background is German philology and theater studies. One of his better known works is the role as the narrator in the German series Ritter Rost. Since 2001, he has been working on his project deutschelyrik.de. (fritzstavenhagen.de, 2021) Until now, he has published recitals of 1,600 German poems (deutschelyrik.de, 2021). The recitals can be listened to for free. For a fee, the recitals can be downloaded as mp3-files. The four recitals chosen as stimuli were recorded between 2001 and 2017. Table 4.2 shows the lengths of the professional recitals, ranging from about 1.5 minutes to 2.3 minutes.

The amateur style was kindly recorded by Julius Uhlmann, a 28 years old acquaintance with experience in radio production. According to himself, between 2013 and 2018, he worked at coloRadio Dresden as a part-time journalist. Between 2015 and 2018, he moderated and produced the "Einhornfutterradioshow". In 2016, he took part in the "EUROPHONICA" exchange program in Strasbourg, where he reported from the European Parliament. He was chosen for his experience, as he had access to the proper equipment and high-quality recordings could be expected. He was instructed to recite the poems as he sees fit. The only guideline was to not listen to any recitals by others before recording his own. The lengths of his recitals can be found in table 4.2. They range from 1.0 minute to 1.65 minutes.

The Text-To-Speech style was synthesized using MaryTTS (Charfuelan, 2012). MaryTTS is an open-source speech synthesizer, commonly used in research projects. Many voices in different languages are available. The most neutral sounding available voice for German was "dfki-pavoque-neutral-hsmm"[5]. This male voice is generated, using a "Hidden Semi-Markov Model" (HSMM). A HSMM "[...] can be considered as an HMM with explicit state duration probability distributions [...]" (Zen et al., 2004, p. 1394). The voice's standard synthesizing settings were not changed. For adequate results, the poems had to be pre-processed. The texts were converted into the XML-format SABLE (Sproat et al., 1998), a proposed markup language for Text-To-Speech from 1998. This format was chosen for its ability to insert short breaks into the audio with the <break>-tag. Using this tag, breaks were inserted at the end of each verse to add a more comprehensive structure to the recitals. Furthermore, all apostrophes were removed, as they introduced unwanted speech artifacts. In case of pronunciation errors, they were tried to be fixed by varying the spelling. This was done to eliminate such errors as a distraction, i.e. confounding factor. Regarding the rhythm, emphasis or stress, no changes were introduced. The duration of the recitals ranged from about 1.0 minute to about 1.5 minutes (cf. table 4.2).

Table 4.2: Lengths of the recitals in the Text-To-Speech style, the amateur style, and the professional style measured in seconds.

Poem	Style		
	Text-To-Speech	Amateur	Professional
Entschuldigungsbrief	62 s	60 s	116 s
Der Fischer	70 s	73 s	140 s
Die Grenadiere	94 s	99 s	138 s
Wahre Liebe	61 s	59 s	92 s

[5] https://github.com/marytts/voice-dfki-pavoque-neutral-hsmm

For the neutral stimuli, the same Text-To-Speech settings were used as for the Text-To-Speech recitals. However, the pre-processing was not as extensive. No manual breaks were added. Punctuation was added at line breaks, to make the flow more natural. Since apostrophes result in incorrect pronunciation, they were removed. No further adjustments were made. The text, i.e. the privacy policy, was sourced from the university's website (uni-weimar.de, 2021). The whole text was synthesized and later cut into 30 seconds long chunks. The entire file had a length of 38 minutes 35 seconds, from which only the first 6.5 minutes were used.

The lengths of the amateur and Text-To-Speech recitals were similar. The professional recitals were up to twice as long. The main reason for this seems to be the slower pace of the professional and him adding more dramaturgical pauses. As intended, the Text-To-Speech style was very monotone and neutral. It also had the highest pitch with an average frequency of 250 Hz. The amateur had an average frequency of 201 Hz and the subjectively youngest sounding voice. With 172 Hz, the professional had the lowest average frequency and the subjectively oldest sounding voice. The professional had the fewest pronunciation mistakes, while the Text-To-Speech often sounds weird or off. In general, the three speakers are very distinct in their abilities and were suitable as stimuli for the task at hand.

4.3 Sample

Having six unique orders for the styles required the sample size to be a multiple of six for balancing. The target sample size was set to 12. Therefore, every order of recital styles was present twice.

As stated earlier, potential participants had to be native speakers of German. They also had to not have impaired hearing or an aphasia. All participants were recruited via convenience sampling and were therefore members of academia: students in Bachelor's and Master's programmes, as well as research associates. Further information on the demographics can be found in chapter 5. Any participant needed to be at least 18 years old to be able to give consent. It was aimed for equal proportions of female to male participants to remove potential influences of gender.

4.4 Apparatus

The study was conducted in-person, in a lab-environment. Having more control over the setting increased the study's reliability (Aziz, 2017). Additionally, in-person was considered to be a more pleasant experience for the participants. In comparison to remote, they did not have to set up the equipment themselves.

Due to the Covid-19 pandemic, a hygiene concept had to be developed. The concept stipulated that the study was conducted in a big, regularly aired room. The examiner had to wear a face mask during the entire time. The participants were allowed to remove theirs during the data recording. It was aimed to space the sessions as far apart from each other as practical. All used surfaces were disinfected after each participant. Furthermore, the examiner took a corona rapid test on each day of the study.

Figure 4.1: Photo of the study set-up. On the left, the mouse for recording the Heart Rate, the eye mask, and the connection for the headphones. In the middle, a comfortable chair. On the right, the device for recording the Skin Conductance Response.

Before the sessions, the participants were provided with the information sheet and the consent form (cf. Appendix). In the beginning of the study, they had some time to ask questions and fill out the consent form.

To be able to fully focus on the recitals, the study was conducted in a closed eye setting. For this, blindfolds were provided, so the participants did not have to focus on closing their eyes. Furthermore, the participants were seated in a comfortable armchair to be able to relax. The audio was played via headphones, which the participants had to bring themselves. This ensured that the participants were already used to the headphones and increased the overall hygiene.

The Skin Conductance data was recorded with a Nexus 4[6] by Mind Media using their galvanic skin response (GSR) sensor[7]. The device is commonly used for biofeedback systems. It was connected to the examiner's computer via Bluetooth. The sensor's two electrodes were attached via hook and loop fasteners to the first phalanxes of the left index finger and the left middle finger.

For the recording of the Heart Rate, the computer mouse Mionix NAOS QG[8] was used. It was connected via USB. After the pilot study, the buttons were disabled to prevent unintended inputs. The participants had to place their right hand on the mouse. The unique feature of this mouse is that it has sensors for recording the users' physiological data. The heart frequency data was obtained via Photoplethysmography. Two kinds were provided: The raw sensor data with an unknown unit and the Heart Rate in beats per minute (bpm), derived from the raw data. These measured values are both

[6] https://www.mindmedia.com/en/products/nexus-4/
[7] https://www.mindmedia.com/en/products/sensors/skin-conductance-sensor/
[8] https://mionix.net/products/naos-qg

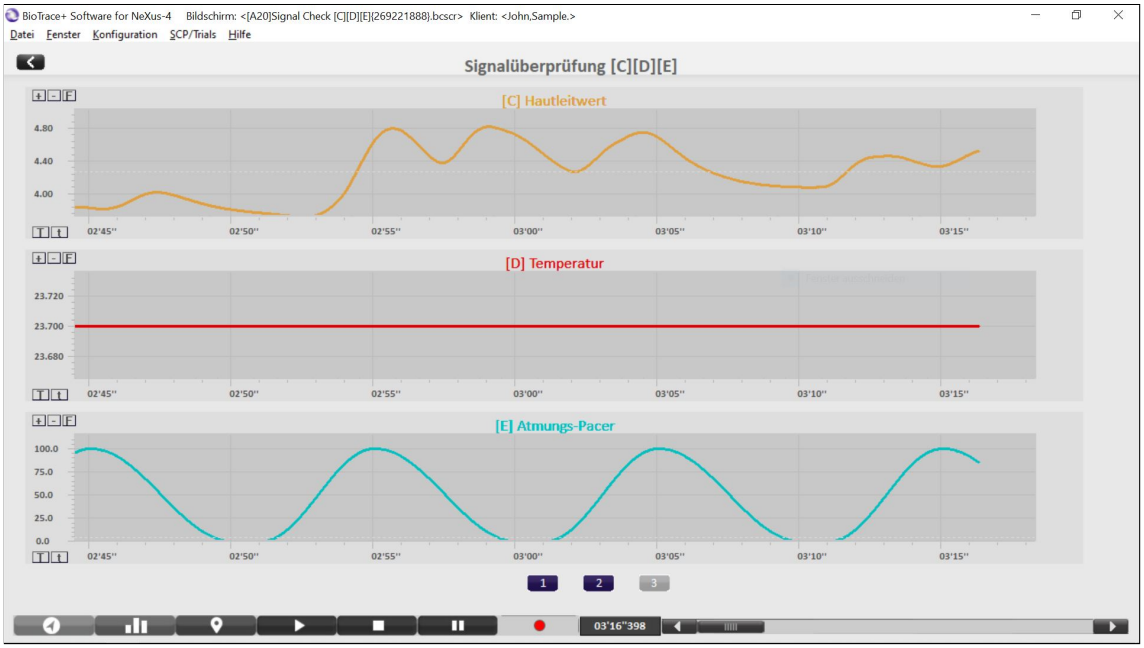


Figure 4.2: Screenshot of the software BioTrace+ while recording Skin Conductance data. The signal course of the Skin Conductance is depicted in orange in the first row. The other two measures below, Temperature and Respiration Pace, were not recorded in the study.

stored and evaluated. Even if the unit of the raw data is unknown, the signal can provide information about the physiological experience. In comparison to the Heart Rate, the data is unprocessed and therefore possibly more detailed. The raw data is called *Raw Mouse data* in the following. Besides the Heart Rate, the mouse can also record Skin Conductance data. However, tests by the Usability Research Group of the university had shown that the Skin Conductance data from the mouse is not satisfactory.

As a result of the pilot study, lower hand rests, i.e. a lower table and chair, were used for increased comfort. Figure 4.1 shows the final set-up.

4.4.1 Implementation

The collection of the physiological data was accomplished with the help of two programs. For the Skin Conductance data, BioTrace+[9] (Version 2018A1), the program by the manufacturer, was used. For the recording of the Raw Mouse data and the Heart Rate, a Python (Version 3.9.5) script framework was developed. The same framework also provided the audio stimuli. The framework can be used for further experiments in the field of measuring physiological reactions to stimuli. The scripts were developed with the aim of easy customizability.

BioTrace+ is the companion program to the Nexus devices by Mind Media. It can be used for simple data recording tasks but also for more complex biofeedback applications. Besides Skin Conductance, Electroencephalography data, Electrocardiogram data and more can be processed. However, only the very basic data collection functionalities for the Skin Conductance were used. All settings remained default. The sample rate was 32 Hz and the unit µSiemens (µS). Figure 4.2 shows the recording screen,

[9] https://www.mindmedia.com/en/products/biotrace-software/

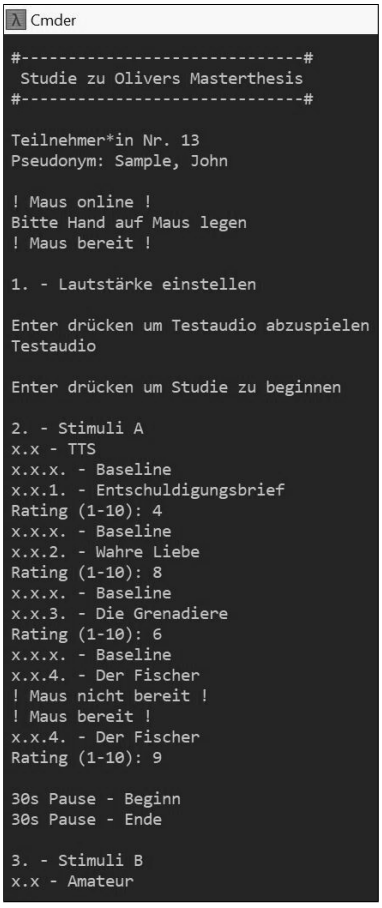


Figure 4.3: Screenshot of the Python script for playback and Heart Rate recording. The participant number is displayed in the second line. Below this, the pseudonym from the demographic questionnaire must be entered. Next, it is checked whether the mouse is connected. A test audio is played to set the volume and to get to know the tool. After confirming with Enter, the playback and recording is started. Following the start, information about the current stimulus is displayed. After an emotional stimulus, the rating of the participant has to be entered. It is checked if the rating is an integer between 1 and 10. The script also informs if the contact between mouse and participant was lost and about breaks.

Temperature and Respiration Pace were ignored. Following the recording, the entire session data, including timestamps, was exported as a tab-separated ASCII text file. Additionally to the timestamps and sensordata, the file included information about the client and session name, the date, the start time, the total duration, the sampling rate, and which sensor was used.

The Python script was used for displaying information for the researcher and handling the playback, and the data recording from the computer mouse. The script was self-written, based on online documentation and the mouse's API[10].

For the playback functionality, the package Pydub [11] was used. Since the neutral stimuli data was output into several files by MaryTTS, Pydub was used to concatenate the files into a single file. Afterwards, the single file was split up into 30 seconds long

[10] https://support.mionix.net/hc/en-us/articles/115000615526-Naos-QG-API
[11] http://pydub.com

snippets. This resulted in abrupt beginnings and ends of the stimuli. To smooth them out, a 1.5 seconds long fade-in and a 2 seconds long fade-out were added. Both values were subjectively chosen by the author.

Furthermore, the package was used to normalize the loudness of all audio files to -22.5 dBFS (decibels relative to full scale (Wettstein, 2018)). This was done to exclude varying levels of loudness as a confounding variable and to make the listening experience more pleasant for the participants.

The Python script selected the order of the three different recital styles automatically based on the participant's number. The number was persisted and got incremented by one after each participant. The order of the poems always got randomized before every style. After each recital, the script required to enter the participants rating. It verified that the rating was an integer value between 1 and 10.

For the recording of the data from the mouse, the script established a connection to the mouse via a WebSocket. A callback was executed on every message sent by the mouse. The messages contained the sensordata in the JSON-format. The Raw Mouse data was collected with a frequency of 21 Hz, the Heart Rate was provided with a frequency of 8 Hz. The data collection ran in a parallel process to not interfere with the audio playback.

The script first checked the type of message, i.e. "bioRaw" (Raw Mouse data) or "bioMetrics" (derived Heart Rate), and checked that the hand was touching the mouse and the Heart Rate was provided by the mouse. If this was the case, timestamps were added and the Raw Mouse data and Heart Rate data were saved.

With the help of Python multiprocessing events, an interrupt system was constructed that stopped the playback of a stimulus if the participant was not touching the mouse anymore. In case of such a disconnect, the current playback was stopped and the "Windows Hardware Remove" sound was played. This sound should be known by the participants and its meaning easily understandable. The script waited until the hand was on the mouse and the Heart Rate derivable again. This could take up to 20 seconds. Afterwards, the "Windows Hardware Insert" sound was played and the last stimulus was restarted.

Before the actual study, a 30 seconds long snippet of the neutral audio was played to test and adjust the volume, and the mouse status check functionality to introduce the participants to the study.

Besides the playback and recording functionalities, the script displayed information about the study and the participant (cf. figure 4.3). In the beginning, it asked for the participant's pseudonym, to easily link the data to the questionnaires. The script informed the examiner about the current stimulus, any mouse disconnects and ongoing breaks. The breaks were modeled according to the study design.

After successfully finishing the data recording, the script generated three output files. Two separate files were generated for the mouse data, i.e. the Raw Mouse data and the derived Heart Rate. The third file contained the participant's number, the pseudonym, the order of the recital styles and poems, the ratings, and the start and end times of every baseline and recital. The data was comma separated. The time in all three files was precise to the microsecond (µs). The precision of the timestamps was reduced in the following analysis.

4.5 Analysis

The collected data was analyzed in several ways, matching the respective types of data. The questionnaire yielded both quantitative and qualitative data. The qualitative data was manually aggregated and intersected with the quantitative data. The physiological data recording generated quantitative data, which was analyzed manually by visualizing it as well as in an automated way.

All questionnaire answers were digitized and transferred into an Excel sheet. The descriptive statistics median, standard deviation, maximum and minimum were calculated. For the nominal data, the distributions over all entries were calculated. The qualitative answers of the questionnaire were manually aggregated. The results were blended with the quantitative data to find patterns.

The participants' ratings were binned into three bins: Ratings between 1 and 4, ratings between 5 and 7, and ratings between 8 and 10. These bins are associated with a low, a medium, and a high rating. This was done to increase the amount of data points per bin and therefore be able to find more meaningful trends. Additionally, fine grained distinctions in the rating were considered to be not important for the discovery of trends.

The three kinds of physiological data were quantitative on a ratio scale. Therefore, they were treated in similar ways. The Raw Mouse data was z-score transformed before further processing. To compute the parameters of the time-series, a Python script was used. For each participant, the script calculated the local baseline mean per poem, using the appropriate preceding neutral stimuli. The baselines were used to normalize the data and therefore to transform absolute differences into relative ones.

The normalized sensor data was used to calculate the parameters amplitude, rise time, and half recovery time. To be able to manually analyze these, interactive boxplots of the data were generated. The plots were grouped by rating and poem, and by style and poem. The plots showed the median values and the first and third quartile. The boxplots' whiskers indicated the upper and lower fences. Additionally, outliers were visualized. As outliers classified were data points three times the inter-quartile range lower than the first or higher than the third quartile. The manual analysis had the goal to identify trends and correlations in the data. It was looked at trends in the outliers, the median values, and the value ranges between the upper and lower fences (the interfence range), and within the interquartile range.

For all participants, the signal course of the individual measures was plotted as a graph for each condition. The respective onset points, peaks, and half recovery points were also visualized. The aim of this visualization was the possibility to analyze the reaction of individual participants. In present thesis, the Heart Rate time-series of the poem "Die Grenadiere" was exemplary analyzed for one participant.

Afterwards, a more automated analysis was done. For the median, standard deviation, maximum, and minimum of the amplitudes, rise times, and half recovery times of each poem, one-way ANOVAs for the ratings and the styles were conducted. On the same data, Spearman's rank correlation coefficients were calculated. The goal was to find parameters that highly correlate with ratings. A high correlation shows

which kind of data is most indicative for a high rating. Spearman's rank correlation coefficient was chosen as metric since it is better suited for unbalanced data with outliers than other correlation metrics (Mukaka, 2012). Important to note is that high inter- and intra-individual differences were to be expected. The available data was influenced by a wide range of factors. Especially due to the long recording lengths, the exact influences likely cannot be pinpointed. This must be taken into account when interpreting the results.

To analyze the inter-individual differences in the physiological reactions, the 95% Confidence Intervals for each condition and measure were visualized.

Finally, heatmaps of the occurrence of peaks in the time-series, correlating with the waveform, i.e. amplitude, of the audio, were built. The peaks of all participants from all three measures were plotted on top of the corresponding waveforms. This was done for every condition. The goal here was to find correlations between the peaks and the audio signal.

4.5.1 Implementation

After the data recording, several scripts for the data analysis were run. They cleaned and truncated the data and calculated several different parameters for comparison. The scripts output the cut data, correlation coefficients and data visualizations.

Because the data was recorded as a continuous stream during the collection, the data outside of the stimulus windows needed to be removed. Some of the data entries were duplicates, i.e. identical timestamp and sensor value. All duplicates were removed. Python timestamps are precise to the microsecond. This level of precision was not needed for the analysis. Therefore, the timestamps got truncated to a precision of 10 milliseconds. With regards to the sampling rates, this level of precision was sufficient to distinguish the samples. The timestamps of the Skin Conductance data were relative to the start of the recording. They got converted to absolute timestamps.
The cleaned data was saved in a list for further processing. This was done for the Skin Conductance data, the Raw Mouse data, and the Heart Rate data. From the cleaned data, time-series were created. In the study, each of the four poems was presented in three recital styles. Additionally, all stimuli had a corresponding neutral stimulus. For all of these 24 stimuli, data from three measures was recorded, resulting in 72 time-series per participant.

After the data cleaning process, the data was normalized and the ratings were binned. For all neutral stimuli, the mean values of the three data recordings were computed. These means served as the local baselines for the respective stimuli and got subtracted from each data point. This was done on a per-participant basis. The expected range of values in the Raw Mouse data was unknown. Therefore, a z-score transformation, a common practice to improve comparability between time-series (Henderi et al., 2021), was applied on the data.

The data was then analyzed using NeuroKit2 (Makowski et al., 2021) for the Skin Conductance data and NeuroKit2 and SciPy (Harris et al., 2020) for the Raw Mouse and Heart Rate data. The applied algorithm was derived from the NeuroKit2 function

eda_process(). The function first filtered the input data and decomposed it into Skin Conductance Level and Response. Peaks and their parameters were searched for on this data and returned. To have the ability to tune more parameters and thereby gain more control over the analysis process, the function was recreated using the lower level functions.

As slow changes and strong shifts were more interesting than small, fast changes, the Skin Conductance data was first smoothed with a lowpass butterworth filter. This also reduced signal noise. A correlation between the unfiltered data and the amplitude of the audio was not found, further supporting the filtering. The smoothed data was decomposed into Skin Conductance Level and Skin Conductance Response using a convex optimization algorithm by Greco et al. (2016). On the Skin Conductance Response, all peaks were identified and a list with the peaks' amplitudes, their indices in the time-series, the onsets' indices, and the half recovery points' indices was created. The indices were converted to timestamps in seconds using the sampling rate and the rise times, and half recovery times calculated.

From this data, descriptive statistics were calculated. For the sensor values, amplitudes, rise times, and half recovery times, median, standard deviation, maximum and minimum values were calculated on a time-series basis.

A similar pipeline was built for both the Raw Mouse and Heart Rate data. The filter function from NeuroKit2 was reused, but the peaks were identified using SciPy. A decomposition was not necessary. From the data, the same set of descriptive statistics was calculated.

For both pipelines, the signal was filtered with the highcut set to 0.075 Hz and the filter order set to 4. For the Skin Conductance Response, amplitudes had to have a minimum height of at least 25% of the time series' highest amplitude. Peaks in the Raw Mouse and Heart Rate data required a minimum width of the sampling rate, i.e. one second, a distance to other peaks of twice the sampling rate, i.e. two seconds, and a prominence of one. These settings were selected, using bisection and manually reviewing the results. The highcut setting was further validated by transforming the data into frequency space using fast Fourier transforms and analyzing the distribution.

The boxplots for the trend analysis were created using the Python package Plotly (Plotly Technologies Inc., 2015). It created an interactive HTML page per plot, allowing the manipulation of the plots, showing values on hover. For the visualizations, the standard settings were used.

