communicated by:
Lehr- und Forschungsgebiet Informatik 9

Prof Dr. Ulrik Schroeder

The present work was submitted to Learning Technologies Research Group
Diese Arbeit wurde vorgelegt am Lehr- und Forschungsgebiet Informatik 9

Voice in Teach-R
Stimme in Teach-R

Master-Thesis
Masterarbeit

presented by / von

Recker, Jona
367356

Prof. Ulrich Schroeder

Prof. Sven Kommer

Adviser: Björn Meißner

Supervised by / Betreut von
Birte Heinemann , M.Sc.

Aachen, March 27, 2024

Contents

List of Figures

List of Tables

Abstract

This thesis introduces the foundation for voice and speech analysis in Teach-R, a virtual environment for teachers-to-be to train their classroom management. Teachers use their voice on a daily basis and have to face various disruptive factors which can cause a strain on the voice. To address this issue, a voice screening can be conducted wherein the acoustic properties of the voice are assessed to ensure proper utilization of the vocal apparatus. Additionally, speech analysis can be applied to investigate prosodic factors of the speech of teachers to convey information to students. To implement these two features, feasibility was examined based on existing approaches, and parameters relevant to the target audience of teachers were discussed. In difference to a professional voice screening, the analysis needs to be done on connected speech and serves the purpose of helping user reflect on their usage of the voice, not to rate it. Several analyses were planned and introduced, from which a foundation for voice and speech analysis was integrated into Teach-R. Through an evaluation the functionality, usability, and helpfulness were assessed, with the result of the features being helpful to reflect on the usage of voice and speech, but further refinements are needed for it to be a pleasant application.

Chapter 1 Introduction

Speaking in front of a bigger audience is a situation that most people face at least once in their educational or professional careers. Especially in the teaching profession, speech is an important tool that is used throughout the day and must counteract many disruptive factors. Through this setting, teachers have a higher risk of developing voice disorders like hoarseness.

This topic is nothing new, but teacher training deals little or not at all with the use of the voice. Several reasons for the burden of teachers' voices have already been found and discussed, but even with recognized disorders, many teachers do not begin treatment. One possible treatment is training the voice to use a tone that puts the least strain on the voice and manages the volume and way of talking in general. To first investigate the cause of a disorder, a voice screening can be done, which is a professional examination of the voice.

1.1. The Burden of the Voice of Teachers

Impact and Causes of Voice Disorders on Teachers

To get a better estimation of the extent and impact of voice disorders on teachers, several studies investigated the situation of teachers in this manner. From this, the necessity of proper training for the usage of the voice should be underlined. International studies show that teachers around the globe experience problems of hoarseness, in a range from 27% to 75% (27% Saudi Arabia[AAA18], 39.6% in Brazil [KAAA15], 42% in Nigeria[ALAS14], 58% in Germany[NSE$^+$20], 66.7% in Latvia [Tri17], 75% in Hong Kong (Tang via [CM19])). Disrupting factors that might lead to a voice disorders of teachers found in studies are age [Tri17], overload [ALAS14], [Wei07], malformation of the larynx, benign lesions such as nodules or polyps or edema or scarring on the vocal folds[Wei07], [MMB17], years of teaching [ALAS14], [KAAA15], [Tri17], gender [ALAS14], [MMB17], [Tri17], [NSE$^+$20], environmental factors (noise[ALAS14], [dGF19], [MMB17], air pollution [ALAS14], [Tri17], [KAAA15], [dGF19]), stress and anxiety [KAAA15], [MMB17], number of pupils in one class [MMB17], [dGF19], [AAA18], voice technique [MMB17], subject area [Tri17], smoking [AAA18], number of hours of teaching per day [dGF19], and type of school [NSE$^+$20].

Differences in Findings

These factors are not consistent in all studies. For example, Akinbode et al. [ALAS14] and Alrahim et al. [AAA18] found no association between voice disorders and years of teaching. Trinite [Tri17] observed a high prevalence of voice problems among music and sports teachers, but not a statistically significant difference among various subjects. No distinction between genders was found by Alrahim et al.[AAA18].