The time series of each physiological measure per condition and participant were visualized with Matplotlib (Hunter, 2007). Onset points, peaks, and half recovery points were plotted in different colors on top of the signal course.

The earlier calculated descriptive statistics were extended with additional information: Besides the mean frequency, the text's EmoScore was added. The one-way ANOVAs and the Spearman's rank correlation coefficients regarding the ratings were calculated for each poem with a Python script and output. Outliers were removed beforehand.

The 95% Confidence Intervals for all styles and all data types, i.e. Raw Mouse data, Heart Rate and Skin Conductance Response, were computed with the help of SciPy and plotted with Matplotlib.

The heatmaps were again generated with Matplotlib. Using Parselmouth (Jadoul et al., 2018) the audios' waveforms were extracted and plotted. The waveform visualizes the sound pressure on an arbitrary scale between 0 and 1. The left audio channel is represented by positive values, the right audio channel by negative values. The peaks of three different measures, i.e. Raw Mouse data, Heart Rate, and Skin Conductance Response, in each condition were plotted on the same graph in different colors. The vertical positions of the peaks were normalized to the maximum height of the audio amplitudes.

5 Results

The description of the collected data is subdivided into the demographic data, the questionnaire, and the physiological data. The last section reports on miscellaneous results of all the collected data.

5.1 Demographics

As described before, the data was collected from 12 participants. Six of the participants identified themselves as male, the other six identified themselves as female. The age ranged from 20 years old to 32 years old. The mean age was 25.3 years with a standard deviation of 4.0 years. All participants were members of academia. Four participants worked as research associates at a university. Seven participants were students at a university and one participant was a recent alumni of a university without a new occupation yet.

5.2 Questionnaire

The gathered questionnaire data is presented in the order of the questions in the questionnaire.

Ratings

In the Text-To-Speech style, the recital of **“Entschuldigungsbrief”** has gotten a mean rating of 3.33, it has a standard deviation of 2.29. The minimum rating of the recital is 1, the maximum rating is 8. The mean rating of **“Der Fischer”** is 2.92, the standard deviation is 1.93. The minimum is 1 and the maximum is 7. **“Die Grenadiere”** has a mean rating of 3.17 and a standard deviation of 1.82. The minimum rating is 7, the maximum rating 1. The mean rating that the recital of **“Wahre Liebe”** has gotten is 2.83, the standard deviation is 1.86. The minimum is 1 and the maximum is 7.
In this style, “Entschuldigungsbrief” received the highest rating, “Die Grenadiere” is the poem with the lowest standard deviation in the ratings.

Table 5.1: Mean ratings of the recitals in the three different recital styles for each poem. The highest mean rating per poem is highlighted in bold.

Poem	Style		
	Text-To-Speech	Amateur	Professional
Entschuldigungsbrief	3.33	5.25	**7.42**
Der Fischer	2.92	4.92	**8.25**
Die Grenadiere	3.17	5.08	**8.08**
Wahre Liebe	2.83	4.83	**8.33**

In the amateur style, **"Entschuldigungsbrief"** received a mean rating of 5.25 with a standard deviation of 1.69. The minimum rating is 3 and the maximum rating is 8. The recital of **"Der Fischer"** has a mean rating of 4.92, the standard deviation is 1.61. The minimum is 2 and the maximum is 7. **"Die Grenadiere"** got a mean rating of 5.08 with a standard deviation of 1.85. The minimum rating is 7 and the maximum rating is 2. **"Wahre Liebe"** received a mean rating of 4.83 with a standard deviation of 1.77. The minimum is 1 and the maximum is 7.
The recital of "Entschuldigungsbrief" is rated highest, while the recital of "Der Fischer" has the lowest standard deviation in the amateur style.

In the professional style, **"Entschuldigungsbrief"** has gotten a mean rating of 7.42, it has a standard deviation of 1.44. The minimum rating is 5 and the maximum rating is 10. The mean rating of **"Der Fischer"** is 8.25, the standard deviation is 1.42. The minimum rating is 5 and the maximum rating is 10. **"Die Grenadiere"** has a mean rating of 8.08 with a standard deviation of 1.11. The minimum rating is 7 and the maximum rating is 10. **"Wahre Liebe"** received a mean rating of 8.33 with a standard deviation of 0.94. The minimum rating is 7, the maximum 10.
From the professional style, but also overall, the recital of "Wahre Liebe" is rated the highest with the lowest standard deviation.

Table 5.1 summarizes the mean ratings again. Overall, the recitals from two different poems were rated the highest in the three styles: "Entschuldigungsbrief" was rated highest in the Text-To-Speech style and the amateur style, and "Wahre Liebe" in the professional style. The mean ratings allow for a clear distinction between the different styles. The mean ratings of the Text-To-Speech style range from 2.83 to 3.33, the mean ratings of the amateur style range from 4.83 to 5.25 and the mean ratings of the professional style from 7.42 to 8.33.
In both the Text-To-Speech and the amateur style, "Wahre Liebe" received the lowest mean ratings. In the professional style, "Entschuldigungsbrief" received the lowest mean rating. The widest range of ratings received "Entschuldigungsbrief" in the Text-To-Speech style (between 1 and 8), this is accompanied by the highest standard deviation in the data set (2.29). The smallest ranges are for "Die Grenadiere" and "Wahre Liebe" in the professional style (between 7 and 10), however, the ratings of "Wahre Liebe" have a lower standard deviation (0.94). The lowest standard deviations can be found in the professional style.
The Average Observed Agreement of the ratings is 0.17. When binning the rating data into the three bins (1-4, 5-7 and 8-10), it increases to 0.53.

Descriptions

10 of the 12 participants described the **Text-To-Speech style** as a mechanical, computer generated voice. One of them compared it to the voice of Google Translate. A monotonous, neutral intonation, and pronunciation difficulties were mentioned in the responses 11 times. On the other hand, one participant praised the intonation of the voice: "for a computer voice, nevertheless partly surprisingly good intonation" (*für Computer-Stimme dennoch teils überraschend gute Betonung*). 12 times, amelodic, arythmic speech was mentioned. The participants described it as "choppy, frail" (*Abgehackt, brüchig*), "partly too fast/too slow" (*teilweise zu schnell/zu langsam*) and "hectic and not pleasant" (*hektisch u. nicht angenehm*). Five participants identified the

gender of the speaker as male. Four participants deemed the speaker poor but funny sounding: "very unpleasant, but also very funny" (*sehr unangenehm, aber auch sehr witzig*). Two think that the Text-To-Speech style is understandable but that it is missing emotions.

The speaker of the **amateur style** was described six times as a young male speaker. Also six times mentioned was a higher pitch in comparison to the other styles. Seven times, the participants called the voice flat and monotonous but with usually a good intonation. Five participants said that the amateur sounds like they did not receive any professional training: "good but amateurish speaker" (*guter aber laienhafter Sprecher*), "'normal' voice" (*'normale' Stimme*). Five times, the speaker was described as not enthusiastic enough: "a little passionless" (*ein wenig leidenschaftslos*). A bit hectic, the speaker was called four times. Three times, the performance was called rehearsed and struggling. Two participants think the speaker sounds nasal. Once it was noted that "The verse meter and the rhymes are uncomfortably brought to the foreground." (*Das Versmaß und die Reime sind unangenehm in den Vordergrund gerückt.*). One participant remarked that the dramaturgical pauses were too short.

The speaker of the **professional style** was seven times described as having an older, mature voice. 10 times, the voice was described as deep, sonorous, and dark. 11 times, it was called melodic, emotional, calm, and pleasant. The intonation was mentioned six times in the answer: "perfect pronunciation and intonation" (*perfekte Aussprache und Betonung*). Five times, the professional was called professional: "sounds like a professional audio book speaker" (*klingt nach einem professionellen Hörbuchsprecher*), "knows how to use the potential of the voice" (*weiß mit den Möglichkeiten der Stimme umzugehen*). Three participants described it as expressive. Three more stated that the speaker has a slow rate of speech. One mentioned that it "could have been presented even more intensively" (*hätte noch intensiver vorgetragen werden können*). One participant stated in their description that "the other sounds that the mouth makes [...] were clearly audible" (*die andere Geräusche, die der Mund macht [...] waren deutlich zu hören*).

When asked what the participants liked about the **Text-To-Speech style**, two did not answer the question. Two explicitly stated they did not like anything. Three people said that they liked the clear pronunciation and the intonation of the Text-To-Speech. Five times, it was answered that the participants liked that the Text-To-Speech sounded funny and ironic: "I think it is brilliant that the person doesn't understand the poems at all and reads them anyway." (*Ich finds genial, dass die Person die Gedichte garnicht versteht und sie trotzdem liest.*). Two participants liked that the speaker sounded soulless according to them.

Regarding what is liked about the **amateur style**, intonation, pronunciation, and fluency were mentioned nine times. One participant liked the irony in the voice. Another participant liked that the speaker sounds "young, approachable, sympathetic" (*jung, nahbar, sympathisch*).

For the **professional style**, pronunciation and intonation were mentioned 10 times: "Intonation according to the text / mood" (*Betonung entsprechend des Textes/der Stimmung*). The deep pitch of the voice was mentioned 6 times. One participant stated about the pitch: "Voice pitch: calm but intense" (*Stimmlage: ruhig aber intensiv*). This coincides with two other participants which had comforting and lively in their descriptions. Four times, the acting component of the speech was brought up: "not a simple speaker, probably more of an actor" (*kein reiner Sprecher, wahrscheinlich eher Schauspieler*),

the four participants liked the resulting harmony. Two participants liked the breaks in the recitals, two more liked the rate of speech. Three participants indicated that they simply liked the overall recitals: "high quality of the recital" (*hohe Qualität der Rezitation*).

Three participants had nothing negative to say about the **Text-To-Speech style**. The other 10 participants agreed on three categories of features they dislike. Six times, wrong pronunciations and inconsistent intonation was listed in the answers: "wrong intonation, partly even incomprehensible" (*falsche Betonungen, z.T. sogar unverständlich*), another participant argued that the intonation impacts the suspense: "flat and no emphasis -> therefore no suspense" (*flache und keine Betonung -> daher keine Spannung*). Six participants disliked that the voice sounded unnaturally and distorted according to them. They thought that the speaker was emotionless, cold, and mechanic. A monotonous melody and a constant rate of speech was mentioned five times.

The answers to the same question on the **amateur style** included the intonation six times: "intonation sometimes better possible (more adapted)" (*Betonung manchmal noch besser möglich (angepasster)*) and the pronunciation: "pronunciation sometimes unclear" (*Aussprache manchmal undeutlich*). Three times, the participants considered the speaker to be too monotonous: "pitch remains the same for all texts and passages" (*Stimmlage bleibt bei allen Texten und Textstellen gleich*). Four criticized that the amateur is "inexperienced" (*Ungeübtheit*). Three times, the breaks, and the rate of speech were mentioned. One participant argued that the amateur was: "too fast [and] a lot of emotion is lost as a result" (*zu schnell, geht viel an Emotionen verloren dadurch*). For two people, the amateur's voice was too high or too weak to fit the poems.

For three participants, the **professional style** was too slow: "Sometimes very slow speech rate" (*teilweise sehr langsames Sprechtempo*). Two participants considered the voice too deep and too manly. Both also criticized that non-speech mouth sounds were audible in the recordings. For one participant, the speaker was too good: "very good voice distracts a little from the content" (*sehr gute Sprecherstimme lenkt etwas vom Inhalt ab*). One participant did not like the voice acting of the speaker and another one thinks the speaker sounds too negative and melancholic. Three participants had nothing negative to say about the professional style.

Overall, the participants mentioned the pronunciation and intonation, as well as breaks and the rate of speech as positive and negative. For the professional style specifically, the professionalism was both liked and disliked.

Comparing the differences between the styles, the participants mentioned the pitch and timbre nine times: "Voice pitch: 1 [the professional] low, 2 [the amateur] high, 3 [the Text-To-Speech] comically undefined" (*Stimmlage: 1 tief, 2 hoch, 3 komisch undefiniert*). Seven times, the differences in intonation were mentioned: "expressiveness/intonation" (*Ausdrucksstärke/Betonung*). Six times, the rate of speech was part of the answer. Six people mentioned that the three speakers differ in their professionalism: "Really deep into poetry 1 [the professional] yes, 2 [the amateur] a bit, 3 [the Text-To-Speech] not at all." (*Im Gedichtgeschehen wirklich tief drin 1 ja, 2 bisschen, 3 gar nicht.*). The emotionality was mentioned three times: "Attempt to convey the mood of the writer" (*Versuch die Stimmung des Schreibenden zu vermitteln*). The age difference between the speakers was pointed out three times. Two times, differences in the

humanness were mentioned. Once, a participant thought that the styles are differently pleasant. One participant saw the difference in the pronunciation, another one in their personal preference.

Preferences

All 12 (100%) participants stated that they most likely would listen to another recital from the professional speaker. No participants chose the Text-To-Speech speaker or the amateur speaker.

When asked to sort the speakers from most to least preferred, 10 participants (83.33%) put the professional speaker in **first place**, two participants (16.67%) chose the Text-To-Speech speaker, and zero participants (00.00%) chose the amateur speaker. For **second place**, eight participants (66.67%) answered with the amateur speaker, two participants (16.67%) chose the professional speaker, and two participants (16.67%) chose the Text-To-Speech speaker. Last ranked, in **third place**, was the Text-To-Speech speaker by eight participants (66.67%), the amateur speaker by four participants (33.33%), and the professional speaker by zero participants (00.00%). The results are shown in table 5.2.

For **“Entschuldigungsbrief”**, seven participants (58.33%) considered the professional the best fitting speaker for the poem, three participants (25.00%) the amateur, and two participants (16.67%) the Text-To-Speech speaker. For **“Der Fischer”**, 11 participants (91.67%) preferred the professional speaker, one participant (08.33%) the amateur, and zero participants (00.00%) the Text-To-Speech speaker. 10 participants (83.33%) considered the professional the best fitting speaker for **“Die Grenadiere”**, one participant (08.33%) the amateur, and one participant (08.33%) the Text-To-Speech speaker. For **“Wahre Liebe”**, the professional was chosen by 10 participants (83.33%), the amateur by one participant (08.33%), and the Text-To-Speech speaker by one participant (08.33%) as well.
For all four poems, the majority chose the professional speaker as the best fitting and the Text-To-Speech speaker as the worst fitting. For “Entschuldigungsbrief”, the agreement was the lowest and for “Der Fischer” the highest.

Table 5.2: Ranking of the styles by preference in percent of the participants. The highest percentage per rank is highlighted in bold.

Style	Rank		
	First	Second	Third
Text-To-Speech	16.67%	16.67%	**66.67%**
Amateur	00.00%	**66.67%**	33.33%
Professional	**83.22%**	16.67%	00.00%

Emotions

Regarding the main emotions of each poem, four participants answered with emotions that are not part of Plutchik's set of basic emotions. For a consistent evaluability, these emotions had to be manually mapped to Plutchik's emotions. This went without any problems. For example, "revulsion" (*Abscheu*) was mapped to disgust, "exhilaration" (*Heiterkeit*) to joy.

The poem **"Entschuldigungsbrief"** was associated with positive emotions 11 times and one time with a negative emotion. The participants chose surprise four times, joy three times, trust two times, anticipation two times, and disgust one time as the main emotion. Out of the answers for **"Der Fischer"**, nine were a positive emotion and three were a negative emotion. As the most prominent emotion, anticipation was selected five times, fear three times, joy two times, surprise one time, and trust one time as well. Regarding **"Die Grenadiere"**, the participants answered five times with a positive emotion and seven times with a negative emotion. They chose sadness five times, trust five times, fear once, and disgust once. **"Wahre Liebe"** got nine positive associations and five negative ones. Of these associations, three were surprise, three were trust, three were joy, two were disgust, and one was anger.

Overall, the poems were more frequently associated with positive emotions (34 times) than negative emotions (16 times). Trust was the most common one (11 times), anger the least common one (one time). The least common positive emotion (anticipation, seven times) occurred more often than the most common negative emotion (sadness, five times). All of Plutchik's eight basic emotions were mentioned at least once.

In the automated emotion analysis (cf. section 4.2), the emotions with the highest probability in **"Entschuldigungsbrief"** were trust and joy, while the majority of the participants chose surprise. For **"Der Fischer"**, the analysis indicated joy, but the participants mostly chose anticipation. For **"Die Grenadiere"**, the most probable emotion in the analysis was sadness, the participants' most selected emotions were sadness and trust. For **"Wahre Liebe"** the emotion was anticipation in the analysis, while the participants most associated emotions were surprise, trust, and joy.

The best in expressing the associated emotion of **"Entschuldigungsbrief"**, the participants nine times (75.00%) considered the professional. Three participants (25.00%) chose the amateur, and zero (00.00%) the Text-To-Speech. For **"Der Fischer"**, nine participants (75.00%) replied with the professional, two participants (16.67%) with the amateur, and one participant (08.33%) with the Text-To-Speech. For **"Die Grenadiere"**, 11 participants (91.67%) said that the professional was the best in expressing the associated emotion, one participant (08.33%) said it was the amateur, and zero

Table 5.3: The best styles in expressing the associated emotion per poem in percent of the participants. The highest percentage per poem is highlighted in bold.

Poem	Style		
	Text-To-Speech	Amateur	Professional
Entschuldigungsbrief	00.00%	25.00%	**75.00%**
Der Fischer	08.33%	16.67%	**75.00%**
Die Grenadiere	00.00%	08.33%	**91.67%**
Wahre Liebe	08.33%	08.33%	**83.33%**

participants (00.00%) the Text-To-Speech. 10 participants (83.33%) considered the professional as the best for **"Wahre Liebe"**, one participant (08.33%) the amateur, and one participant (08.33%) the Text-To-Speech. The results are shown in table 5.3
Overall, the majority considered the professional speaker the best in expressing the emotion and the Text-To-Speech speaker the worst. The answers for "Der Fischer" had the lowest agreement and the answers for "Die Grenadiere" the highest.

When asked to indicate how the speakers conveyed the emotions, the participants mentioned seven different ways. One participant did not provide any answer. Generally, changing and varying the following characteristics was perceived to show different emotions. Nine times, the emphasis and intonation of syllables, words, and phrases were mentioned. One participant described this as "An attempt was made to adopt the intonation and style of performance of a person who is having these emotions at the moment." (*Es wurde versucht die Betonung und die Vortragsweise einer Person anzunehmen, die diese Gefühle gerade hat.*). The pitch of a speaker's voice was mentioned eight times. Breaks, for example for creating suspense (*künstlicher Spannungsaufbau*), were mentioned seven times. Four participants listed the speed or rate of speech. Pronunciation was listed by two participants. Two participants identified the volume. Two participants described the speakers' techniques simply as emotional, the participants got "captivated" (*mit in den Bann gezogen*).

Content

Judging by its content, eight participants (66.67%) considered "Die Grenadiere" the **most** emotional poem, two participants (16.67%) "Wahre Liebe", one participant (08.33%) "Der Fischer", and one participant (08.33%) "Entschuldigungsbrief". Four participants (33.33%) considered "Der Fischer" the **second most** emotional poem, four participants (33.33%) "Wahre Liebe", three participants (08.33%) "Die Grenadiere", and one participant (08.33%) "Entschuldigungsbrief". Six participants (50.00%) chose "Entschuldigungsbrief" as the **third most** emotional poem, four participants (33.33%) chose "Wahre Liebe", two participants (16.67%) chose "Der Fischer", and zero participants (00.00%) "Die Grenadiere". "Der Fischer" was answered as the least, i.e. **fourth most**, emotional poem by five participants (41.67%), "Entschuldigungsbrief" by four participants (33.33%), "Wahre Liebe" by two participants (16.67%), and "Die Grenadiere" by one participant (08.33%).
The participants had the highest agreement on the most emotional poem and the lowest agreement on the second most emotional poem.

Table 5.4: Ranking of the poems by preference in percent of the participants. The highest percentage per rank is highlighted in bold.

Poem	Rank			
	First	Second	Third	Fourth
Entschuldigungsbrief	25.00%	33.33%	33.33%	08.33%
Der Fischer	**33.33%**	**41.67%**	08.33%	25.00%
Die Grenadiere	25.00%	08.33%	**41.67%**	25.00%
Wahre Liebe	16.67%	16.67%	16.67%	**41.67%**

Four participants (33.33%) stated that they liked “Der Fischer” the **most** based on its content. Three participants (25.00%) chose “Entschuldigungsbrief”, three participants (25.00%) “Die Grenadiere”, and two participants (16.67%) “Wahre Liebe”. Five participants (41.67%) selected “Der Fischer” as the **second most** liked poem, four participants (33.33%) “Entschuldigungsbrief”, two participants (16.67%) “Wahre Liebe”, and one participant (08.33%) “Die Grenadiere”. “Die Grenadiere” was chosen as the **third most** liked poem by five participants (41.67%), “Entschuldigungsbrief” by four participants (33.33%), “Wahre Liebe” by two participants (16.67%), and “Die Grenadiere” by one participant (08.33%). Five participants (41.67%) considered “Wahre Liebe” the least, i.e. **fourth most**, liked poem, three participants (25.00%) “Der Fischer”, three participants (25.00%) “Die Grenadiere”, and one participant (08.33%) “Entschuldigungsbrief”. The results are summarized in table 5.4.

Bodily Changes

When the participants were asked about physical changes, three of them stated that they did not notice any changes. Three participants reported increasing relaxation over time, of which one “at some point [...] realised that [they] had forgotten to breathe in deep enough.” (*irgendwann habe ich gemekt, dass ich vergessen hatte tief genug zu atmen*). Four participants were bored or nervous during the recitals of the Text-To-Speech style and the amateur style but more relaxed during the professional style: “From impatience (speakers 1+2 [amateur and Text-To-Speech]) to excitement and enjoyment while listening (speaker 3 [professional])” (*Von Ungeduld (Sprecher 1+2) zu Spannung und Genuss beim Hören (Sprecher 3)*). One participant could only relax during the breaks and was stressed during the recitals. Another participant got tired over time.

Concerning changes of the participants’ attention, one participant did not notice any changes. Eight participants stated that their attention differed between the three styles. The attention was the lowest during the Text-To-Speech style: “Attention to text and performance decreased during speaker 3 [Text-To-Speech]” (*Aufmerksamkeit auf Text und Vortrag während Sprecher 3 nahm ab*) and the amateur style: “[...] during the 2nd speaker [amateur] it was difficult to stay attentive” (*[...] während dem 2. Sprecher war es schwierig aufmerksam zu bleiben*) and the highest during the professional style: “Higher attention with speaker 3 [professional]” (*Höhere Aufmerksamkeit bei Sprecher 3*). Two participants recognized changes in their attention correlating with bodily changes. They noticed positive changes: “I became more and more attentive, which was also [...] ‘arriving’ in the situation and the eye mask.” (*Ich wurde immer aufmerksamer, was auch [...] am ‘Ankommen’ in der Situation und der Augenmaske [lag].*), as well as negative changes: “In the middle, my attention was somewhat reduced due to the beginning tiredness” (*In der Mitte war meine Aufmerksamkeit aufgrund der einsetzenden Müdigkeit etwas reduziert*). One participant mentioned getting distracted by outside events, i.e. hearing birds chirping or hearing the experimenter move.

Regarding any specific technique helping with concentration, seven participants did not give an answer. Three participants stated that they focused on certain properties to help stay concentrated. For example, they focused on the voice, and the technique and pronunciation. One participant answered that they did light tapping movements with their fingers to keep themselves concentrated. Another one stated that they kept their eyes actively shut even though wearing the eye mask.

Out of the 12 participants, eight did not know any of the poems previous to the study. Four participants already knew "Der Fischer". Two of them from school, two did not provide from where they knew it. The other three poems were unknown to everyone.

Comments

The question if objective ratings of poem recitals are possible was answered by all participants. Four participants thought that it is **possible**, one of them did not justify their answer. The rest argued that people with similar background should have similar expectations for good recitals, especially regarding the emphasis and pronunciation: "[Because] all people (of a cultural space/language space) have had similar experiences with the connection of language, speech, character, emotionality of individuals." (*[Weil] alle Menschen (eines kulturellen Raumes/eines Sprachraumes) ähnliche Erfahrungen bei den Zusammenhang von Sprache, Sprechweise, Charakter, Emotionalität von Individuen gemacht haben.*). For an objective evaluation, eight different features were suggested by the participants. The most suggested features include breaks, rhythm, and speed, emphasis, and pitch. Less frequently, volume and sympathy were mentioned. Once each, "dialect freeness" (*Dialektfreiheit*) and "emotional possibilities" (*emotionale Möglichkeiten*) were mentioned. Four participants believed that it is **in parts possible** to rate objectively. They thought that many people are likely used to a certain recital style: "because we are accustomed to this way of speaking through media consumption, theater, radio, etc." (*weil wir durch Medienkonsum, Theater, Radio, etc. diese Art des Sprechens gewöhnt sind.*) but "[...] quality is comprised of objective and subjective characteristics." (*[...] Qualität sich aus objektiven und subjektiven Merkmalen zusammensetzt.*). Three participants thought that objective ratings of recitals are **not possible**, since people have different taste and subjective experiences: "[Because] different people, with different experiences and preferences, also perceive recitals differently." (*[Weil] unterschiedliche Personen, mit unterschiedlichen Erfahrungen und Vorlieben auch Rezitationen unterschiedlich wahrnehmen.*). Another participant rejected objective ratings as well: "one hundred percent objectivity is an illusion." (*hunderprozentige Objektivität ist eine Illusion.*).

The opportunity to provide comments was used by eight participants. Five participants indicated that they liked participating in the study. For example, one participant answered "I had great fun participating!" (*Ich hatte großen Spaß an der Teilnahme!*). Two participants stated that they think that rating recitals objectively is inherently difficult. Two other participants considered the tasks at hand rather difficult, one of them stated "Generally very difficult to concentrate on (evaluating) the recital but not the text." (*Generell sehr schwierig, sich auf die Rezitation zu konzentrieren (diese zu bewerten), nicht aber auf den Text.*). Three participants commented on the Text-To-Speech style. Two in a negative way: "Fortunately, there are better computer voices." (*Zum Glück gibt es bessere Computerstimmen.*), but one participant said "The tape voice was not so bad in its intonations." (*Die Tonbandstimme war garnicht so schlecht in ihren Betonungen.*) and stressed that it "[...] leaves room for one's own interpretation" (*Außerdem lässt sie Raum für die eigene Interpretation*).

5.3 Physiological Data

In the following section, the trends, discovered in the manual boxplot analysis, are presented. First, the amplitude, rise time, and half recovery time data from section 4.5 sorted by rating and poem, and afterwards, the data sorted by style and poem are described.

In the boxplots, it was looked for trends in the outliers, the median values and the value ranges between the upper and lower fences (the interfence range) and within the interquartile range.

Figure 5.1a through figure 5.4b show the boxplots of the data. The first row always shows the amplitude data, the second row the rise time data, and the third row the half recovery time data. In the columns, the poems can be found in alphabetical order (i.e. “Entschuldigungsbrief”, “Der Fischer”, “Die Grenadiere”, and “Wahre Liebe”). The blue plots are of the Text-To-Speech style or bin 1. The red plots are of the amateur style or bin 2. The green plots are of the professional style or bin 3.

Rating x Poem

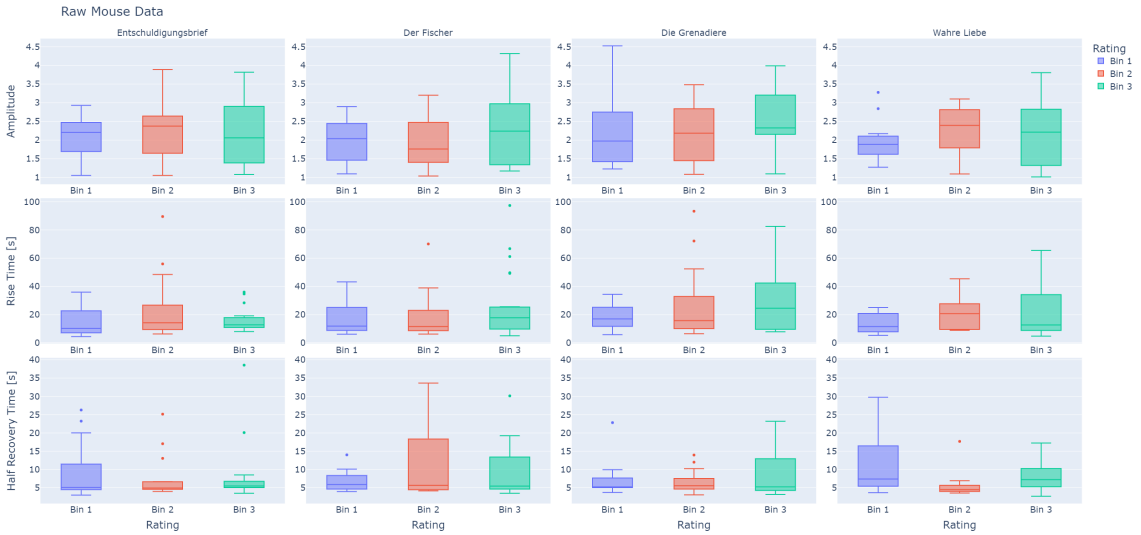
Figure 5.1a shows the **Raw Mouse data**, sorted by rating and poem.

The amplitude data of the poem **“Entschuldingungsbrief”** shows no trends in the median, in the interfence range, or in the outliers. In the interquartile range, it has an upward trend (bin 1: 0.78, bin 2: 0.99 , bin 3: 1.51). The rise times show a trend only in the number of outliers above the upper fence (bin 1: 0, bin 2: 2, bin 3: 3). The half recovery times have a downward trend in the interquartile range (bin 1: 7.00 s, bin 2: 2.05 s, bin 3: 1.63 s).

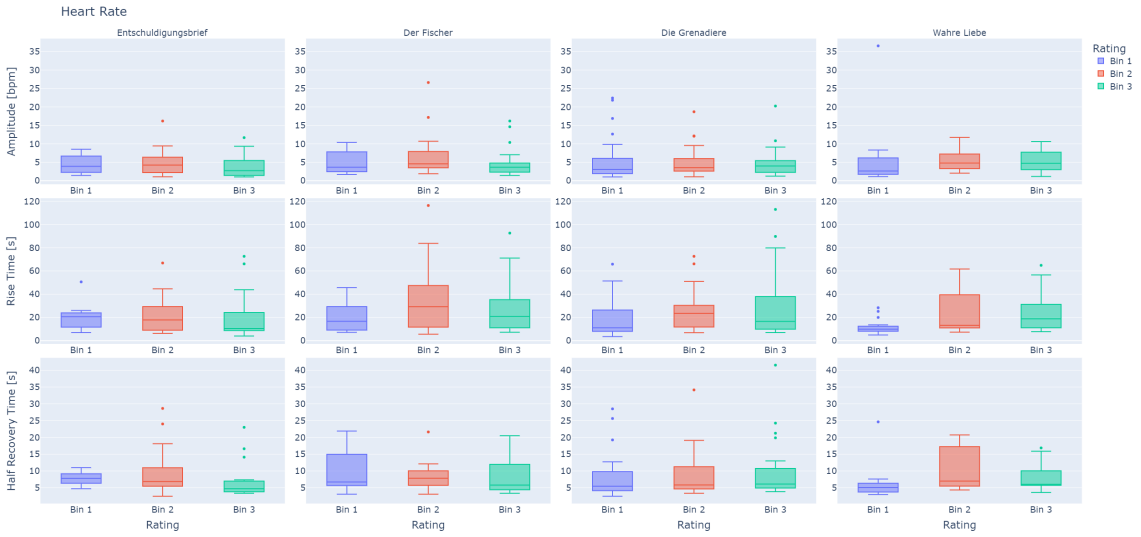
“Der Fischer” has two trends in the amplitude values, both are upward. The interquartile range (bin 1: 0.98, bin 2: 1.07, bin 3: 1.63) and the interfence range (bin 1: 1.80, bin 2: 2.17, bin 3: 3.14) go up the higher the rating is. The outliers in the rise times get more with higher ratings (bin 1: 0, bin 2: 2, bin 3: 4), the maximum outlier value increases as well (bin 1: -, bin 2: 70.09 s, bin 3: 97.32 s). Additionally, the interfence range of the rise times decreases (bin 1: 37.05 s, bin 2: 32.68 s, bin 3: 20.55 s). For the half recovery times, the median decreases (bin 1: 5.91 s, bin 2: 4.47 s, bin 3: 4.36 s).

The poem **“Die Grenadiere”** has one trend in its amplitude values. The median increases with higher values (bin 1: 1.98, bin 2: 2.19, bin 3: 2.33). Both the interfence range (bin 1: 28.55 s, bin 2: 45.86 s, bin 3: 74.59 s) and the interquartile range (bin 1: 13.45 s, bin 2: 26.38 s, bin 3: 32.73 s) of the rise times increase. The same is true for the half recovery times: the interfence range (bin 1: 6.23 s, bin 2: 7.18 s, bin 3: 20.03 s) and in the interquartile range (bin 1: 2.61 s, bin 2: 2.91 s, bin 3: 8.68 s) go up.

There are two consistent trends in the amplitude data of **“Wahre Liebe”**. The interquartile range (bin 1: 0.48, bin 2: 1.02, bin 3: 1.50) and the interfence range (bin 1: 0.90, bin 2: 2.00, bin 3: 2.79) increase with higher ratings. For the rise times regarding this poem, the interfence range (bin 1: 19.68 s, bin 2: 36.45 s, bin 3: 60.73 s), as well as the interquartile range (bin 1: 13.02 s, bin 2: 18.10 s, bin 3: 25.34 s) increase. The half recovery times show no trends.



(a) Boxplots of the Raw Mouse data, sorted by rating and poem.



(b) Boxplots of the Heart Rate, sorted by rating and poem.

Figure 5.1: Boxplots of the Raw Mouse data (a) and the Heart Rate (b), sorted by rating and poem. Each column represents a poem. The first row depicts the amplitudes (in (b) in beats per minute), the second row the rise times in seconds, and the third row the half recovery times in seconds. The blue boxplots are of rating bin 1, the red boxplots of bin 2, and the green boxplots of bin 3. The boxplots visualize the median, the first and third quartile, the lower and upper fences, and outliers. Outliers are data points three times the inter-quartile range lower than the first or higher than the third quartile.

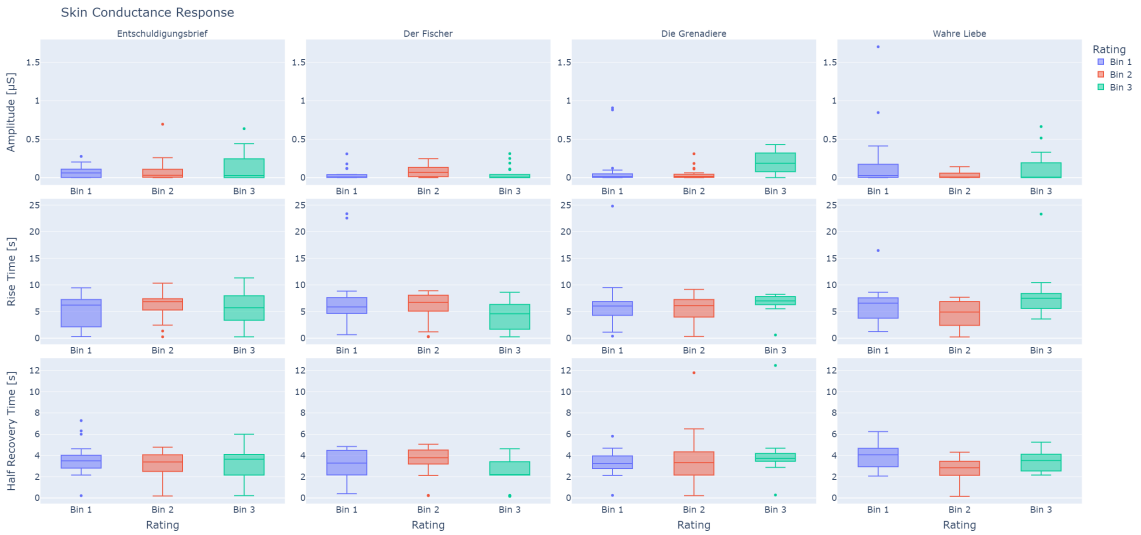


Figure 5.2: Boxplots of the Skin Conductance Response, sorted by rating and poem. Each column represents a poem. The first row depicts the amplitudes in microsiemens, the second row the rise times in seconds, and the third row the half recovery times in seconds. The blue boxplots are of rating bin 1, the red boxplots of bin 2, and the green boxplots of bin 3. The boxplots visualize the median, the first and third quartile, the lower and upper fences, and outliers. Outliers are data points three times the inter-quartile range lower than the first or higher than the third quartile.

The **Heart Rate**, sorted by rating and poem, is visualized in figure 5.1b.
Of the poem **"Entschuldingungsbrief"**, the interquartile range of the amplitude values decreases with increasing rating (bin 1: 4.44 bpm, bin 2: 4.19 bpm, bin 3: 4.05 bpm). The rise times' median goes down with higher ratings (bin 1: 20.63 s, bin 2: 17.88 s, bin 3: 10.5 s). The number of outliers above the upper fence increases (bin 1: 1, bin 2: 1, bin 3: 2), as well as their maximum value (bin 1: 50.63 s, bin 2: 66.88 s, bin 3: 72.63 s). The half recovery times have a downward trend in the median (bin 1: 7.81 s, bin 2: 6.88 s, bin 3: 4.75 s) and an upward trend in the number of outliers above the upper fence (bin 1: 0, bin 2: 2, bin 3: 3).

"Der Fischer" has two trends in the amplitude values. The number of outliers above the upper fence (bin 1: 0, bin 2: 2, bin 3: 3) increases and the interquartile range (bin 1: 5.38 bpm, bin 2: 4.42 bpm, bin 3: 2.46 bpm) is decreasing. Both the rise times and the half recovery times show no consistent trends.

The amplitude values' median of **"Die Grenadiere"** has an upward trend (bin 1: 3.06 bpm, bin 2: 3.52 bpm, bin 3: 4.00 bpm). Furthermore, the interfence range (bin 1: 8.85 bpm, bin 2: 8.49 bpm, bin 3: 7.87 bpm) and the interquartile range (bin 1: 4.15 bpm, bin 2: 3.40 bpm, bin 3: 3.19 bpm) decrease. The rise times of this poem have no trends. Of the half recovery times, the median goes up with higher ratings (bin 1: 5.5 s, bin 2: 5.88 s, bin 3: 6.13 s).

The poem **"Wahre Liebe"** shows only one trend. The median of the rise times increases (bin 1: 9.94 s, bin 2: 10.97 s, bin 3: 18.88 s). The amplitude values and the half recovery times have no indications for trends.

The **Skin Conductance Response**, sorted by rating and poem, can be found in figure 5.2. The first poem, **"Entschuldingungsbrief"**, has an upward trend in the interfence range (bin 1: 0.20 µS, bin 2: 0.26 µS, bin 3: 0.44 µS) of the amplitude values. Its rise times on the other hand indicate no trends. The half recovery times show two trends: An increasing interfence range (bin 1: 2.47 s, bin 2: 4.59 s, bin 3: 5.78 s) and an increasing interquartile range (bin 1: 1.20 s, bin 2: 1.56 s, bin 3: 1.94 s).

"Der Fischer" shows no consistent trends for its amplitude values or the rise times. The half recovery times of "Der Fischer" however, has several trends. The interfence range goes down (bin 1: 4.44 s, bin 2: 2.94 s, bin 3: 2.47 s). Similarly, the interquartile range decreases (bin 1: 2.31 s, bin 2: 1.31 s, bin 3: 1.25 s). Lastly, the number of outliers below the lower fence increases (bin 1: 0, bin 2: 1, bin 3: 2), while their minimum value decreases (bin 1: -, bin 2: 0.25 s, bin 3: 0.16 s).

The amplitudes' median value of **"Die Grenadiere"** increases with higher ratings (bin 1: 0.01 µS, bin 2: 0.02 µS, bin 3: 0.19 µS). In the rise times, no trends are found. The half recovery times have an increasing median (bin 1: 3.23 s, bin 2: 3.33 s, bin 3: 3.72 s) and an increasing maximum value of the outliers (bin 1: 5.81 s, bin 2: 11.75 s, bin 3: 12.47 s).

The last poem of the set, **"Wahre Liebe"**, has only one trend: The median value of the amplitudes goes down (bin 1: 0.03 µS, bin 2: 0.01 µS, bin 3: 0.00 µS). Both rise times and half recovery times show no indication of consistent trends.

Style x Poem

Figure 5.3 visualizes the **Raw Mouse data**, sorted by style and poem.
In the amplitude values of **"Entschuldingungsbrief"**, the interfence range goes up with increasing skill (TTS: 1.605, amateur: 1.613, professional: 2.83). The interfence range of the rise times goes up as well (TTS: 31.46 s, amateur: 37.18 s, professional: 40.82 s). For the half recovery times, the number of outliers above the upper fence goes up (TTS: 0, amateur: 1, professional: 5).

Regarding **"Der Fischer"**, neither the amplitude values nor the rise times show trends. For the half recovery times, the median decreases the more skilled the speaker is (TTS: 8.00 s, amateur: 5.48 s, professional: 4.53 s).

The amplitude values of **"Die Grenadiere"** do not show signs of trends. The rise times have an increasing range of values in the interquartile range as well (TTS: 11.00 s, amateur: 21.65 s, professional: 37.93 s). Additionally, the interfence range goes up (TTS: 21.96 s, amateur: 45.86 s, professional: 85.36 s). The half recovery times have no trends.

Of the poem **"Wahre Liebe"**, the amplitude values do not show signs of trends. The rise times' interquartile range (TTS: 12.74 s, amateur: 16.47 s, professional: 23.96 s) and the interfence range (TTS: 19.68 s, amateur: 37.86 s, professional: 60.73 s) go up as well. Additionally, the median increases (TTS: 11.64 s, amateur: 12.00 s, professional: 12.77 s). In the half recovery times, no trends are found.

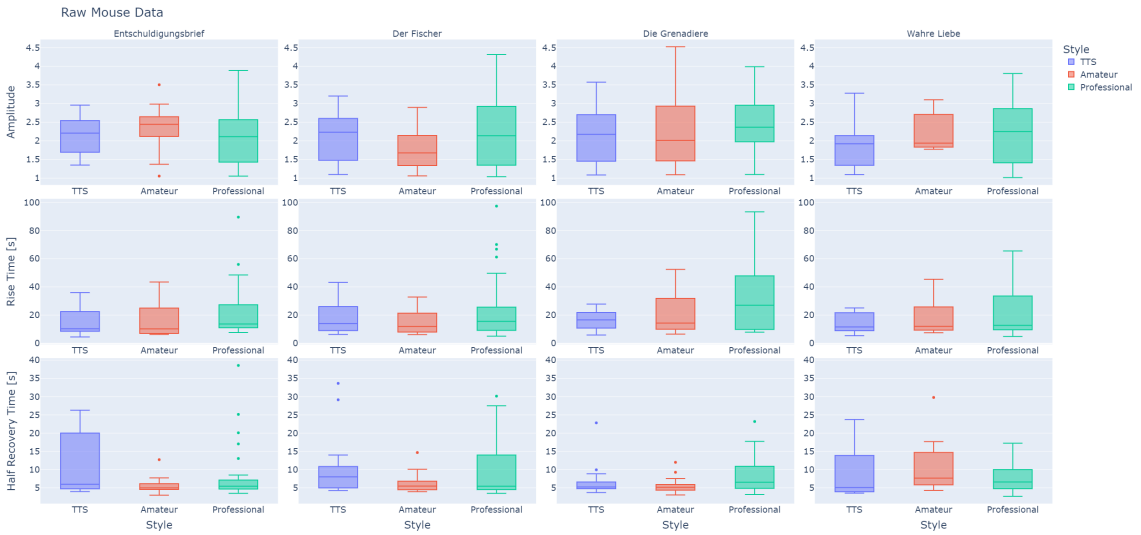


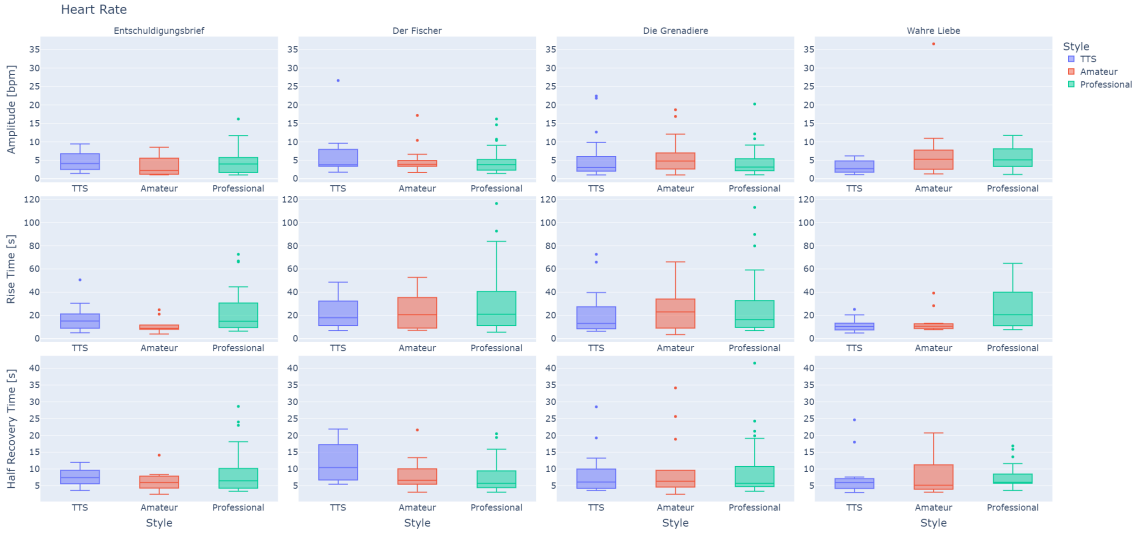
Figure 5.3: Boxplots of the Raw Mouse data, sorted by style and poem. Each column represents a poem. The first row depicts the amplitudes, the second row the rise times in seconds, and the third row the half recovery times in seconds. The blue boxplots are of the Text-To-Speech style (TTS), the red boxplots of the amateur style, and the green boxplots of the professional style. The boxplots visualize the median, the first and third quartile, the lower and upper fences, and outliers. Outliers are data points three times the inter-quartile range lower than the first or higher than the third quartile.