Some studies claim that fewer men report having or had problems with their voice compared to women (cf. [KAAA15]). One reason for the gender factor could be the smaller size of their larynx and their lower vocal power, according to de Sousa et al. [dGF19].

In contrast, Nusseck et al. [NSE^{+}20] established that fewer female teachers have physiologic voice problems compared to male teachers. They point out men recognize a change in voice rarer as a potential problem than women. Additionally, Nusseck et al. [NSE^{+}20] state the hypothesis that younger teachers have a less general awareness of their voice. They derive that by an increase in both age group and emotional stress and voice changes. In general, Korn et al. [KAAA15] found only 32.5% of patients with hoarseness are seeking medical advice. They underline their finding with reports from Da Costa et al. [DCPRC12] of 32.6% of elementary teachers looking for medical care and another outcome from Roy et al. with only 14.3% of elementary and secondary school teachers searching for medical advice because of voice complaints.

Apart from disruptive factors that cannot be influenced by a teacher and their classroom management to control the noise of the pupils, the voice of the teacher can be trained to better withstand these challenges. To first get an impression of a teacher's voice, what it is capable of, and where it lacks some training, a speech analysis can be done.

Stain on the Voice through Inconvenient Usage

"Voice disorders are mainly characterized by a persistent disturbance in the sound of the voice. [...] voice disorders are also defined by limited laryngeal efficiency or vocal endurance and sensations of laryngeal discomfort." [AWNN20]. Voice disorders have an influence on our voice and speech in the sense of voice quality or hoarseness, pitch, or loudness. (cf. [AWNN20], [Roy03]). This is especially difficult for people in job domains with a high share of speech, like teachers. Voice disorders or vocal dysfunctions can be caused or amplified by vocal abuse and misuse, deterioration of general health, environmental factors such as dry air, dust, classroom acoustics, excessive background noise, psycho-emotional factors, and stress, lack of vocal education and training, length of teaching service (cf. [vCWv12]).

Consequences for Teachers and Students

One aspect to keep in mind is that a restricted voice might have different consequences for both teachers and pupils. It is to say that not only teachers are influenced when they suffer from dysphonia, but also pupils. In a paper by Chui and Ma [CM19] they investigate the influence on pupils of teachers with dysphonia. They found a significant decrease in students' performance when they listened to a recording of a person with dysphonic voices. This is supported by outcomes from Nusseck et al. [NSE^{+}20]. Interesting about this observation is, that they did not find a significant difference in the influence of the performance when listening to a recording of a person with mild or severe dysphonia. One possible explanation Chui and Ma state, is that listening to a person with a dysphonic voice might be more difficult to process which takes the cognitive load needed to process the understood information.

Influnce of Dysphonia on Different Languages

Another aspect that should be addressed, is the impact of different languages. Chui and Ma [CM19] conducted a study on whether there is a significant difference in the performance of pupils who are listening to English speaking vs. Cantonese speaking persons without dysphonia vs. mild vs. severe dysphonia. The idea was to examine a language with different segmental or suprasegmental features. English is a stress or accent language, where syllables are emphasized. In contrast, Cantonese is a lexical tonal language, which uses pitch to distinguish lexical or grammatical meaning. They did not find a significant distinction between English and Cantonese, but an important limitation was, that they only had Cantonese natives with English as a second language. If there is a possible influence on the type of language, e.g. stress vs. lexical tonal, future research needs to be carried out. As German is similar to English, it is a stress-timed language, which means that its rhythm relies on the regular occurrence of stressed syllables rather than on a consistent timing between syllables like in syllable-timed languages. Hence, stress is not as relevant as in the English language, but the result could be partially generalized to German.