The **Heart Rate**, sorted by style and poem, is visualized in figure 5.4a.

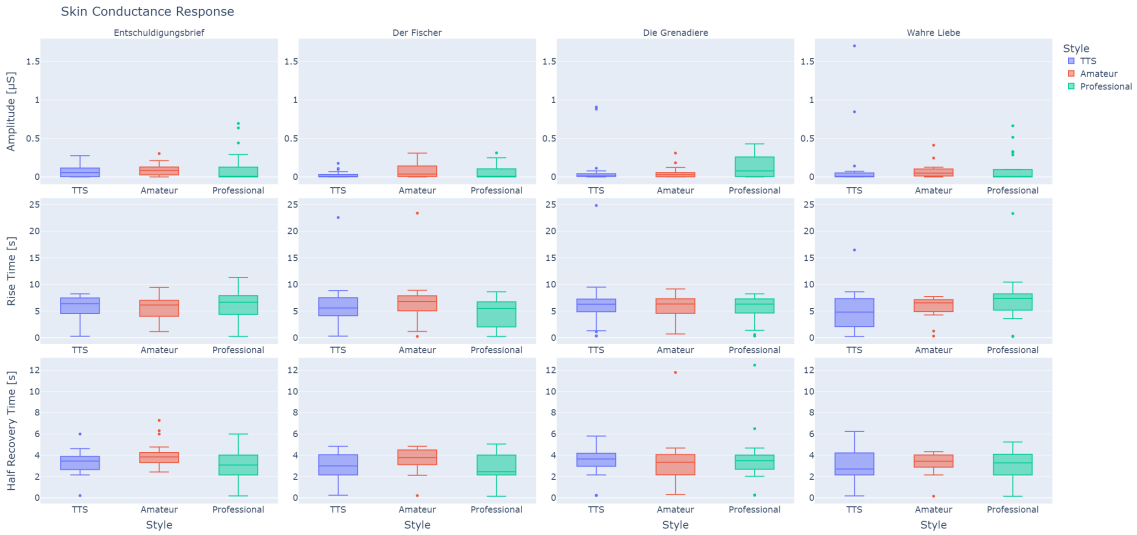
The amplitudes of the poem **"Entschuldingungsbrief"** do not have any trends. The rise times' numbers of outliers above the upper fence increase with increasing skill of the speaker (TTS: 1, amateur: 2, professional: 3). The half recovery times' number of outliers above the upper fence increase as well (TTS: 0, amateur: 1, professional: 3). Furthermore, the outliers' maximum value increases (TTS: -, amateur: 14.13 s, professional: 28.63 s).

Regarding **"Der Fischer"**, the number of outliers above the upper fence in the amplitude data increases (TTS: 1, amateur: 2, professional: 4). The maximum outlier values, however, are decreasing (TTS: 26.62 bpm, amateur: 17.16 bpm, professional: 16.19 bpm). The poem's rise times have an increasing median, as well as increasing values within the fences (TTS: 41.63 s, amateur: 45.63 s, professional: 78.13 s) and an increasing interquartile range (TTS: 21.13 s, amateur: 26.25 s, professional: 34.94 s). The half recovery times' median goes down (TTS: 10.44 s, amateur: 6.63 s, professional: 5.75 s), while the number of outliers above the upper fence goes up (TTS: 0, amateur: 1, professional: 2).

"Die Grenadiere" has no signs of consistent trends in the amplitudes or the rise times. Of the half recovery times, the number of the outliers above the upper fence increases (TTS: 2, amateur: 3, professional: 4), their maximum value increases as well (TTS: 28.5 s, amateur: 34.13 s, professional: 41.5 s).



(a) Boxplots of the Heart Rate, sorted by style and poem.



(b) Boxplots of the Skin Conductance Response, sorted by style and poem.

Figure 5.4: Boxplots of the Heart Rate (a) and the Skin Conductance Response (b), sorted by style and poem. Each column represents a poem. The first row depicts the amplitudes (in (a) in beats per minute, in (b) in microsiemens), the second row the rise times in seconds, and the third row the half recovery times in seconds. The blue boxplots are of the Text-To-Speech style (TTS), the red boxplots of the amateur style, and the green boxplots of the professional style. The boxplots visualize the median, the first and third quartile, the lower and upper fences, and outliers. Outliers are data points three times the inter-quartile range lower than the first or higher than the third quartile.

Of **"Wahre Liebe"**, the amplitude values' interfence range shows a positive trend (TTS: 5.06 bpm, amateur: 9.63 bpm, professional: 10.59 bpm). The rise times' median has an upward trend (TTS: 10.50 s, amateur: 10.69 s, professional: 20.63 s). The half recovery times on the other hand do not show trends at all.

The visualization of the **Skin Conductance Response**, sorted by style and poem, is found in figure 5.4b.
"Entschuldingungsbrief" has an increase in the number of outliers in the amplitude data above the upper fence (TTS: 0, amateur: 1, professional: 3). The outliers' maximum value increases as well (TTS: -, amateur: 0.30 µS, professional: 0.70 µS). The rise times' interfence range goes up (TTS: 7.97 s, amateur: 8.31 s, professional: 11.06 s). The interquartile range increases as well (TTS: 2.91 s, amateur: 2.98 s, professional: 3.5 s). However, there are no trends in the half recovery times.

The poem **"Der Fischer"** has no trends in the amplitude values, the rise times or the half recovery times.

Both the amplitude values and the rise times of **"Die Grenadiere"** do not have any trends. In the half recovery times, the number of outliers above the upper fence (TTS: 0, amateur: 1, professional: 2) and their maximum value (TTS: -, amateur: 11.78 s, professional: 12.47 s) increase.

The amplitudes of **"Wahre Liebe"** have an upward trend in the interquartile range (TTS: 0.05, amateur: 0.09, professional: 0.10). The median of the rise times increases with increasing skill (TTS: 4.83, amateur: 6.55, professional: 7.38). In the half recovery times, no trends are found.

Overall, there are 19 trends present for the poem **"Entschuldigungsbrief"**. The most (5 trends) are in the Heart Rate, sorted by rating. **"Der Fischer"** has 17 trends, most of which are in the Heart Rate, sorted by style (6 trends). For **"Die Grenadiere"**, 16 trends are found, most of them in the Raw Mouse data, sorted by rating (5 trends). Finally, **"Wahre Liebe"** has 13 trends. Most of these are found in the Raw Mouse data, sorted by rating (4 trends). Table 5.5 gives an overview of the results of the trend analysis. The check mark indicates the presence of a trend, the cross indicates the absence of one.

Table 5.5: Overview of the results from the trend analysis. The bold check mark represents the existence of a trend, the grayed out cross represents the lack of a trend. The names of the individual poems are abbreviated (E is "Ensthuldingungsbrief", F is "Der Fischer", G is "Die Grenadiere", W is "Wahre Liebe").

			Trend in Rating				Trend in Style			
Measure	Parameter	Stat. Value	E	F	G	W	E	F	G	W
Raw Mouse	Amplitude	Median	×	×	✓	×	×	×	×	×
		Interquartile	✓	✓	×	✓	×	×	×	×
		Interfence	×	✓	×	✓	✓	×	×	×
		Outliers	×	×	×	×	×	×	×	×
	Rise Time	Median	×	×	×	×	×	×	×	✓
		Interquartile	×	×	✓	✓	×	×	✓	✓
		Interfence	×	✓	✓	✓	✓	×	✓	✓
		Outliers	✓	✓	×	×	×	×	×	×
	Half Rec. Time	Median	×	✓	×	×	×	✓	×	×
		Interquartile	✓	×	✓	×	×	×	×	×
		Interfence	×	×	✓	×	×	×	×	×
		Outliers	×	×	×	×	✓	×	×	×
HR	Amplitude	Median	×	×	✓	×	×	×	×	×
		Interquartile	✓	✓	✓	×	×	×	×	×
		Interfence	×	×	✓	×	×	×	×	✓
		Outliers	×	✓	×	×	×	✓	×	×
	Rise Time	Median	✓	×	×	✓	×	✓	×	✓
		Interquartile	×	×	×	×	×	✓	×	×
		Interfence	×	×	×	×	×	✓	×	×
		Outliers	✓	×	×	×	✓	×	×	×
	Half Rec. Time	Median	✓	×	✓	×	×	✓	×	×
		Interquartile	×	×	×	×	×	×	×	×
		Interfence	×	×	×	×	×	×	×	×
		Outliers	✓	×	×	×	✓	✓	✓	×
SCR	Amplitude	Median	×	×	✓	✓	×	×	×	×
		Interquartile	×	×	×	×	×	×	×	✓
		Interfence	✓	×	×	×	×	×	×	×
		Outliers	×	×	×	×	✓	×	×	×
	Rise Time	Median	×	×	×	×	×	×	×	✓
		Interquartile	×	×	×	×	✓	×	×	×
		Interfence	×	×	×	×	✓	×	×	×
		Outliers	×	×	×	×	×	×	×	×
	Half Rec. Time	Median	×	×	✓	×	×	×	×	×
		Interquartile	✓	✓	×	×	×	×	×	×
		Interfence	✓	✓	×	×	×	×	×	×
		Outliers	×	✓	✓	×	×	×	✓	×

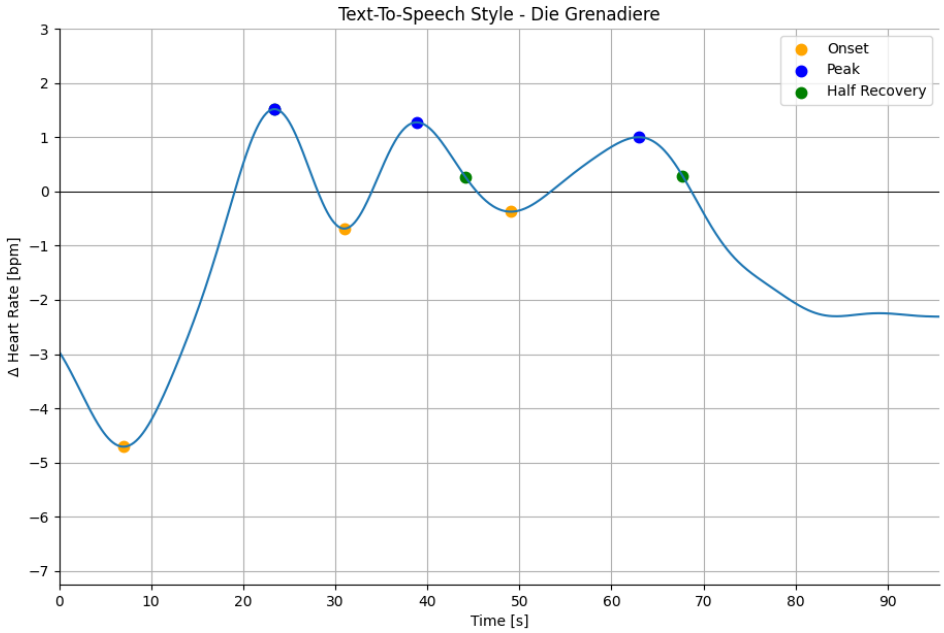


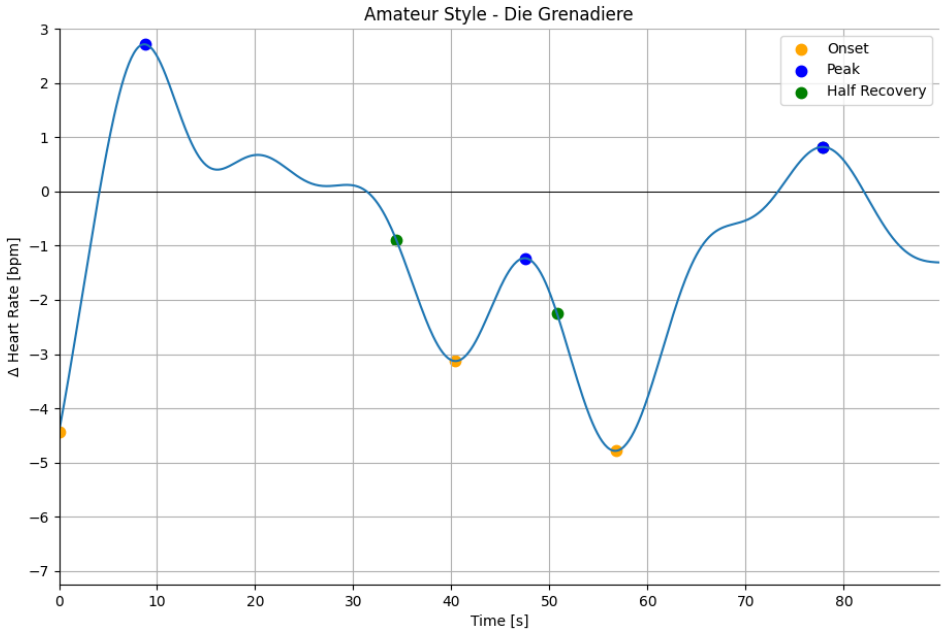
Figure 5.5: Example Heart Rate time-series of “Die Grenadiere” in the Text-To-Speech style. Depicted are differences to baseline mean in beats per minute. The blue line is the signal course. Plotted on top are the onset points (orange), the peaks (blue), and the half recovery points (green).

Example Time-Series

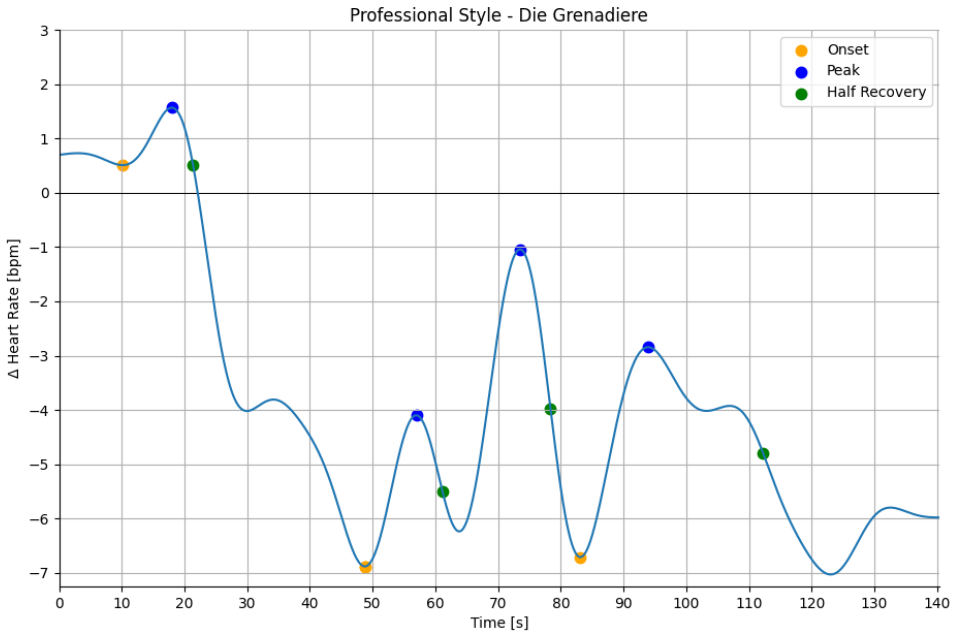
In the following, the Heart Rate time-series of one participant are exemplarily compared for one poem. Figure 5.5, figure 5.6a, and figure 5.6b show the Heart Rate time-series of participant 6 for the poem “Die Grenadiere”. The orange dots indicate the onset, the blue dots the amplitude, and the green dots the half recovery point.

Participant 6 is a female student in the younger half of the participants. The styles were played in the order “Professional - Amateur - Text-To-Speech”. In the professional style, she heard “Die Grenadiere” on third position. She gave it a rating of 7. In the amateur and the Text-To-Speech styles, she listened to the poem on first position. Both received a rating of 3. The example participant stated that she would most likely listen to the professional speaker again. She liked the professional the most, the Text-To-Speech the second most, and the amateur the least. Her preferred style for “Die Grenadiere” was the professional. The professional also transferred the main emotion (trust) the best, according to her. In terms of perceived emotionality and her personal liking, she ranked the poem fourth.

The **Text-To-Speech style** shows three peaks. The mean amplitude is 3.18 bpm, the mean rise time is 12.71 s, and the mean half recovery time is 5.00 s. The **amateur style** has three peaks, a mean amplitude of 4.88 bpm, a mean rise time of 12.34 s, and a mean half recovery time of 14.44 s. The **professional style** shows four peaks. The mean amplitude is 3.39 bpm, the mean rise time is 12.91 s, and the mean half recovery time is 7.69 s.



(a) Example Heart Rate time-series of “Die Grenadiere” in the amateur style.



(b) Example Heart Rate time-series of “Die Grenadiere” in the professional style.

Figure 5.6: Example Heart Rate time-series of “Die Grenadiere” in the amateur style (a) and the professional style (b). Depicted are differences to baseline mean in beats per minute. The blue line is the signal course. Plotted on top are the onset points (orange), the peaks (blue), and the half recovery points (green).

Table 5.6: Overview of the results from the ANOVAs. For parameters with a p-value <0.05, the value is indicated, otherwise replaced by a gray "≥0.05". The names of the individual poems are abbreviated (E is "Ensthuldingungsbrief", F is "Der Fischer", G is "Die Grenadiere", W is "Wahre Liebe").

		p-value							
		Rating				Style			
Measure	Parameter	E	F	G	W	E	F	G	W
Raw Mouse	Amplitude	≥0.05	≥0.05	≥0.05	≥0.05	≥0.05	≥0.05	≥0.05	≥0.05
	Rise Time	≥0.05	≥0.05	≥0.05	≥0.05	≥0.05	≥0.05	0.006	≥0.05
	Half Rec. Time	0.030	≥0.05	≥0.05	≥0.05	0.008	≥0.05	0.022	0.170
HR	Amplitude	≥0.05	0.048	≥0.05	≥0.05	≥0.05	≥0.05	≥0.05	0.042
	Rise Time	≥0.05	≥0.05	≥0.05	0.011	0.032	≥0.05	≥0.05	0.001
	Half Rec. Time	0.042	≥0.05	≥0.05	0.022	≥0.05	≥0.05	≥0.05	≥0.05
SCR	Amplitude	0.072	0.000	0.000	≥0.05	≥0.05	≥0.05	0.000	≥0.05
	Rise Time	≥0.05	0.030	0.035	0.017	≥0.05	≥0.05	≥0.05	≥0.05
	Half Rec. Time	≥0.05	0.006	≥0.05	0.002	≥0.05	≥0.05	≥0.05	≥0.05

Correlation

Following the results from the manual analysis, the results from the subsequent automated analyses are presented.

The one-way ANOVAs show significance ($p < 0.05$) for 21 of the 72 checked parameters. The most significant parameters (8) are present for the Skin Conductance Response. The highest p-value is for the rise time of the Skin Conductance Response for the poem "Entschuldigungsbrief" and the recital style ($p = 0.900$). The lowest p-value is for the amplitude of the Skin Conductance Response of the poem "Die Grenadiere" and the rating ($p = 1.160e^{-11}$). An overview of the significant results can be found in table 5.6.

As mentioned in the previous chapter, Spearman's Rank Correlation Coefficient was calculated for 53 variables against the rating. Table 5.7 shows how many coefficients each poem has in a certain value range. For simplicity, the ranges are denoted as absolute values.

Of the poem **"Entschuldigungsbrief"**, eight variables have a coefficient lower than -0.5 or higher than 0.5. These include the mean frequency of the recital (-0.61), the number of peaks of the Heart Rate (0.65), the minimum amplitude of the Heart Rate (-0.53), the standard deviation of the amplitudes of the Heart Rate (0.56), the minimum rise time of the Heart Rate (-0.56), the standard deviation of the rise times of the Heart Rate (0.55), the minimum half recovery time of the Heart Rate (-0.52), and the standard deviation of the half recovery times of the Heart Rate (0.60).

The poem **"Der Fischer"** has three variables with a coefficient below -0.5 or above 0.5. These include the mean frequency of the recital (-0.79), the number of peaks of the Heart Rate (0.53), and the standard deviation of the rise times of the Heart Rate (0.56).

Regarding **"Die Grenadiere"**, lower than -0.5 or higher than 0.5 is only the coefficient of the mean frequency of the recital (-0.77).

For the poem **"Wahre Liebe"**, a total of 10 variables have a coefficient lower than -0.5 or above 0.5. These 10 are the mean frequency of the recital (-0.79), the maximum rise times of the Heart Rate (0.59), the median rise time of the Heart Rate (0.56), the standard deviation of the half recovery times of the Heart Rate (0.59), the number of peaks of the Skin Conductance Response (0.56), the minimum value of the Skin Conductance

Table 5.7: Number of variables for each poem with a Spearman's Rank Correlation Coefficient in a certain value range. For improved readability, positive and negative values have been combined using absolute values.

	Spearman's Rank Correlation Coefficient				
Poem	$<\|0.2\|$	$\geq\|0.2\| <\|0.3\|$	$\geq\|0.3\| <\|0.4\|$	$\geq\|0.4\| <\|0.5\|$	$\geq\|0.5\|$
Entschuldigungsbrief	30	12	3	0	8
Der Fischer	26	12	10	2	3
Die Grenadiere	25	12	12	3	1
Wahre Liebe	28	7	4	4	10

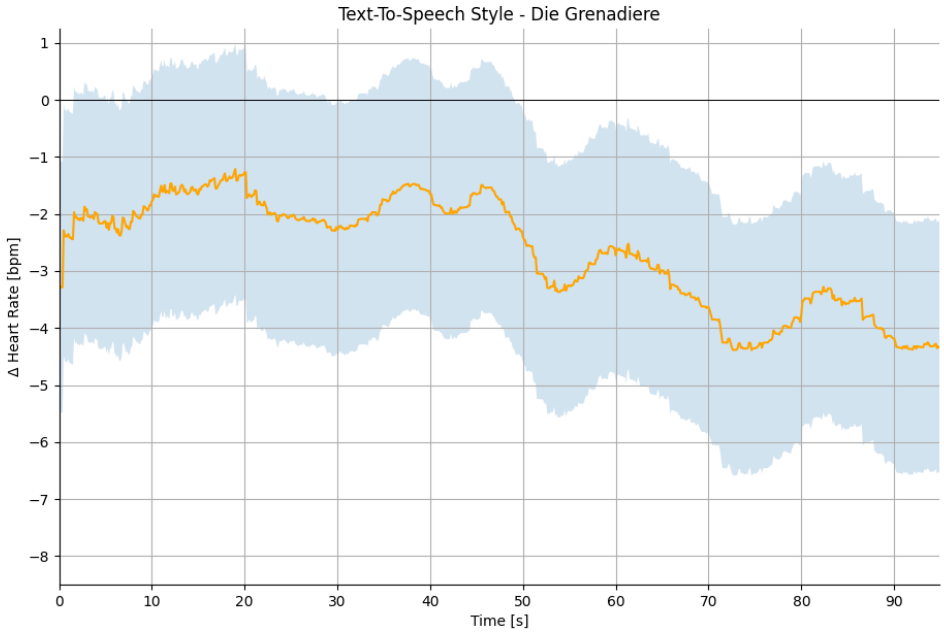


Figure 5.7: 95% Confidence Interval of the Heart Rate for “Die Grenadiere” in the Text-To-Speech style. Depicted are differences to baseline mean in beats per minute. The orange line is the mean signal course over all participants. The Confidence Interval is visualized as the area in blue.

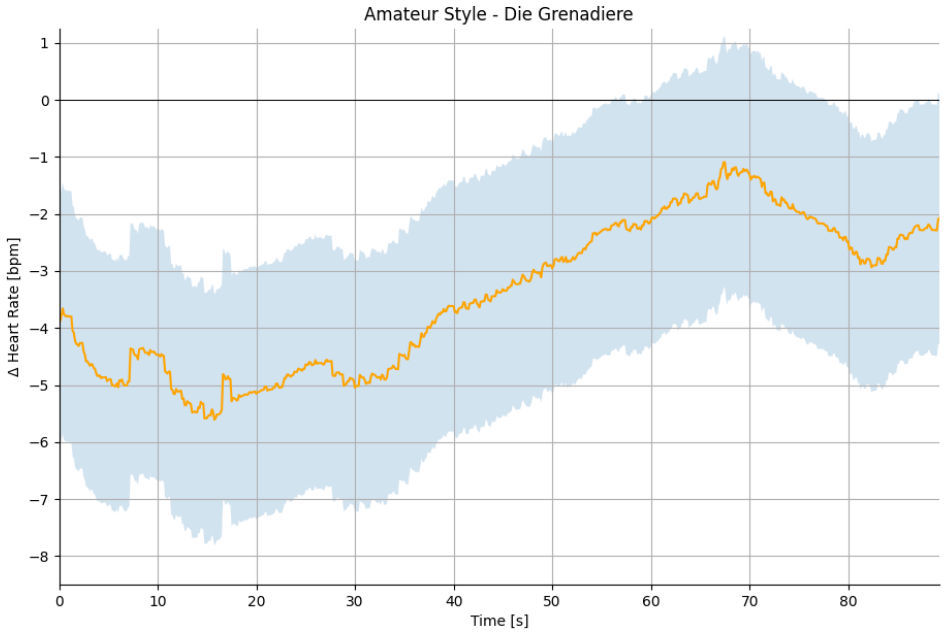
Response (-0.52), the standard deviation of the amplitudes of the Skin Conductance Response (0.80), the maximum rise time of the Skin Conductance Response (0.62), the standard deviation of the rise times of the Skin Conductance Response (0.78), and the standard deviation of the half recovery times of the Skin Conductance Response (0.71).

Overall, 16 distinct variables have a coefficient lower than -0.5 or above 0.5. Four times, these are the mean frequency of a recital. Twice each, it is the number of peaks of the Heart Rate, the standard deviation of the rise times of the Heart Rate, and the standard deviation of the half recovery times of the Heart Rate. The rest of the variables have a correlation lower than -0.5 or above 0.5 only once. The Raw Mouse data never has such a relevant coefficient.

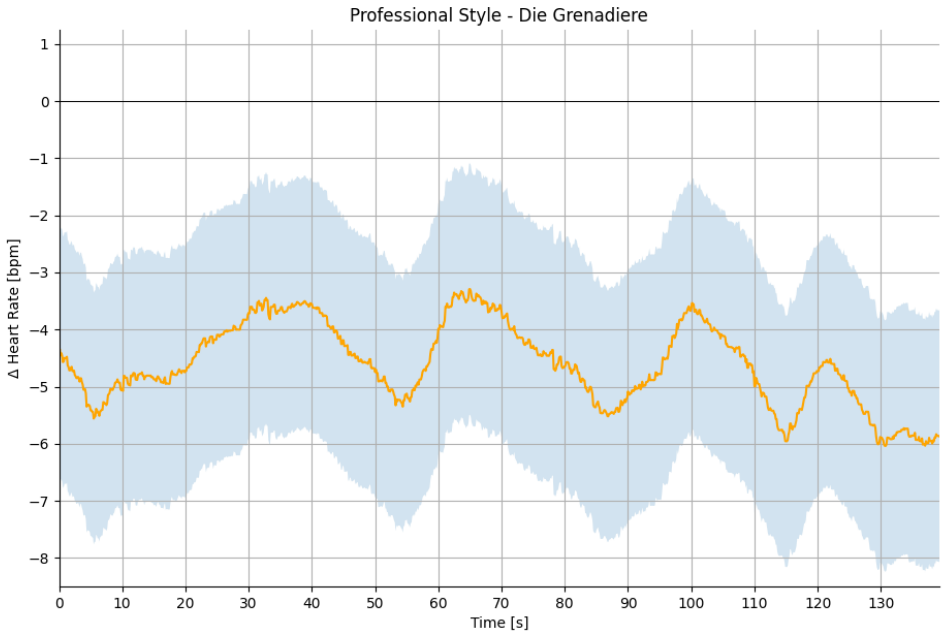
Confidence Interval

The 95% Confidence Interval visualizes the range in which 95% of the measured values of all participants lie. In the presented graphs (figure 5.7, figure 5.8a, and figure 5.8b), this is represented by the blue areas. The orange lines visualize the arithmetically averaged values of all participants.

The 95% Confidence Intervals were calculated for all conditions and physiological measures. The presented graphs show the intervals for the Heart Rate during the recitals of “Die Grenadiere”. Figure 5.7 visualizes the interval for the Text-To-Speech style. Over the signal course, the values have a downward trend. For the amateur



(a) 95% Confidence Interval of the Heart Rate for “Die Grenadiere” in the amateur style.



(b) 95% Confidence Interval of the Heart Rate for “Die Grenadiere” in the professional style.

Figure 5.8: 95% Confidence Interval of the Heart Rate for “Die Grenadiere” in the amateur style (a) and the professional style (b). Depicted are differences to baseline mean in beats per minute. The orange line is the mean signal course over all participants. The Confidence Interval is visualized as the area in blue.

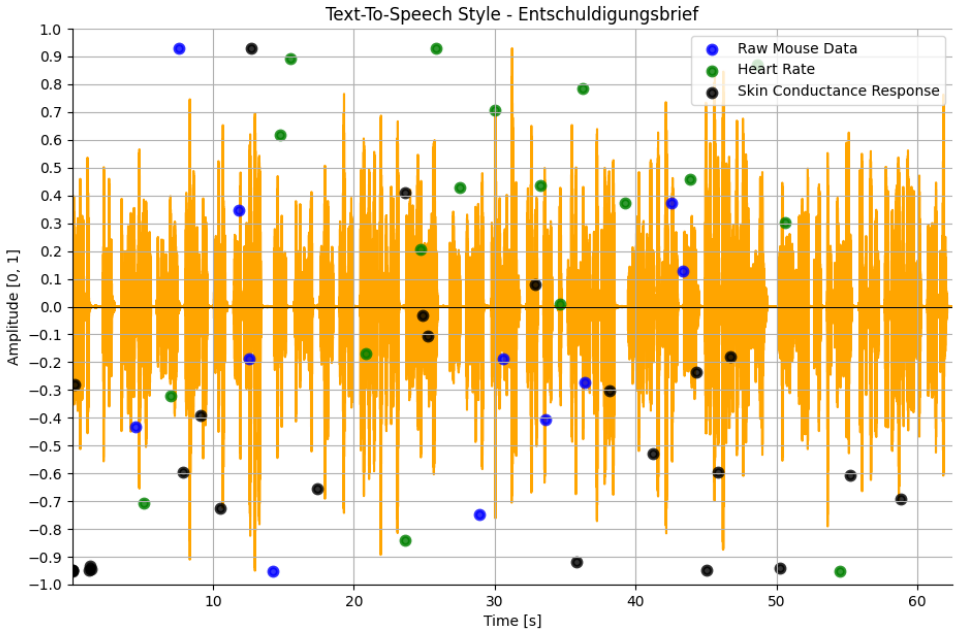


Figure 5.9: Heatmap of the peaks for and waveform of “Entschuldigungsbrief” in the Text-To-Speech style. The audio amplitude of the recital is depicted in orange in the background. It visualizes the sound pressure on an arbitrary scale between 0 and 1. The left audio channel is represented by positive values, the right audio channel by negative values. The peaks from the Raw Mouse data (blue), the Heart Rate (green), and the Skin Conductance Response (black) are plotted on top. They are scaled to be within the maximum audio amplitude.

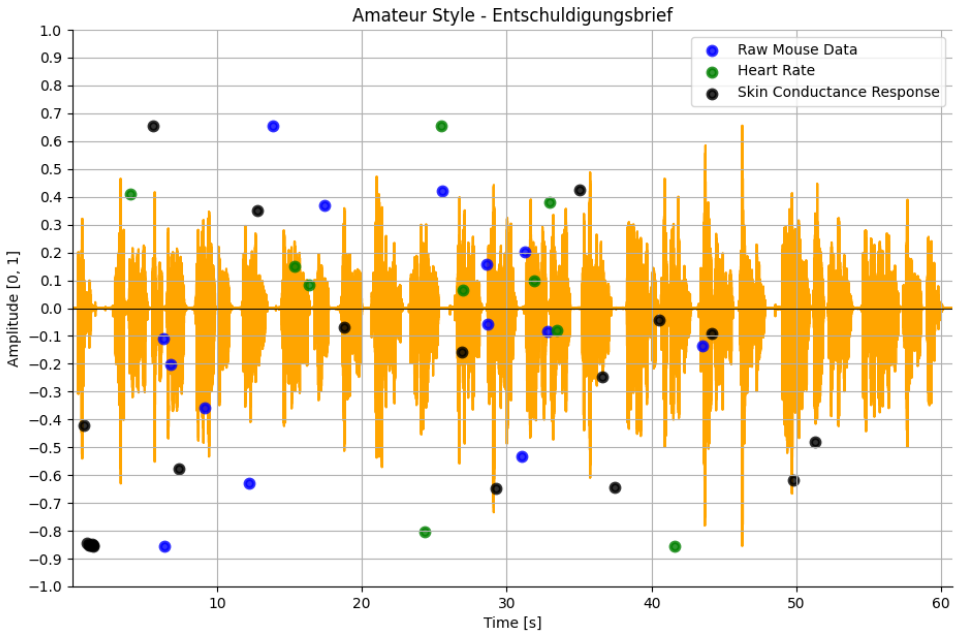
style (cf. figure 5.8a), a large peak towards the end can be seen. The values have an upward trend. The professional style (cf. figure 5.8b) has four distinct peaks. The signal course appears more variable than in the other two styles.

Heatmap

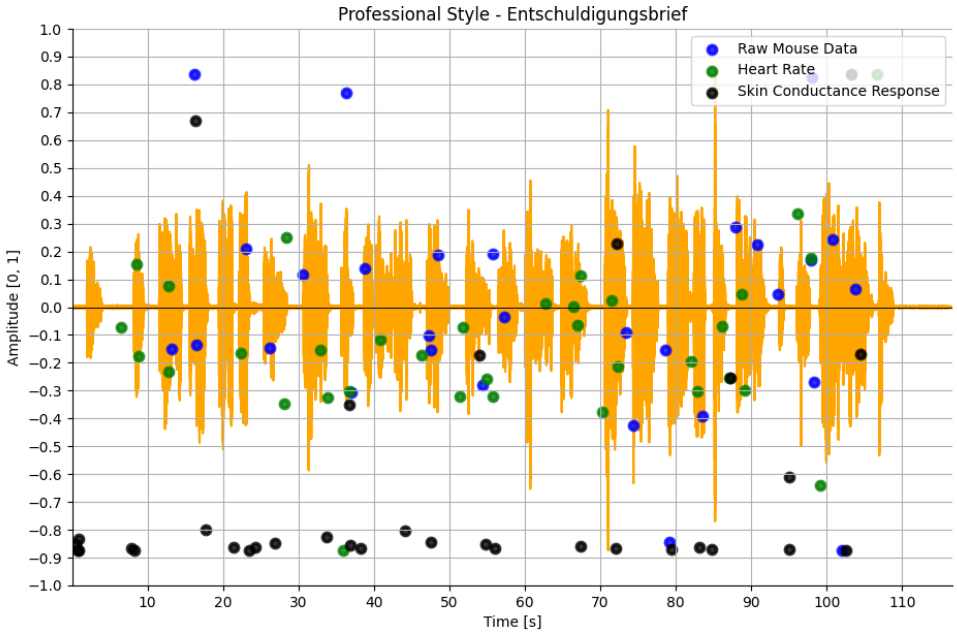
The heatmaps visualize the peaks of the three measures of all participants on the belonging audio amplitude. The amplitude, i.e. the sound pressure, is visualized in orange in the background. The peaks of the Raw Mouse data are marked in blue, of the Heart Rate in green, and of the Skin Conductance Response in black. The vertical positions of the peaks are scaled to lie between the minimum and maximum amplitude of the audio. Similar to the Confidence Intervals, the heatmaps are presented exemplarily for the poem “Entschuldigungsbrief” (cf. figure 5.9, figure 5.10a, and figure 5.10b). Fewer peaks occur towards the end of the recitals and most peaks occur when the speakers are audible, i.e. not during dramaturgical breaks.

5.4 Miscellaneous

The probabilities of the poems being played in a certain position of a style are shown in table 5.8. Assuming a perfect distribution, the expected probability would be 0.25. “Entschuldigungsbrief” is exceptionally rare in the first position. The same is true for



(a) Heatmap of the peaks for and waveform of “Entschuldigungsbrief” in the amateur style.



(b) Heatmap of the peaks for and waveform of “Entschuldigungsbrief” in the professional style.

Figure 5.10: Heatmap of the peaks for and waveform of “Entschuldigungsbrief” in the amateur style (a) and the professional style (b). The audio amplitude of the recital is depicted in orange in the background. It visualizes the sound pressure on an arbitrary scale between 0 and 1. The left audio channel is represented by positive values, the right audio channel by negative values. The peaks from the Raw Mouse data (blue), the Heart Rate (green), and the Skin Conductance Response (black) are plotted on top. They are scaled to be within the maximum audio amplitude.

"Die Grenadiere" in the second position. However, in the second position "Entschuldigungsbrief" occurs unexpectedly frequent, as well as "Die Grenadiere" in first position. Three participants listened to the same order of poems twice, for one participant in two adjacent styles.

In the **Text-To-Speech style**, three poem orders appeared twice, while the poem "Wahre Liebe" never occurred first. In the **amateur style**, the sequence "Wahre Liebe - Entschuldigungsbrief - Der Fischer - Die Grenadiere" appeared four times, another order appeared two times. "Entschuldigungsbrief" never occurred in the first position. In the **professional style**, the order "Die Grenadiere - Entschuldigungsbrief - Wahre Liebe - Der Fischer" appeared four times. The poem "Entschuldigungsbrief" never appeared in the first position, but nine times in the second position. "Die Grenadiere" was never listened to in the second position.

Of all ratings, 56 times (38.89%) a rating in **bin 1** was made. Of these, 36 (64.29%) were for the Text-To-Speech style (75.00% of all TTS ratings), 20 (35.71%) for the amateur style (41.67% of all amateur recitals), and zero (00.00%) for the professional style (00.00% of all professional ratings). The mean length of the recitals in this bin is 71 s, they have a mean frequency of 215 Hz.

A rating in **bin 2** was received 51 times (35.41%). To the Text-To-Speech style belong 11 (21.57%; 22.92% of all TTS ratings) of these, 26 (50.98%) belong to the amateur style (54.17% of all amateur ratings), and 14 (27.45%) to the professional style (29.17% of all professional ratings). They have a mean length of 89 s. Their mean frequency is 165 Hz.

37 (25.69%) of the ratings were in **bin 3**. Only one (02.70%) is part of the Text-To-Speech style (02.08% of all TTS ratings), two (05.41%) are part of the amateur style (04.16% of all amateur ratings), and 34 (91.89%) are part of the professional style (70.83% of all professional ratings). The mean length of the recitals with a rating in bin 3 is 121 s, the mean frequency 132 Hz.

The recitals in the **Text-To-Speech style** have an average length of 72 s and a mean frequency of 250 Hz. 36 times (75.00%), these recitals got a rating in bin 1, 11 times (22.92%) in bin 2, and one time (02.08%) they got a rating in bin 3.

In the **amateur style**, the recitals have an average length of 71 s and a mean frequency of 152 Hz. 20 times (41.67%), they received a rating in bin 1, 26 times (54.17%) a rating in bin 2, and two times (04.17%) a rating in bin 3.

The four recitals of the **professional style** have an average length of 128s. Their mean frequency is 126 Hz. Zero times (00.00%) they got a rating in bin 1, 14 times (29.17%) a rating in bin 2, and 34 (70.83%) times they received a rating in bin 3.

Table 5.8: Probabilities of the poems appearing a certain position during playback in each recital style. With a perfect distribution, the positions each would have a probability of 0.25.

Poem	Position			
	First	Second	Third	Fourth
Entschuldigungsbrief	0.08	0.50	0.19	0.22
Der Fischer	0.22	0.17	0.28	0.33
Die Grenadiere	0.42	0.08	0.25	0.25
Wahre Liebe	0.28	0.25	0.28	0.19

The mean rating of the male participants is 5.17, the mean rating of the female participants is 5.57. Of the six younger participants (20 - 24), the mean rating is 5.02. Of the six older participants, the mean rating (26 - 32) is 5.72.

The most liked poem (“Der Fischer”) has an EmoScore of 0.06. “Entschuldingungsbrief”, the second most liked poem, has an EmoScore of 0.05. The third most liked poem (“Die Grenadiere”) has an EmoScore of 0.00. “Wahre Liebe”, the least liked poem, has an EmoScore of 0.08.

6 Machine Learning

Considering that the results in chapter 5 appear promising, in the next step of the project, two sets of Machine Learning models were trained using the data aggregated in the data analysis. One set for predicting a recital's rating and one for predicting a recital's style. Both models are based on the corresponding physiological measures. The goal of training the models was to answer the questions "Can you predict a participant's rating based on their physiological response?" and "Can you predict the style (TTS, amateur or professional) based on the physiological response?". The models can add to the results of the trend and correlation analysis and thereby help to support the hypothesis.

In chapter 4, it was explained why the data could only be analyzed on an intra-poem level. This limitation also applies to the Machine Learning tasks. Therefore, all models were built on a per poem basis.
The data for each poem included one data point for each amplitude, rise time, and half recovery time value of the Raw Mouse data, Heart Rate, and Skin Conductance Response. Any invalid values (NaN) were dropped. Besides the value, each data point included the type of measure, the parameter, i.e. amplitude, rise time or half recovery time, and the number of peaks in the value's time-series. To be able to be processed, the first two additional values were transformed to categorical integers.

In the analysis, the ratings were binned into three bins. This binning was kept to increase the amount of samples per class in the rating task. This made the models easier to train, since data sparseness is a common issue in Machine Learning tasks. As in the trend analysis, exact results were less important than their order of magnitude. Even after the binning, the data sets are sparse with only between 461 and 659 data points per poem. At least Deep Learning requires generally a lot more samples for meaningful and generalizable results. Therefore, only statistical models were examined in this first trial. The evaluation should be able to indicate the feasibility of training a Machine Learning model for the tasks at hand. To counter the data sparseness further, 10-fold cross-validation was used and the order of the data points shuffled. All evaluation metrics were averaged over all 10 folds.

Predicting a poem's rating can be considered a regression task. The ratings are placed on an ordinal scale, even though only integer values were accepted during the experiment. Additionally to the regression algorithms, classification algorithms were tested. To evaluate the models, the Mean Absolute Error (MAE) was calculated. A baseline model predicted a training set's median rating for all test samples.
For the prediction of the styles, only classification algorithms made sense as the style classes were strictly nominal. For the evaluation of the classification models, the F1-score, i.e. the harmonic mean of Precision and Recall, was calculated. A model predicting the classes according to their distribution in the training set served as the baseline.

In this trial, 14 different statistical Machine Learning algorithms were tested, most of them both as classifiers and regressors. These included, amongst others, Random Forest, Decision Tree, Multilayer Perceptron, Logistic Regression, Stochastic Gradient Descent, Support Vector Machine, and Naïve Bayes. The models were implemented using scikit-learn (Pedregosa et al., 2011). They were all initialized with their standard parameters. For reproducibility, their random state was each set to 97. Regarding the rating prediction task, the three algorithms with the lowest MAE were selected per poem. For the style prediction task, the same was done with the highest F1-score.

The parameters of the top three models were then tuned, using scikit-learn's Randomized Search. This function randomly tests parameters in a predefined range and selects the best fitting set of parameters. The search was performed for up to 500 iterations. Afterwards, a model was refit with the best parameters and evaluated. The parameters' ranges were set based on preceding manual tests. As the Randomized Search adds another layer of cross-validation (3-fold), the best parameter settings differed between each fold and a single set of the best parameters for the entire data set cannot be presented.

6.1 Results

In the following, the best fitting models and the respective baselines are presented. The results of the models predicting the rating can be found in table 6.1, the results of the models predicting the style are listed in table 6.2.

Concerning the prediction of the rating, a Support Vector Classifier based model works best (MAE: 0.52) for the poem **"Entschuldigungsbrief"**. The MAE of the model is 0.11 (21.15%) lower than the baseline's (MAE: 0.63). For **"Der Fischer"**, an Extra Trees Classifier model (MAE: 0.48) performs the best. Its MAE is 0.21 (43.75%) lower than the baseline's (MAE: 0.69). **"Die Grenadiere"** has the best performance with a Support Vector Classifier model (MAE: 0.54) as well. The MAE is 0.07 (12.96%) lower than the baseline's (MAE: 0.61). For **"Wahre Liebe"**, a Random Forest Classifier model works best (MAE: 0.50). The MAE of the model is 0.20 (40.00%) lower than the baseline's (MAE: 0.70).

Table 6.1: Results of the rating prediction Machine Learning task. Shown are the results of the baseline and the best-performing algorithm for each poem. The models were evaluated with the metric Mean Absolute Error (MAE).

Poem	Algorithm	MAE
Entschuldigungsbrief	Baseline	0.63
	Support Vector Classifier	0.52
Der Fischer	Baseline	0.69
	Extra Trees Classifier	0.48
Die Grenadiere	Baseline	0.61
	Support Vector Classifier	0.54
Wahre Liebe	Baseline	0.70
	Random Forest Classifier	0.50

All four models beat their baselines. On average, the trained models are 0.1475 (29.47%) better than the baseline. The average Mean Absolute Error is 0.51. The best algorithms are always classifiers. Bin 2 is the overall most predicted rating (38.68%).

Regarding the style prediction task, the poem **"Entschuldigungsbrief"** has the best results with a Support Vector Classifier model (F1-score: 0.5455). It performs 0.2280 points (71.50%) better than the baseline (F1-score: 0.3175). The best results on **"Der Fischer"** are achieved with an Extra Trees Classifier model (F1-score: 0.5595). Its F1-score is 0.2465 points (78.75%) higher than the baseline's (F1-score: 0.3130). For **"Die Grenadiere"**, the highest F1-score is achieved with a Random Forest Classifier model (F1-score: 0.5034). It performs 0.1702 points (51.08%) better than the baseline (F1-score: 0.3332). For **"Wahre Liebe"**, an Extra Trees Classifier model (F1-score: 0.5556) works best. Its F1-score is 0.2377 points (74.77%) better than the baseline's (F1-score: 0.3179).
The baselines get beaten by all of the models. The F1-scores are, on average, 0.2206 points (69.03%) higher than the baselines'. The average F1-score is 0.5408. Overall, the professional style is the most predicted style (48.56%).

Table 6.2: Results of the style prediction Machine Learning task. Shown are the results of the baseline and the best-performing algorithm for each poem. The models were evaluated with the metric F1-score.

Poem	Algorithm	F1-score
Entschuldigungsbrief	Baseline	0.3175
	Support Vector Classifier	0.5445
Der Fischer	Baseline	0.3130
	Extra Trees Classifier	0.5595
Die Grenadiere	Baseline	0.3332
	Random Forest Classifier	0.5034
Wahre Liebe	Baseline	0.3179
	Extra Trees Classifier	0.5556

7 Discussion

To answer the research questions and to try to support the hypothesis, the results are interpreted below. They are discussed in the same order as they were presented in the previous chapter. Again, the results can only be compared on an intra-poem level (cf. chapter 4).

7.1 Demographics

With 12 participants, the size of the study was rather small. The study is not representative and the results are not generalizable. As intended, it only serves for first impressions and the discovery of trends.

The participants' gender was perfectly balanced between male and female. For each condition, one sample was from a female and one sample from a male participant. People of other genders did not participate. A participant's gender is therefore not needed to be considered as a confounding variable. The cohort was generally of younger age. Hence, the results would potentially be different for older people. All of the participants were members of academia, they received scientific training and can be considered as well educated. They might have a better understanding of the poems than less educated people. Furthermore, they might appreciate poetry differently. This might affect the participants' ratings.
The demographics of the participants do not allow for a generalization on a wider population. However, the homogeneity of the group favors the discovery of trends. The participants potentially reacted more similar than a more heterogeneous group would, which makes the discovery of trends more easy.

7.2 Questionnaire

Ratings

The possibility of reevaluating the given ratings was used by seven participants. They only applied small changes to the ratings, the biggest change was by 3 points.
The initial assessments directly after a stimulus seem to reflect the opinions quite well. When repeating the experiment, the possibility for adjustments to the ratings should nevertheless not be removed. It gives the participants the security that the initial assessment does not have to be perfect and can still be changed later. The focus is thus not too much on the correct quantification of the opinion and the participants can fully focus on experiencing the recitals.

The increase of the Average Observed Agreement for the binned ratings (cf. section 5.3) supports the grouping of the ratings. However, the agreement is still rather low. This highlights the high inter-individual differences in the experience of poetry

recitals and is an argument against objective ratings. As the participants did not have a common rating scheme, they might use different ratings for a similar perception, thus leading to the inter-individual differences.
This also raises the question of how consistently participants rate the recitals. It is unclear whether they would rate similarly in a repetition. The binning into coarser ratings can counteract such intra-personal inconsistencies. It also ensured that enough data per rating was available for the trend analysis.

As expected, the Text-To-Speech style received the lowest mean ratings (2.83 - 3.33) and the professional style the highest, i.e. best, mean ratings (8.33 - 7.42). Additionally, the mean ratings allow for a clear distinction between the styles. This supports the considerations made during the planning of the study (cf. chapter 4), i.e. to assume an increase in quality from the Text-To-Speech style over the amateur style to the professional style. This is important for the validity of the design and the reproducibility of the study. The few outliers in the ratings show that the ratings are subjective assessments.

By mean rating, “Entschuldigungsbrief” was rated the highest in the Text-To-Speech style and the amateur style but the lowest in the professional style. In the professional style, “Wahre Liebe” was rated the highest but was rated the lowest in the Text-To-Speech style and the amateur style. Therefore, the recitals appear to have been rated independently of their content, as requested in the Info Sheet (cf. Appendix). This increases the validity of the results and limits the content of the poems as a confounding variable. It also shows that the Text-To-Speech style and the amateur style were perceived more similar than both to the professional style. This is also reflected by the fact that the mean ratings in the Text-To-Speech and amateur styles are closer to each other than both to the professional style.

Descriptions

The descriptions show how the different recital styles were perceived. With the help of the positive and negative judgments, it can be determined which characteristics make a good recital.

The **Text-To-Speech style** appeared very mechanical and monotonous to the participants. The participants could easily recognize that it was a computer-generated voice. In the descriptions, the participants mentioned some factors that made the voice sound unnatural. For example, the rhythm, intonation, and pronunciation. This led to the voice being perceived as absurd and funny, and not taken seriously.
The fewest positive things were said about this style. Four participants gave no answer to the question what they liked about the Text-To-Speech speaker or explicitly stated that they did not like anything. Aside from that, the clear pronunciation and intonation were mentioned. However, both of these could also simply be euphemistic and might have been stated for the sake of answering something. Both the pronunciation and intonation were also heavily criticized by other participants. That participants liked that the style sounds funny and ironic or soulless, supports this assumption.
When asked, the participants criticized the mispronunciation and inconsistencies in the intonation. In the future, when adapting Text-To-Speech to poetry, the voice should be made more natural and less distorted in addition to these points. Variations in speech

melody and rate of speech could help improve the recitals. After implementing the improvements, the new Text-To-Speech system should also be compared with the other two styles (cf. chapter 9).

The descriptions show that the Text-To-Speech style was received poorly by most participants and is only liked ironically. The ratings, with the exception of a few outliers, support this view. Before a commercial use for poetry recitals, the voice needs to be heavily improved.

The **amateur style** was received in a mixed way. The participants correctly observed that the speaker has not received any professional training as a speaker. In their descriptions, the participants already criticized a few points. These seem to have stood out to them particularly. The lack of enthusiasm, the length of dramaturgical pauses, and the monotony in the voice were criticized. The intonation on the other hand was praised. In the criticism, it should not be forgotten that this was an amateur.

When the participants listed what they liked, only two indicated that they did not like anything. One of the two people did not like anything about the Text-To-Speech style either. In addition to fluency, pronunciation, and intonation, the participants stated that the speaker sounded approachable and sympathetic. Both of these are qualities that were not mentioned for the Text-To-Speech style. The fluent performance also stands out from the Text-To-Speech style, which was criticized for its irregular way of speaking.

When looking at what the participants disliked about the amateur style, it becomes apparent, that poetry recitals are a subjective matter. As some people liked the pronunciation and intonation of the speaker, some disliked it. The voice was also called monotonous and too high pitched for the poems. Similar to the Text-To-Speech style, the rate of speech was criticized.

The amateur style was perceived considerably better than the Text-To-Speech style, which is also reflected in the ratings. Some points of criticism from the Text-To-Speech style were positively remarked upon here. It stands to reason that an amateur speaker has a gut feeling for the recital of poems. However, it also became clear that there is still room for improvement.

When describing the **professional style**, mainly positive aspects were mentioned. For example, the voice was described as melodic, pleasant, and emotional. In general, the descriptions pointed to a professionally trained speaker.

The answers to what the participants liked about the professional speaker implied that the speaker must be a professional as well. For example, the acting component of the recitals was emphasized. In general, pronunciation, and intonation were mentioned again. The low pitch of the voice also seems to have been a factor. In contrast to the other two styles, the speaker was able to use dramaturgical pauses and the rate of speech effectively. Overall, a more broad range of characteristics was perceived as positive than in the other two styles.

Some points that the participants did not like about the professional style are surprising. Besides understandable points, such as the speaker being too slow or too low and some finding the acting exaggerated, the speaker was also perceived as too good. The sentiment regarding the depth of the voice is very subjective, as some liked it and some did not.

It is clear from the descriptions and the ratings that the professional style recitals were perceived best by the participants. It can be concluded that the other two styles should follow this style to improve their quality of performance.

The comparison of the styles again highlights the characteristics that make a good recital. Furthermore, it was found that the skill of the speakers could easily be identified by the participants and was attributed according to the author's expectations. This indicates a successful study design. The results show that pronunciation and intonation, dramaturgical pauses, the rate of speech, and pitch and timbre are important qualities of reciters. It is also important that the reciter speaks in a melodic way. These qualities lead to different perceptions of emotionality, a key point of the reasoning behind the hypothesis.

The findings match the results of Rebordão et al. (2009) and Schröder (2001). Both suggest pitch or fundamental frequency, and the rate of speech as emotional features of speech. Schröder (2001) additionally name the loudness. In comparison, none of the participants of the present study mentioned the loudness. However, this was to be expected, as the loudness was normalized in the present study.

Preferences

The unanimous result that the participants would most likely listen to the professional speaker is in accordance with the mean ratings. This supports the overall tendency in the ratings. The outliers, i.e. particularly high ratings of the Text-To-Speech style, are also in line with this result, all participants gave the majority of their highest ratings to the professional style.

The results from the ranking of the most favorite to least favorite speaker is again in accordance with the previous results, as the most favorite speaker is the professional one. However, the ranking is not unanimous. Participants might have chosen the Text-To-Speech speaker as their most favorite because it sounds "funny" or "comical", but indicated that they would most likely listen to the professional speaker again, because they actually prefer listening to the professional. As before, the majority of the highest ratings of each participant was given to the professional style, supporting this assumption.

For every poem, the best fitting speaker is always the professional. "Entschuldigungsbrief" had the lowest agreement. This fits that "Entschuldigungsbrief" was rated the highest in the Text-To-Speech style and the amateur style but not the professional style. For this poem, the Text-To-Speech and amateur styles seem to perform comparatively well. Overall, the results are as expected from the earlier findings.

Emotions

Since four participants, contrary to the wording of the question, answered with emotions that are not part of the eight basic emotions by Plutchik, the question design should be adapted in case of a repeat study. Instead of an open answer, checkboxes could be used. Nevertheless, all emotions could be transferred into basic emotions, so that the analysis was possible without problems.

The valence that was automatically determined in the preparation matches the valence of the participants' evaluations of all four poems. However, only in one case, the most associated emotions correspond to the in section 4.2 automatically determined ones. This shows that although the valence is quite clear, the actual associated emotion cannot be represented by the sum of the emotions of the individual words.

"Die Grenadiere" was only marginally attributed negative valence (seven times negative vs. five times positive). Also, a negative emotion (sadness) is ranked together with a positive emotion (trust) as the most associated emotion. The perception appears to be very subjective. For "Der Fischer", the attribution of the valence agrees with the automated assessment but not with the author's one (cf. chapter 4). The attribution is also more unambiguous than for the previous poem (nine times positive vs. three times negative). Since this is a personal assessment, the subjective component comes into play here. The intended balancing of positive and negative valenced poems was not achieved. The trend analysis, as well as the other results are still valid, but reasoning based on valence is not possible.

The majority of participants (9) chose the professional as the best in expressing emotions. Few individual participants chose the amateur (3) and none the Text-To-Speech (0). This result was expected. On the one hand, the professional did best in the rating of the recitals, on the other hand, the professional has proper theatrical training and experience in voice acting. The amateur speaker is human and has an intuition for conveying emotions. The Text-To-Speech synthesizer is agnostic of emotions.

Opposed to the fact that the emotion association for "Der Fischer" has the lowest agreement is the fact that the recital of "Der Fisher" has the second best rating in the professional style and only the third best rating in the Text-To-Speech and amateur styles. The three participants who went against the majority, however, rated the professional recitals always higher than their best speaker in expressing emotions (2, 6, and 4 rating points higher). Either the emotionality was not important to them in the rating or they wanted to make the answers more diverse and it can be attributed to demand characteristics. However, this anomaly is normalized by the majority.

Regarding the features that were used to convey the emotions, all results were mentioned above in the context of the positive characteristics of the professional speaker. The positive characteristics therefore seem to be associated with emotionality. As with the positive characteristics, intonation was mentioned most frequently, hence it appears to be the most important feature for emotionality from the perspective of the participants.

Content

The participants chose "Die Grenadiere" as the most emotional poem by content. However, it ranked third in the ranking of the most liked poem by content. The most liked poem by content is "Der Fischer". For all other places in the ranking, the participants had a higher agreement.

A correlation between the emotionality of the content and the liking of the content is not apparent. Nor does there seem to be a correlation with the ratings of the recitals. They seem to have been perceived independently of the text.

Bodily Changes

As explained in chapter 4, the first question on bodily changes aimed to find out whether participants perceived changes themselves, i.e. they noticed very strong physiological reactions, even though they did not know about the hypothesis. Noticeable reactions would support the hypothesis. Additionally, the question concerned whether the recorded data had to be cleaned of confounding factors, e.g. distractions.

All but three participants noticed changes. The fact that one participant was stressed when listening to the recitals is likely explained by the fact that the participants were in an unusual and unfamiliar situation. They also might have felt pressure to give a good performance. As the participant felt the stress during all recitals, potential effects on the physiological measures should have normalized out. The fact that three participants became more relaxed over time and four participants were bored by the amateur and Text-To-Speech recitals suggests that there must be differences in the physiological reactions, at least between the Text-To-Speech and the amateur style, and the professional style. It also supports the assumption that the professional recitals are of better quality. Boredom should not occur with good recitals.

The participants of the present study did not report any chills. However, based on the results of Wassiliwizky et al. (2017) such reports were expected. In their publication, they report that "[...] nearly 77% of the naïve participants experienced chills in response to unfamiliar poems [...]" (Wassiliwizky et al., 2017, p. 1237). The difference may be due to the fact that Wassiliwizky et al. (2017) explicitly asked about chills and the participants paid more attention to them. It is possible that similar results could have been observed if the questionnaire had asked about chills. This could be investigated in future work.

The attention was inquired about because it could have had an affect on the quality of the data. More participants noticed changes in attention than did in physical changes. The two types of changes are not exclusively mutually dependent. Only one participant did not notice any changes, this participant also did not notice any physical changes.

Eight of the participants noticed differences in attention between the styles, i.e. the lowest attention during the Text-To-Speech and the amateur recitals, and the highest during the professional recitals. This finding is consistent with the ones on the bodily changes and further supports that the professional recitals are of higher quality.

That one participant was distracted by outside events shows that the study was not conducted in a perfect lab environment. As the distraction was only caused by chirping birds and the slight movement of the experimenter, the data of the participant did not need to be discarded.

The question on the use of specific concentration techniques was answered with No by the majority of the participants. This might indicate that the question was not understood correctly. Answers like visualizing the recitals or focusing on certain properties were expected. It is also possible that the participants indeed did not use any specific techniques. Differences in people using a technique and people not using a technique were not further investigated as they did not appear promising. The one participant that kept their eyes actively shut reported increasing tiredness in the previous questions. The tiredness might have been amplified by keeping the eyes actively shut. The light tapping movements of one participant might constitute a displacement activity. The use of such activities as an indicator for concentration is addressed in chapter 9.

In the best case scenario, the participants would not have known any of the poems beforehand. However, four participants already knew “Der Fischer”. “Der Fischer” was indeed the overall most liked poem. Two of the four participants ranked the poem in first place, the other two in second place. However, not only the four who already knew the poem rated it well. There were also no outliers in the ratings of these four in comparison to all participants. No direct effect of knowing poems beforehand is visible. This might be different if the participants knew the poem from a context other than school. For all participants, the school ended probably more than two years ago and memories of the poems might have faded.

Comments

As mentioned in the introduction (cf. chapter 1), the objective rating of art is a very controversial topic. This is reflected by the answers of the participants, all had an opinion on objective ratings of poetry recitals. Four think that these are possible, four think they are in parts possible and four think that they are not possible. The features for such ratings, proposed by the proponents, were all mentioned as positive or negative features of the speakers as well (cf. Descriptions). This suggests a certain objectivity of them but ignores any subjective component.

The question of the possibility of objective ratings cannot be conclusively answered, as this was also not intended. The results suggest tendentially that the subjective component in ratings predominates.

As no criticism of the study design was raised in the free comment section, the study does not require major modifications for the participants if repeated.

It was mentioned that it is difficult to separate the recitals and their text. The comment is contradicting the results from Ratings and Content. However, this was noted by only one participant. This participant’s ratings suggest that the content did not play a predominant role in the ratings, as the poems performed differently in each style. Thus, the person managed to separate text and recital despite the difficulties.

Two participants have stated that there are better Text-To-Speech voices available today. The Text-To-Speech voice was used because of its neutrality and lack of emotions. A direct comparison with more modern systems would be interesting in further research.

7.3 Physiological Data

In the boxplot trend analysis, a lot of plots had to be analyzed. For each, “Rating x Poem” and “Style x Poem”, 12 plots per physiological measure had to be analyzed, totalling up to 72 plots. Regarding the parameters, it is important to mention that in the analysis slow, strong shifts with in some cases comparatively high rise times and half recovery times (cf. Dawson et al., 2000) were considered. In the context of the study, these were of greater interest than small, fast changes. Importantly, the results could only be compared on an intra-poem level (cf. chapter 4).

The best suited physiological measures for each of the four poems, found in the analysis, differ. Some measures have trends for almost all poems, some only for individual ones. Due to the comparatively long recording times and the nature of the physiological data, there are many factors that may have influenced the measured values. Exact statements about the individual influences cannot be made. The discussion focuses on

similarities, however, they are not intended as comparisons. Generalizations are not made and differences are emphasized. The trends are exemplarily explained using the shared commonalities. All individual results can be found in chapter 5. The commonalities may anecdotally indicate generalizable trends, but need to be re-examined in a larger-scale follow-up study with a design that allows inter-poem comparisons. It must be ensured that the trends are not only coincidental (cf. chapter 9).
In particular, no judgment is made as to how well individual trends are suitable for the quantification of the experience, as the study design does not allow for this. However, statements are made about how the physiological measures seem to be appropriate for the task. Recommendations for one type of physiological data, based on the majority, do not exclude that others are also suitable or that individual ones are not suitable for other poems.

Interruptions in the connection to the mouse only occurred in two cases, and in each case within the first 10 seconds of the recital. This can be considered negligible for the trend analysis and no recordings were discarded due to connection interruptions.

Rating x Poem

The **Raw Mouse data** showed the most trends in the analysis of Rating and Poem (17). In the amplitudes, the interquartile range was trending for three out of the four poems (“Enschuldigungsbrief”, “Die Grenadiere”, and “Wahre Liebe”). Each time, the range goes up with increasing ratings. Reactions get more varied amplitudes with higher ratings. Of the rise times, the interfence range has the most promising trends. For two poems (“Die Grenadiere” and “Wahre Liebe”) it increases, i.e. the rise times get more varied. For one poem (“Der Fischer”) it decreases, i.e. the rise times get more similar. In general, the Raw Mouse mouse data changes for all poems with increasing rating in a predictable way.

In the **Heart Rate** data, there are fewer trends (12). Only the interquartile range of the amplitudes is applicable for at least three poems (“Entschuldingungsbrief”, “Der Fischer”, and “Die Grenadiere”). It decreases for all of them with increasing rating and therefore, the reactions get more similar.

Noticeable about the **Skin Conductance Response** is that there are no trends at all present for the rise times. Apparently, the quality of a recital has no influence on how fast the peak of the arousal, reflected in the Skin Conductance Response, is reached. For the amplitudes, there are only two matches in the median (“Die Grenadiere” and “Wahre Liebe”), and one match in the interfence range (“Entschuldigungsbrief”). The half recovery times have some trends, but not for all poems in each instance. For two poems (“Entschuldingungsbrief” and “Der Fischer”), the interfence and interquartile ranges go either up or down, i.e. the values get more or less varied, and for one poem (“Die Grenadiere”), the median increases, i.e. the state of arousal lasts longer. For two other poems (“Der Fischer” and “Die Grenadiere”), the outliers have trends.

Unfortunately, there is no kind of trend of a parameter that occurs for all poems. Quantifying the experience of poetry recitals is difficult, but the results are promising. The best approach is to use a combination of several parameters, which differs between the four poems. Some parameters which are promising for one poem are unsuitable for

other poems. For **“Entschuldingungsbrief”**, all parameters except the rise times of the Skin Conductance Response reflect the experience. For **“Der Fischer”**, only the Raw Mouse data, the amplitudes of the Heart Rate and the half recovery times of the Skin Conductance Response reflect the experience. For **“Die Grenadiere”**, every parameter besides the rise times of Heart Rate and Skin Conductance Response reflect the experience. For **“Wahre Liebe”**, the amplitudes and the rise times of the Raw Mouse data, the rise times of the Heart Rate and the amplitudes of the Skin Conductance Response reflect the experience.

Trends in the parameters are already evident with just 12 participants. The occurrence of trends with larger cohorts is subject to further research.

Style x Poem

Because the mean ratings of the recitals in the three styles do not overlap and therefore rating and style are somewhat linked, similar results to “Rating x Poem” were expected. However, the inter-individual component cannot be neglected here.

Of the **Raw Mouse data**, the rise times’ interfence range works well for the inference of the reciter’s skill for three poems (“Entschuldigungsbrief”, “Die Grenadiere”, “Wahre Liebe”). For these poems, it increases, meaning the rise times vary more with increasing skill.

Regarding the **Heart Rate** only the outliers of the half recovery times show trends for at least three poems (“Entschuldigungsbrief”, “Der Fischer”, and “Die Grenadiere”). The number of outliers increases. With increasing skill, some participants have notably longer reactions. The rise times are very suited for inferring the style of “Der Fischer”. Three of the four kinds of possible trends apply here. The rise times get longer on average but also more varied. More than half of the participants take longer to reach the peak of the reaction.

Of the **Skin Conductance Response**, neither amplitudes nor rise times or half recovery times have trends for more than one poem each.

In some cases, the data of one kind of measure is only feasible for the quantification for one specific poem. For the application of the findings, that means that the selection of measure is dependent on the poem. For **“Entschuldingungsbrief”**, all parameters except the amplitudes of the Heart Rate and the half recovery times reflect the style. For **“Der Fischer”**, the half recovery times of the Raw Mouse data, all parameters of the Heart Rate and no parameters of the Skin Conductance Response reflect the style. For **“Die Grenadiere”**, the rise times of the Raw Mouse data, the half recovery times of the Heart Rate and the half recovery times of the Skin Conductance Response reflect the style. For **“Wahre Liebe”**, the rise times of the Raw Mouse data, the amplitudes and the rise times of the Heart Rate, and the amplitudes and the rise times of the Skin Conductance Response reflect the style. As before, this observation should be verified in a larger study in future work.

The results show that the skill of the reciter affects various physiological parameters. In general, the quality of a recital, i.e. the rating, has a greater impact (more trends in the data) on the physiological experience than the style. There is no obvious correlation between the trends from the two analyses.

Example Time-Series

The Heart Rate time-series examined for “Die Grenadiere” stems from participant number 6. Her answers to the questionnaire are for the most part in line with the majority. She gave the professional the highest rating, the amateur and the Text-To-Speech received the same rating. Except for the amateur, the ratings were all within one standard deviation of the mean rating. This shows that she thought the amateur was worse than most of the other participants thought.

Based on the ratings, it was expected that the physiological response to the Text-To-Speech style and amateur style would be similar, and much stronger to the professional style.
The number of peaks in the data fulfill this expectation. The reaction to the Text-To-Speech and amateur style each have three peaks and to the professional style four peaks. However, the reaction to the amateur has the highest mean amplitude (4.88 bpm). The mean rise times are all close to each other, but the reaction to the amateur style has the lowest mean rise time (12.34 s). Furthermore, the amateur style reaction has the longest mean half recovery time (14.44).

The results indicate that participant 6 liked the amateur style recital the best. However, as mentioned above, this does not match her ratings.
The analysis shows that for this participant, the Heart Rate is not suitable for inferring the degree of liking of this recital from the physiological experience. This is generally consistent with the results from the trend analysis. However, this does not rule out the possibility that the Heart Rate may still be a suitable measure for other participants or poems.

Correlation

Regarding the one-way ANOVAs, 21 of the 72 tested parameters are significant.
When looking at the rating, the highest number of significant parameters are in the Skin Conductance Response. However, as with the trends before, no parameter is significant for all poems. **“Entschuldigungsbrief”** has significant parameters in all physiological measures. **“Der Fischer”** has significant parameters in the Heart Rate and Skin Conductance Response. **“Die Grenadiere”** has only significant parameters in the Skin Conductance Response. **“Wahre Liebe”** has significant parameters in the Heart Rate and the Skin Conductance Response. Contrary to the results of the trend analysis, the Raw Mouse data is the most unsuitable physiological measure regarding the rating, i.e. has the least significant measures. In the trend analysis, it showed the most trends.
When looking at the style, **“Entschuldigungsbrief”** has significant parameters in the Raw Mouse data and the Heart Rate. **“Der Fischer”** has significant parameters in none of the physiological measures. **“Die Grenadiere”** has significant parameters in the Raw Mouse data and the Skin Conductance Response. **“Wahre Liebe”** has significant parameters in the Raw Mouse data and the Heart Rate.
As before, different parameters suit different poems. Rating and style have different parameters with significance. The combination of parameters needs to be adapted to each specific poem.

In the discussion of the analysis with Spearman's Rank correlation coefficient, only the 16 variables with a comparably higher coefficient, i.e. lower than -0.5 or above 0.5, are considered.

For all poems, the mean frequency of a recital is negatively correlated with the rating. Recitals with lower mean frequencies have higher ratings. This was to be expected, as the professional style has the lowest mean frequencies and the highest ratings, and the Text-To-Speech style the highest mean frequencies and the lowest ratings. However, it is not possible to draw general conclusions about the rating from the average frequency. It stands to reason that speakers with a deep voice can also recite poorly. Whether there is a general bias for deeper voices would have to be analyzed in a separate study. Some of the participants did indeed explicitly like the deep voice of the professional and some did explicitly dislike it.

For "Entschuldigungsbrief" and "Der Fischer", more peaks in the Heart Rate are associated with higher ratings. For these two poems, a higher standard deviation of the rise times is associated with higher ratings as well. The better the recital is, the more varied are the rise times. For poorer recitals, participants have more similar responses. For "Entschuldigungsbrief" and "Wahre Liebe" a higher standard deviation of the half recovery times of the Heart Rate is correlated with higher ratings.

All other variables with a correlation lower than -0.5 or above 0.5 only occur for one poem each. For "Entschuldigungsbrief", higher ratings are associated with higher, more varied amplitudes and shorter rise times in the Heart Rate, as well as shorter and more diverse half recovery times in the Heart Rate.

For "Wahre Liebe", the rise times increase with higher ratings. The half recovery times of the Heart Rate get more varied with higher ratings. With higher ratings, and therefore better recitals, more peaks, generally higher values, more varied amplitudes, higher and more diverse rise times and more diverse half recovery times in the Skin Conductance Response are associated.

The poem "Die Grenadiere" has no physiological measures with a coefficient above the threshold.

In the case of "Wahre Liebe", the Skin Conductance Response has several indicators for the rating in addition to the Heart Rate. In the Heart Rate, the rise times of the Heart Rate increase with higher ratings and the half recovery times get more varied. In the Skin Conductance Response, higher rated recitals tend to have more peaks and the amplitudes, rise times, and half recovery times get more varied.

Here, the Heart Rate performs better than earlier and the Raw Mouse data considerably worse than in the trend analysis. No coefficients in the Raw Mouse data make the threshold. Different perspectives on the same data may produce different results. The findings of the correlation analysis and the trend analysis are not mutually exclusive. A combination of the different perspectives is suggested for quantifying the physiological experience of poetry recitals.

Confidence Interval

The 95% Confidence Intervals reaffirm the great inter-individual differences in the physiological reactions, supporting that the experience of poetry recitals is subjective.

Heatmap

The heatmaps nicely show that the peaks occur while the audio signal, i.e. the recital, is audible. This suggests that the peaks are related to what is heard and not triggered by other events. The expected clusters of peaks were not detected, which again highlights the inter-individual differences and diverse preferences of the participants.

7.4 Miscellaneous

The order distribution of the poems was not perfectly balanced. A better distribution was expected for "Entschuldigungsbrief" and "Wahre Liebe". However, in random distributions, outliers are to be expected. Order effects should still be negligible when interpreting the results.

The distribution of the recital styles in the three bins is according to the expectations. Most of the ratings in bin 1 are of the Text-To-Speech style (64.29%), most of the ratings in bin 2 are of the amateur style (50.98%), and most of the ratings in bin 3 are of the professional style (91.89%). The same is true when looking at the distribution of the ratings in the styles. Most of the Text-To-Speech recitals are in bin 1 (75.00%), most of the amateur recitals are in bin 2 (54.17%), and most of the professional recitals are in bin 3 (70.83%). When looking at the mean frequencies and the lengths of the recitals, longer and deeper recitals achieved higher ratings. This is in line with the results from the correlation analysis and is mainly influenced by the professional speaker, who spoke very slowly and low.
Overall, fewer high ratings were given, resulting in 38.89% of all ratings being in bin 1 and 25.69% of all ratings being in bin 3. The participants seem to be more hesitant to give high ratings, as they might think that there is always room for improvement.

The female participants gave slightly higher ratings on average (5.57) than the male participants (5.17). The older half of the participants gave higher ratings (5.72) than the younger half (5.02). However, these effects are minor. They likely cannot be attributed to their demographics. Effects of rhyme scheme, metre or cadence were not discovered.

7.5 Machine Learning

Even though predicting the rating can be considered a regression task, classifier algorithms performed best for all four poems. With an average improvement of 29.47% over the baseline, the models perform better than expected. The results still leave room for improvements. However, they are sufficient as a first orientation. The average Mean Absolute Error of 0.51 fits to the fact that bin 2 is the most predicted bin. With 38.68% of the data classified as bin 2, the classification is not overly skewed towards one single class.

The average improvement of the style prediction, seems high (69.03%). However, the average F1-score is only 0.5408. Considering the limited amount of data and the high inter-individual differences of the participants, these results are still good. The classification is a bit skewed towards the professional style, as 48.56% of the predictions are the professional style. In the training data, 47.12% are of this class. With more data, it

would have been possible to balance the sets, but it was more important to keep the number of training data points as high as possible. After all, the aim was only to gain a first insight.

The results on the small amount of data are very promising. They suggest that with training on more data and with better tuning, further improvements of the models can be achieved. Possibly, a custom architecture could also be developed. Based on the results so far, it can be stated that predicting the rating, as well as predicting the skill of a reciter with the help of physiological measures is feasible. This also supports the findings in section 7.3, i.e. that the experience is encoded in the physiological data.

7.6 Hypothesis

The trends found in the physiological data, as well as the results from the correlation analysis support the presented hypothesis. The quality of poetry recitals can be quantified measuring physiological reactions. The literature research from chapter 3 and the findings in the Machine Learning task also support the hypothesis.
The results show that different physiological measures are suitable for the quantification of each poem. With different analysis methods, different measures are suitable. This suggests that depending on the poem, a different combination of physiological measures and analysis methods should be used to quantify the experience. The single one indicator for all poems was not found. However, the study design would not have allowed a generalization to other poems anyway. To further support the findings and to be able to make general statements on the topic, more research is needed (cf. chapter 9). The present study is only a first step on quantifying the experience of poetry recitals.

Section 7.2 shows how the different recital styles are perceived by the participants. The features that make a good recital are also mentioned there. Section 7.3 highlights the high intra-individual differences in the reactions. These make the automated quantification difficult. The perception of the recitals is very individual, although section 7.2 shows that there are many common preferences and perspectives. The assessments of the three recital styles, e.g. via the rating (cf. section 7.2), show that the assumptions about the recital quality were correct when designing the study.
Section 7.3 shows that Heart Rate and Skin Conductance Response indeed are adequate measures for the experience of poetry recitals. However, the suitability highly depends on the poem.
No direct statement can be made about whether better recitals elicit stronger emotions with the data analysis carried out. The literature suggests such a relationship (cf. chapter 3). In addition, trends that allow for the quantification were found in the physiological data.

8 Conclusion

This thesis provided a small glimpse into the topic of *quantifying the experience of poetry recitals by physiological means*.

First, a novel hypothesis was presented. It was derived from existing work on related topics. The hypothesis states “The quality of poetry recitals can be quantified, measuring physiological reactions”. In the literature review, further existing research was discussed and its relevance to the study was explained. A small scale study was designed and conducted as a method to test the hypothesis. Four German poems in three different recital styles were played to 12 participants while recording their physiological reactions. The study also included a questionnaire. As part of the study, a Python framework was developed, that can be reused in other physiological studies that want to investigate reactions to stimuli. Tools for analyzing the data are also part of this framework. As the physiological data, the Heart Rate and Skin Conductance Response were recorded. Psychological data was obtained via questionnaires. The data was analyzed using various methods, such as a trend analysis, a correlation analysis, and a qualitative analysis, and presented. Furthermore, Machine Learning models were trained on the data to further test the hypothesis. In the last chapter, an outlook on future research possibilities is given.

The study shows that the proposed idea is feasible and suggests that the experience of poetry recitals can be quantified by physiological means. A number of trends and correlations were found that can be used to draw conclusions about the experience and thus the liking or disliking of a poetry recital. The fact that in poetry recitals, as in art in general, the subjective perception is of high importance can be seen in the large inter-individual differences of the reactions. This is also why the chosen physiological measures and the method of analysis must be adapted to each individual poem in order to be able to quantify the experience. In future research, the findings need to be verified with a larger scale study, so that generalizing statements can be made and the results can be applied in practice.

The questionnaire showed that the pronunciation and intonation of the speaker are most important to the participants. Other important characteristics are, for example, dramaturgical pauses, and the rate of speech. These distinguish a bad recital from a good one and should therefore be taken into account by Text-To-Speech systems for poetry recitals.

9 Future Work

As the present study was only small scale to assess the topic, the data collection could be extended and improved. However, the first results already can be used in real world applications.

9.1 Research

The presented data analysis was not fully exhaustive. In a further analysis, the recitals could be separated into early, mid, and late segments, which would reveal trends within the recitals. The segmentation could also be done based on text features like stanzas. This could be useful for comparing the results to Wassiliwizky et al. (2017). Based on the collected data, it could also be looked at where the peaks occur in the texts, i.e. correlate the texts with the peaks, which can inform which text features are relevant for the emotional experience. As in Leite et al. (2019), a sliding window approach could be tested for the analysis of changes in physiological measures.

For a higher external validity, i.e. generalizability, of the results, the presented study should be expanded. A more diverse cohort could be part of such a follow up study. Most importantly, the number of participants should be increased to be able to make more generally valid observations. In the present study, all participants were between 20 and 32 years old. Older participants with more life experience might have a higher interest in poetry and experience the recitals differently than a younger population. The participants were also quite homogeneous in terms of their occupations. Different levels of education might lead to different perceptions too.

The physiological data was limited to two kinds, the Heart Rate and the Skin Conductance Response. Additionally, an eye tracker could be used to record measures like pupil size, blinking or eye movement. However, this would only be possible if the closed eye setting was discarded, which might introduce visual distractions. As an additional measure, it could be looked at the brain activation with fMRI, like Wassiliwizky et al. (2017). Such additions could help to better understand the affective perception. Also, a professional measuring device could be used for the Heart Rate and thereby more precise information on the changes in the Heart Rate could be collected.
As mentioned in the discussion (cf. chapter 7), the participants did not experience any chills, unlike in Wassiliwizky et al. (2017). It should be investigated whether this is due to the fact that in Wassiliwizky et al. (2017) the participants were asked explicitly about chills and therefore focused on them, or whether there are other reasons for the disparity in the results.

Besides expanding the cohort and introducing additional measures, there are variables of the recitals that could be changed to achieve different perceptions. Speakers with different gender or cultural backgrounds could be used. In a conversation after the study, a female participant suggested that female speakers potentially could get rated worse, as the standard voice in media and advertisement is "male" and people are more used to it. This would be an interesting topic for interdisciplinary research with other disciplines like Gender Studies.

One of the speakers of this study, the professional, might have been familiar to some participants. Either from advertising or from his work as a voice actor. In a follow up study, the recognition of a voice could be factored in as a potentially influencing variable.

Some participants suggested that culture has an influence on the perception: "All people (of a cultural space/language space) have had similar experiences with the connection of language, speech, character, emotionality of individuals." (*Alle Menschen (eines kulturellen Raumes/eines Sprachraumes) [haben] ähnliche Erfahrungen bei den Zusammenhang von Sprache, Sprechweise, Charakter, Emotionalität von Individuen gemacht [...].*) Such an impact could be investigated. As all poems were presented in German only, no statements can be made for other languages. Depending on if the speaker or listener is native or non-native, the perception could differ. A speaker's dialect might have an influence as well.

The analysis of the distribution of the poem order showed that it was not perfectly balanced. In a follow-up study, the balancing should be improved.

Additionally to the three tested styles, the reactions to more recital styles or speakers could be analyzed. For example, the reactions to a more modern Text-To-Speech system, a specially improved Text-To-Speech system for poetry (cf. section 9.2) or to a musical interpretation could be analyzed.

Further, the possibility to measure the experience of recitals of other, similar emotional, genres of text could be investigated. It stands to reason that similar reactions could be expected. However, the preferred characteristics for the speaker may differ.

The study only looked at auditory stimuli. The physiological experience of input over different sensory channels could be compared. The different input channels could be combined for a new poetry experience, building on the work of Arellano et al. (2014).

As a measure for the concentration level of the participants, the mouse clicks could be counted. During the study, the author noticed that some participants excessively clicked the Heart Rate measuring mouse. One participant even mentioned "[...] light tapping movements with fingers" (*[...] leichte Tippbewegungen mit Fingern*) when asked about their technique for concentrating in the questionnaire. Higher click rates could indicate more or less attention. The results of van Drunen et al. (2009) suggest such a correlation. Other displacement activities, like lip biting, might serve as indicators of attention as well. However, the quantification might be difficult.

9.2 Application

As a practical application of the findings, a specialized Text-To-Speech system for poetry recitals could be developed. Such a system should be able to incite similar or stronger physiological reactions compared to the amateur reciter in this study. For maximum usefulness, the tool could function as post-processing unit for existing Text-To-Speech systems.

The tool could mimic several voices, matching the poem's different characters. By this, it would add to the acting component of a recital. The different voices could be made more natural trough random changes in the voices' characteristics in a predefined range. This would result in unique vocalizations, as they always sound a little bit different and thereby more natural (cf. Scherer, 1995).

The workflow of such a tool could look like this: First, the different speaking characters of a poem are identified, for example, with the help of Speaker Recognition (e.g. Yeung and Lee, 2017). For the different speakers, the voice styles are computed by extracting the involved emotions, rhyme, metre, patterns, et cetera from the text. In parallel, a generic Text-To-Speech recital is generated by a standard system. The poem's text and the recital's audio are automatically aligned, using Text-To-Speech Alignment (e.g. Serrière et al., 2016). Afterwards, the voices are adapted according to the computed styles and finally, the random factor (see above) is added. Important features for adaption are, amongst others, the pronunciation, the intonation, the rate of speech, and dramaturgical pauses.

Such a poetry Text-To-Speech system should be evaluated similarly to the present study. It could help blind and dyslexic people to experience poetry without the help of others. Furthermore, it could be applied to other, similarly emotional, genres of text. Implemented in a voice assistant, it could make the generated speech more emotional and more natural.

Alternatively, the results can be used to improve human recitals. For example, the amateur could pay more attention to dramatic pauses. Moreover, based on the Machine Learning models, it could be tried to estimated how well a person liked a recital, i.e. quantify their individual experience. Using additional data, e.g. from people with similar reactions and text based features, poetry recitals could be suggested that potentially trigger a similar experience.

Acknowledgements

I would like to thank several people with respect to my thesis. Throughout the last months they supported me and my work in different ways. Without them, the project would not have been the same.

Sophie Grimme was one of the very first people that I got to know when I moved to Weimar. At the time of writing she wrote on her own thesis in HCI. Over the last months we had a great exchange and were able to support us mutually.

Fabian Post is a colleague at the chair of transport system planning. We started working there at about the same time and soon became friends. He proofread all of my content critically and always supplied me with chocolate milk.

Luise Kraaz is a another colleague from work. She joined the team only recently but we quickly became friends. As a native English speaker, she volunteered to proofread my text as well.

Julius Uhlmann is the third colleague from work that supported me. He recorded the amateur recitals for me. When thinking about potential amateur narrators, he was the first that came to my mind. As expected, he has done a very good job.

When looking for professional poetry recitals, I quickly stumbled upon the website of **Fritz Stavenhagen**. Later I noticed that I already know his voice from my childhood as he narrated the children's book series "Ritter Rost". He kindly provided the four amazing professional recitals.

Prof. Dr.-Ing. Eva Hornecker is the professor for Human-Computer Interaction and the head of the Master's programme. She let me use her templates for both the info sheet and the consent form.

Finally, I would like to thank **Jun.-Prof. Dr. phil. Jan Ehlers** and **Dr. phil. Magdalena Wolska**. They supervised this thesis. They gave me a lot of freedom to realize my ideas and valuable feedback on my work.

Bibliography

amazon.com. 2021. Sophisticated Ape Robotic Poetry Reading [online]. Accessed: 21.10.2021, 3:45pm.

Oscar Araque, Lorenzo Gatti, Jacopo Staiano, and Marco Guerini. 2018. DepecheMood++: a Bilingual Emotion Lexicon Built Through Simple Yet Powerful Techniques. *arXiv preprint arXiv:1810.03660*.

Diana Arellano, Simon Spielmann, and Volker Helzle. 2014. The Muses of Poetry-In search of the poetic experience. In *CHI 2014 Extended Abstracts on Human Factors in Computing Systems*, page 383–386, New York, USA. ACM.

Hassan A. Aziz. 2017. Comparison between Field Research and Controlled Laboratory Research. *Archives of Clinical and Biomedical Research*, 1(2):101–104.

Archana Balyan, SS Agrawal, and Amita Dev. 2013. Speech synthesis: a review. *International Journal of Engineering Research & Technology (IJERT)*, 2(6):57–75.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. O'Reilly Media Inc., Sebastopol, USA.

Daniel T. Burley and Stephanie H M van Goozen. 2020. Pupil Response to Affective Stimuli: a Biomarker of Early Conduct Problems in Young Children. *Journal of abnormal child psychology*, 48(5):693–701.

Michel Cabanac. 2002. What is emotion? *Behavioural processes*, 60:69–83.

Cambridge Advanced Learner's Dictionary & Thesaurus. 2021. recital [online]. Accessed: 15.11.2021, 09:30pm.

Marcela Charfuelan. 2012. MARY TTS HMM-based voices for the Blizzard Challenge 2012. In *Blizzard Challenge*, Portland, USA. Carnegie Mellon University.

Marcela Charfuelan and Ingmar Steiner. 2013. Expressive speech synthesis in MARY TTS using audiobook data and EmotionML. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pages 1564–1568, Lyon, France. International Speech Communication Association.

George I. Christopoulos, Marilyn A. Uy, and Wei Jie Yap. 2019. The Body and the Brain: Measuring Skin Conductance Responses to Understand the Emotional Experience. *Organizational Research Methods*, 22(1):394–420.

Michelle Cohn, Eran Raveh, Kristin Predeck, Iona Gessinger, Bernd Möbius, and Georgia Zellou. 2020. Differences in Gradient Emotion Perception: Human vs. Alexa Voices. In *Proc. Interspeech 2020*, pages 1818–1822, Shanghai, China. International Speech Communication Association.

Michael E. Dawson, Anne M. Schell, and Diane L. Filion. 2000. The electrodermal system. In John T. Cacioppo, Louis G. Tassinary, and Gary G. Berntson, editors, *Handbook of Psychophysiology*, 2nd edition, pages 200–223. Cambridge University Press.

Thierry Declerck. 2017. A Set of Annotations for supporting a TTS Application for Folktales. In *Proceedings of the Workshop on Corpora in the Digital Humanities (CDH 2017)*, pages 58–63, Bloomington, USA.

Rodolfo Delmonte. 2013. Computing Poetry Style. In *Proceedings of the First International Workshop on Emotion and Sentiment in Social and Expressive Media: approaches and perspectives from AI (ESSEM 2013)*, pages 148–155, Turin, Italy. Italian Association for Artificial Intelligence.

Rodolfo Delmonte. 2019. SPARSAR recites Shakespeare's Sonnets–and Coping with Early Modern English variants. In *Proc. 8th AIUCD Conference 2019*, pages 46–48, Udine, Italy. Associazione per l'Informatica Umanistica e la Cultura Digitale.

Rodolfo Delmonte and Anton Maria Prati. 2014. SPARSAR: An Expressive Poetry Reader. In *Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 73–76, Gothenburg, Sweden. Association for Computational Linguistics.

deutschelyrik.de. 2021. Projekt Deutsche Lyrik [online]. Accessed: 14.06.2021, 08:30pm.

Annemiek van Drunen, Egon L. van den Broek, Andrew J. Spink, and Tobias Heffelaar. 2009. Exploring workload and attention measurements with uLog mouse data. *Behavior Research Methods*, 41(3):868–875.

Christian Eisenreich, Jana Ott, Tonio Süßdorf, Chistian Willms, and Thierry Declerck. 2014. From Tale to Speech: Ontology-based Emotion and Dialogue Annotation of Fairy Tales with a TTS Output. In *Proceedings of ISWC 2014*, pages 153–156, Riva del Garda, Italy. Springer.

Paul Ekman. 1999. Basic Emotions. In T. Dalgleish and M. Power, editors, *Handbook of Cognition and Emotion*, chapter 3, pages 45–60. John Wiley & Sons, Sussex, Great Britain.

Florian Eyben, Sabine Buchholz, and Norbert Braunschweiler. 2012. Unsupervised clustering of emotion and voice styles for expressive TTS. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4009–4012, Kyoto, Japan. IEEE.

fritzstavenhagen.de. 2021. Sprecher [online]. Accessed: 14.06.2021, 08:30pm.

Jack Gandour, Donald Wong, Mario Dzemidzic, Mark Lowe, Yunxia Tong, and Xiaojian Li. 2003. A cross-linguistic fMRI study of perception of intonation and emotion in Chinese. *Human Brain Mapping*, 18(3):149–157.

Alberto Greco, Gaetano Valenza, Antonio Lanata, Enzo Pasquale Scilingo, and Luca Citi. 2016. cvxEDA: A Convex Optimization Approach to Electrodermal Activity Processing. *IEEE transactions on bio-medical engineering*, 63(4):797–804.

Charles R. Harris, K. Jarrod Millman, Stéfan J van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. 2020. Array programming with NumPy. *Nature*, 585:357–362.

Henderi, Tri Wahyuningsih, and Efana Rahwanto. 2021. Comparison of Min-Max normalization and Z-Score Normalization in the K-nearest neighbor (kNN) Algorithm to Test the Accuracy of Types of Breast Cancer. *International Journal of Informatics and Information Systems*, 4(1):13–20.

J. D. Hunter. 2007. Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, 9(3):90–95.

Yannick Jadoul, Bill Thompson, and Bart de Boer. 2018. Introducing Parselmouth: A Python interface to Praat. *Journal of Phonetics*, 71:1–15.

David Maxwell Kaplan. 2007. Computational Analysis and Visualized Comparison of Style in American Poetry. *Princeton University*.

Jan Kercher. 2013. *Verstehen und Verständlichkeit von Politikersprache Verbale Bedeutungsvermittlung zwischen Politikern und Bürgern*. Springer VS, Wiesbaden, Germany.

Evgeny Kim and Roman Klinger. 2018. Who Feels What and Why? Annotation of a Literature Corpus with Semantic Roles of Emotions. In *Proceedings of COLING 2018, the 27th International Conference on Computational Linguistics*, pages 1345–1359, Santa Fe, USA. Association for Computational Linguistics.

Stefan Koelsch, Thomas Fritz, D. Yves v. Cramon, Karsten Müller, and Angela D. Friederici. 2006. Investigating emotion with music: An fMRI study. *Human Brain Mapping*, 27(3):239–250.

Sofia Leite, Miguel S. Dias, Sara Eloy, João Freitas, Sibila Marques, Tiago Pedro, and Lázaro Ourique. 2019. Physiological Arousal Quantifying Perception of Safe and Unsafe Virtual Environments by Older and Younger Adults. *Sensors*, 19(11):2447.

Philipp Lorenz-Spreen, Bjarke Mørch Mønsted, Philipp Hövel, and Sune Lehmann. 2019. Accelerating dynamics of collective attention. *Nature Communications*, 10(1):1759.

Suhas R. Mache, Manasi R. Baheti, and C. Namrata Mahender. 2015. Review on Text-To-Speech Synthesizer. *International Journal of Advanced Research in Computer and Communication Engineering*, 4:54–59.

Dominique Makowski, Tam Pham, Zen J. Lau, Jan C. Brammer, François Lespinasse, Hung Pham, Christopher Schölzel, and S. H. Annabel Chen. 2021. NeuroKit2: A Python toolbox for neurophysiological signal processing. *Behavior Research Methods*.

Merriam-Webster. 2021. recite [online]. Accessed: 15.11.2021, 09:30pm.

John Stuart Mill. 1860. Thoughts on Poetry and Its Varieties. *The Crayon*, 7(4):93–97.

Saif M. Mohammad. 2021. NRC Word-Emotion Association Lexicon [online]. Accessed: 14.06.2021, 02:00pm.

Saif M. Mohammad and Peter D. Turney. 2013. Crowdsourcing a Word-Emotion Association Lexicon. *Computational Intelligence*, 29(3):436–465.

M. M. Mukaka. 2012. Statistics corner: A guide to appropriate use of correlation coefficient in medical research. *Malawi medical journal : the journal of Medical Association of Malawi*, 24(3):69–71.

Clifford Nass, Ulla Foehr, Scott Brave, and Michael Somoza. 2001. The Effects of Emotion of Voice in Synthesized and Recorded Speech. In *Proceedings of the AAAI symposium emotional and intelligent II: The tangled knot of social cognition*, North Falmouth, USA. Association for the Advancement of Artificial Intelligence.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Joan Peskin. 2008. The genre of poetry: Secondary school students' conventional expectations and interpretive operations. *English in Education*, 41:20–36.

Plotly Technologies Inc. 2015. Collaborative data science [online]. Accessed: 13.07.2021, 10:00pm.

Robert Plutchik. 1980. *Emotion, a psychoevolutionary synthesis*. Harper and Row, New York, USA.

ranieriandco.com. 2021. Changing Attention Span and What it Means for Content in 2021 [online]. Accessed: 24.10.2021 23:15pm.

Antonio Rebordão, Mostafa Shaikh, Keikichi Hirose, and Nobuaki Minematsu. 2009. How to Improve TTS Systems for Emotional Expressivity. In *10th Annual conference of the International Speech Communication Association (Interspeech 2009)*, pages 520–523, Baixas, France. International Speech Communication Association.

Nikki S. Rickard. 2004. Intense emotional responses to music: a test of the physiological arousal hypothesis. *Psychology of Music*, 32(4):371–388.

Klaus Scherer. 1995. Expression of Emotion in Voice and Music. *Journal of Voice*, 9(3):235–248.

Marc Schröder. 2001. Emotional Speech Synthesis: A Review. In *Proceedings of the 7th European Conference on Speech Communication and Technology: Vol. 1. Eurospeech 2001*, pages 561–564, Aalborg, Denmark. International Speech Communication Association.

Marc Schröder, Laurence Devillers, Kostas Karpouzis, Jean-Claude Martin, Chatherine Pelachaud, Christian Peter, Hannes Pirker, Björn Schuller, Jianhua Toa, and Ian Wilson. 2007. What should a generic emotion markup language be able to represent? In *Affective Computing and Intelligent Interaction*, pages 440–451, Lisbon, Portugal. Springer.

Guillaume Serrière, Christophe Cerisara, Dominique Fohr, and Odile Mella. 2016. Weakly-supervised text-to-speech alignment confidence measure. In *Proceedings of*

COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 2042–2050, Osaka, Japan. The COLING 2016 Organizing Committee.

Oliver Singler. 2019. Automatic Annotation of Emotions on German Metaphors. Bachelor's thesis, Heidelberg University, Germany.

Richard Sproat, Andrew Hunt, Mari Ostendorf, Paul Taylor, Alan W. Black, Kevin Lenzo, and Mike Edington. 1998. SABLE: A Standard For TTS Markup. In *Prof. of 5th International Conference on Spoken Language Processing (ICSLP-1998)*, Sydney, Australia. International Speech Communication Association. Paper 0040.

Fritz Stavenhagen. 2019. *Einführung in die Lyrik, Grundlagen und Formen der Poesie*. BoD - Books on Demand, Norderstedt, Germany.

Gerhard Stemmler. 2004. Physiological processes during emotion. In *The regulation of emotion*, pages 48–85. Psychology Press.

Thales Teixeira, Michel Wedel, and Rik Pieters. 2012. Emotion-Induced Engagement in Internet Video Advertisements. *Journal of Marketing Research*, 49(2):144–159.

uni-weimar.de. 2021. Datenschutz [online]. Accessed: 18.04.2021, 00:45pm.

Chin-An Wang, Talia Baird, Jeff Huang, Jonathan D. Coutinho, Donald C. Brien, and Douglas P. Munoz. 2018. Arousal Effects on Pupil Size, Heart Rate, and Skin Conductance in an Emotional Face Task. *Frontiers in Neurology*, 9.

Christian Wartena. 2019. A Probabilistic Morphology Model for German Lemmatization. In *Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019)*, pages 40–49, Erlangen, Germany. German Society for Computational Linguistics & Language Technology.

Eugen Wassiliwizky, Stefan Koelsch, Valentin Wagner, Thomas Jacobsen, and Winfried Menninghaus. 2017. The emotional power of poetry: neural circuitry, psychophysiology and compositional principles. *Social Cognitive and Affective Neuroscience*, 12(8):1229–1240.

John B. Watson and J.J.B. Morgan. 1917. Emotional Reactions and Psychological Experimentation. *American Journal of Psychology*, 28:163–174.

Peter Weir. 1989. Dead Poets Society. Touchstone Pictures.

Thom Wettstein. 2018. Loudness Messung endlich verständlich! [online]. Accessed: 05.08.2021, 22:15pm.

David Yanofsky, Barry van Driel, and James Kass. 1999. "Spoken Word" and "Poetry Slams": the voice of youth today. *European Journal of Intercultural studies*.

Chak Yan Yeung and John Lee. 2017. Identifying Speakers and Listeners of Quoted Speech in Literary Works. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 325–329, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Heiga Zen, Keiichi Tokuda, Takashi Masuko, Takao Kobayashi, and Tadashi Kitamura. 2004. Hidden Semi-Markov Model Based Speech Synthesis. In *Proc. of 8th International Conference on Spoken Language Processing (INTERSPEECH 2004 - ICSLP)*, pages 1393–1396, Jeju Island, Korea. International Speech Communication Association.

Deyu Zhou, Xuan Zhang, Yin Zhou, Quan Zhao, and Xin Geng. 2016. Emotion Distribution Learning from Texts. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 638–647, Austin, USA. Association for Computational Linguistics.

Appendix

Info Sheet

Bauhaus-Universität Weimar

Oliver Singler
+49 176 XX XX XX XX
oliver.singler@uni-weimar.de

Informationsdokument – Studie zur Masterthesis

Die Studie
In meiner Masterthesis befasse ich mich mit physiologischen (= körperliche) Reaktionen auf Gedichtsrezitationen (= Gedichtsvorträge). Diese Studie dient dazu eine meiner Hypothesen zu überprüfen.
Du wirst verschiedene Gedichte in verschiedenen Rezitationsstilen hören und anschließend einige Fragen dazu beantworten.

Aus Gründen des Infektionsschutzes findet die Studie in einem großen, gut belüfteten Raum statt. Ort der Studie ist Raum 305 in der Marienstraße 13C. Deine Teilnahme wird ca. eine Stunde dauern. Ich trage während der gesamten Zeit einen medizinischen Mund-Nasen-Schutz. Du darfst diesen während des Experiments ablegen. Der Versuchspersonensessel und alle Gegenstände werden nach jeder Sitzung gründlich desinfiziert.

Bitte bringe **deine eigenen Kopfhörer** (3,5mm Klinkenanschluss) mit.

Teilnahmevoraussetzungen sind Volljährigkeit und Deutsch als Muttersprache. Des Weiteren dürfen keine Aphasien (Beeinträchtigungen des Hör-Verstehens) oder starke Beeinträchtigungen des Hörens vorliegen.
Durch deine Teilnahme bestätigst du, dass du diese Voraussetzungen erfüllst.
Für die Teilnahme an dieser Studie und die Datenerfassung bitte ich dich um dein Einverständnis. Als Dankeschön gibt es nach der Teilnahme eine kleine Überraschung.

Aufgabenbeschreibung
Während der Studie wirst du in einem gemütlichen Sessel sitzen. Um visuelle Ablenkungen zu verhindern, wirst du eine Augenmaske tragen. Über deine Kopfhörer wirst du Rezitationen von vier Gedichten in drei Stilen hören.

Während der Rezitationen, musst du nur entspannt zuhören. Vor einer Rezitation wirst du jeweils einen 30-sekündigen, neutralen Text hören. Nach jeder Rezitation wirst du gebeten diese auf einer Skala von 1 (gefällt mir nicht) bis 10 (gefällt mir sehr) subjektiv zu bewerten. Wichtig ist, dass du die Rezitation und nicht das Gedicht bewertest. Zwischen den drei Stilen hast du je 30 Sekunden Pause. Die Pausen sage ich an. Insgesamt wird dieser Teil ca. 30 Minuten dauern.

Vor Beginn machen wir einen kurzen Probedurchlauf zur Gewöhnung und zum Einstellen der Lautstärke.

Nachdem alle Rezitationen gehört wurden, folgt ein ausführlicherer Fragebogen. Um die Fragen beantworten zu können, musst du bei den Rezitationen lediglich entspannt zuhören. Zur Erinnerung werden dir die Texte der Gedichte vorgelegt.

Bauhaus-
Universität
Weimar

Oliver Singler
+49 176 XX XX XX XX
oliver.singler@uni-weimar.de

Datensammlung
Vor Beginn der Studie gibt es einen Fragebogen mit demographischen Fragen. Während des Hörens der Gedichtsrezitationen werden Herzfrequenz, mittels Geräts in der Hand, und Hautleitwerte, über externe Sensoren an den Fingern, gemessen.

Wichtig ist der dauerhafte Kontakt zum Herzfrequenzmesser. Im Falle eines Kontaktverlustes ertönt ein Signalton und die Rezitation wird abgebrochen. Bei Wiederherstellung des Kontakts ertönt erneut ein Signalton und die letzte Rezitation wird neu gestartet. Eine erfolgreiche Kontaktwiederherstellung dauert ca. 10 Sekunden. In den Pausen darf der Kontakt unterbrochen werden.

Alle Daten werden pseudonymisiert und ausschließlich gemäß der Einverständniserklärung verwendet.

Datenschutz und Anonymität
Dein Name wird nicht mit den Forschungsergebnissen in Verbindung gebracht. Wo nötig, wird ein Pseudonym verwendet. Daten und Informationen, die Rückschlüsse auf einzelne Teilnehmende zulassen, werden nicht in der Thesis oder anderweitig veröffentlicht, sofern nicht explizit gestattet.

Teilnahme
Die Teilnahme an dieser Studie ist freiwillig. Du kannst deine Teilnahme jeder Zeit, ohne Angabe von Gründen, abbrechen und dein Einverständnis zurückziehen. Du darfst jeder Zeit Fragen zur Studie stellen. Du kannst dich dazu entscheiden Fragen von mir nicht zu beantworten oder die Verwendung deiner Daten zu untersagen, auch nach Beginn der Studie.

Wenn du mit der Teilnahme einverstanden bist, fülle bitte die Einverständniserklärung aus. Bei weiteren Fragen stehe ich gerne zur Verfügung.

Vielen Dank für Deine Unterstützung,

Oliver Singler

Consent Form

Bauhaus-
Universität
Weimar

Oliver Singler
+49 176 XX XX XX XX
oliver.singler@uni-weimar.de

Einverständniserklärung – Studie zur Masterthesis

Freiwilligkeit der Studie / Vertraulichkeit:
Die Teilnahme an dieser Studie ist freiwillig. Du kannst deine Teilnahme jeder Zeit, ohne Angabe von Gründen, abbrechen und dein Einverständnis zurückziehen.
Du darfst jeder Zeit Fragen zur Studie stellen. Du kannst dich dazu entscheiden Fragen von mir nicht zu beantworten oder die Verwendung deiner Daten zu untersagen, auch nach Beginn der Studie.

Dein Name wird nicht mit den Forschungsergebnissen in Verbindung gebracht. Wo nötig, wird ein Pseudonym verwendet. Daten und Informationen, die Rückschlüsse auf einzelne Teilnehmende zulassen, werden nicht in der Thesis oder anderweitig veröffentlich, sofern nicht explizit gestattet.

Einverständnis zur Teilnahme

Ich, _____

stimme der Teilnahme an dieser Studie zu.

Mir wurde der Zweck der Studie erklärt.
Ich hatte die Möglichkeit Fragen zur Studie zu stellen. Diese wurden mir zufriedenstellend beantwortet.
Ich wurde informiert, dass ich zu jedem Zeitpunkt, ohne Angabe von Gründen, die Teilnahme abbrechen und diese Einwilligung zurückziehen kann.
Mir wurde die Vertraulichkeit meiner persönlichen Daten zugesichert.
Ich stimme zu, dass die von mir zur Verfügung gestellten Informationen pseudonymisiert für Forschungszwecke, einschließlich Veröffentlichungen, verwendet werden dürfen. Hierbei werden meine persönlichen Daten vertraulich behandelt.

Ich habe verstanden, dass ich mich bei Fragen, Problemen oder Bedenken an
Oliver Singler, +49 176 XX XX XX XX, oliver.singler@uni-weimar.de
wenden kann.

Datum: _____ Unterschrift: _____

Demographics Questionnaire

Bauhaus-
Universität
Weimar

Oliver Singler
+49 176 XX XX XX XX
oliver.singler@uni-weimar.de

Fragebogen – Demographie – Studie zur Masterthesis

Pseudonym:	
Alter:	
Geschlecht:	
Tätigkeit:	

Final Questionnaire

Bauhaus-
Universität
Weimar

Oliver Singler
+49 176 XX XX XX XX
oliver.singler@uni-weimar.de

Fragebogen – Gedichtsrezitationen – Studie zur Masterthesis

Pseudonym:

In der dritten Spalte hast du die Möglichkeit deine Bewertungen (1-10) nochmals zu verändern.

Sprecher 1

_____: _____ _____
_____: _____ _____
_____: _____ _____
_____: _____ _____

Sprecher 2

_____: _____ _____
_____: _____ _____
_____: _____ _____
_____: _____ _____

Sprecher 3

_____: _____ _____
_____: _____ _____
_____: _____ _____
_____: _____ _____

Bauhaus-
Universität
Weimar

Oliver Singler
+49 176 XX XX XX XX
oliver.singler@uni-weimar.de

Bitte beschreibe die drei unterschiedlichen Sprecher für jemanden, der*die die Rezitationen nicht gehört hat.

Sprecher 1

Sprecher 2

Sprecher 3

Wie unterscheiden sich die Sprecher?

Bauhaus-
Universität
Weimar

Oliver Singler
+49 176 XX XX XX XX
oliver.singler@uni-weimar.de

Was gefällt dir an den Sprechern?

Sprecher 1

Sprecher 2

Sprecher 3

Was gefällt dir an den Sprechern nicht?

Sprecher 1

Sprecher 2

Sprecher 3

m

Bauhaus-
Universität
Weimar

Oliver Singler
+49 176 XX XX XX XX
oliver.singler@uni-weimar.de

Von welchem Sprecher würdest du dir am ehesten weitere Gedichtsrezitationen anhören?

- ☐ Sprecher 1
- ☐ Sprecher 2
- ☐ Sprecher 3

Bitte sortiere die Sprecher nach deiner persönlichen Vorliebe (1. = mag ich am meisten).

1. _____
2. _____
3. _____

Welcher Sprecher passt am besten zu den jeweiligen Gedichten?

Entschuldigungsbrief: _____

Der Fischer: _____

Die Grenadier: _____

Wahre Liebe: _____

Ordne den Gedichten (Text) die jeweils bedeutendste Emotion zu.
Mögliche Emotionen: Angst, Ekel, Erwartung, Freude, Trauer, Vertrauen, Wut, Überraschung

Entschuldigungsbrief: _____

Der Fischer: _____

Die Grenadier: _____

Wahre Liebe: _____

Bei welchem Sprecher wurden diese Emotionen am deutlichsten?

Entschuldigungsbrief: _____

Der Fischer: _____

Die Grenadier: _____

Wahre Liebe: _____

n

Bauhaus-
Universität
Weimar

Oliver Singler
+49 176 XX XX XX XX
oliver.singler@uni-weimar.de

Wodurch wurden diese Emotionen deutlich?

Bitte sortiere die Gedichte (Text) nach der Stärke ihrer Emotionalität (1. = am stärksten).

1. _____
2. _____
3. _____
4. _____

Bitte sortiere die Gedichte (Text) nach deiner persönlichen Vorliebe (1. = mag ich am meisten).

1. _____
2. _____
3. _____
4. _____

Hast du während des Hörens körperliche Veränderungen wahrgenommen?

☐ Ja
☐ Nein

Falls ja, welche? Wann?

o

Bauhaus-
Universität
Weimar

Oliver Singler
+49 176 XX XX XX XX
oliver.singler@uni-weimar.de

Hast du während des Hörens Veränderungen deiner Aufmerksamkeit wahrgenommen?

☐ Ja
☐ Nein

Falls ja, welche? Wann?

Hast du dich mit Hilfe einer bestimmten Technik auf die Rezitationen konzentriert?

Welche der Gedichte kanntest du schon vor der Studie? Woher?

Lässt sich die Qualität von Gedichtsrezitationen objektiv bewerten?

☐ Ja …
☐ Nein …
…, weil:

Falls ja, welche Merkmale könnten hierfür genutzt werden?

p

Bauhaus-
Universität
Weimar

Oliver Singler
+49 176 XX XX XX XX
oliver.singler@uni-weimar.de

Platz für weitere Gedanken:

q

Eidesstattliche Erklärung[1]

Ich gebe hiermit die eidesstattliche Erklärung ab, dass ich meine Masterarbeit über: “Quantifying the Experience of Poetry Recitals by Physiological Means” selbständig angefertigt, keine anderen als die angegebenen Quellen und Hilfsmittel benutzt und alle wörtlichen oder sinngemäß übernommenen Textstellen als solche kenntlich gemacht habe.

Weimar, 20. November 2021 Oliver Singler

[1] Text vom Gemeinsamen Prüfungsamt der Universität Heidelberg aus dem Jahr 2018