Influnce of Dysphonia on Quality of Life and Job Satisfaction

Voice complaints do not only influence the professional aspect of a teacher, but some findings also state the impact on the quality of life in general or job satisfaction. This is established by Nusseck et al. [NSE^{+}20] on findings by Lu D et al., Martinello et al., and, Hummel C et al. In their study, they found that teachers in Germany miss 24% of their workdays which are 1.2 days per year due to voice problems. A study by Roy et al., mentioned by Chui and Ma [CM19], conducted a telephone questionnaire with 2,400 teachers and non-teachers in the United States. They state that teachers could be more likely to have reduced job performance, miss more work days, and consider quitting jobs due to voice problems.

With all this at hand, educating people on the matter of proper voice usage, especially teachers, could improve teaching and to some part the personal life of teachers. This is supported by Chui and Ma [CM19], who state that with the correct use of voice and vocal hygiene, teachers have a lower risk of developing a voice problem.

1.2. Voice Screening

To investigate the correct usage of the voice or potential disorders, a voice screening can be done. Voice screening is an examination used to assess various aspects of a person's vocal abilities and characteristics. It is commonly utilized in fields such as speech therapy, singing, acting, and broadcasting.

Purpose of Voice Screening

The purpose of voice screening depends on the context but generally includes evaluating parameters like pitch, volume, tone, clarity, resonance, and overall vocal health. For the health aspect, parameters like jitter and shimmer, roughness and other irregularities are of importance. In the context of speech therapy, voice screening may be used to identify any disorders in an individual's voice production, such as hoarseness, vocal nodules, or vocal cord dysfunction. This can help speech therapists tailor treatment plans to address specific vocal issues.

In singing, voice screening can be used to assess a singer's vocal range, flexibility, and technique. It may also help identify any vocal habits or tendencies that could potentially lead to strain or injury. For this purpose, several tests utilize singing as the basic data source for investigations aimed at constructing a voice range profile. In acting and broadcasting, voice screening is often used to evaluate an individual's speaking voice for clarity, expressiveness, and suitability for different roles or broadcasting purposes. This may involve exercises to improve articulation, diction, and vocal projection. Similarly to these two application areas, a voice screening is offered at the RWTH Aachen University.

Voice Screening for Teachers-To-Be

The Chair for Contemporary German Language [Sti] conducts this screening and it is a mandatory event for teachers-to-be at the RWTH Aachen. They use the software lingWaves by WEVOSYS [WEV] and perform the vocal load test with the participants. To do so, the participants have to hold different vowels for a certain time or as long as possible and read the standardized text *Nordwind und Sonne*. Next to the assessment of the voice, workshops take place to train the participants in the basics of voice and speech training. The complete goal is to get teachers-to-be in first touch with their voice, a short professional assessment of their usage of the voice, and how they can train to get resilient and persistent voice function. Similar offers exist at other universities, for example at the Regensburg University [Anga].

Overall, voice screening provides valuable insights into an individual's vocal capabilities and can be used to inform various aspects of vocal training, therapy, or performance. Different measures are used, based on singing, holding a tone or vowel, reading standardized texts, interviews, and self-reporting.

1.3. Teach-R

Teach-R [Mmv23] is a VR application developed and in constant development by a team of different universities in Germany. The main idea of the application is for teachers-to-be to train their classroom management and especially build up knowledge on how to deal with disruptions in the VR classroom. To do so, the users are placed in a virtual school with different kinds of classrooms where they can conduct lessons.

Application of Teach-R

One big advantage of a simulation is the possibility of creating repeatable or rare situations to practice how to deal with this kind of situation. Nevertheless, it does not substitute for the real environment. Instead, it could be considered more of a helpful addition (cf. [HS23]). In more detail, users can train discussions and presentations in class and how to deal with disruptive behavior in a classroom mostly caused by students (cf. wiki of Git project [Anw23]). For instance, some disruptive behaviors are students talking, throwing paper balls, or more specific to a subject, e.g. wrong handling of equipment in the chemistry classroom. Several rooms are already implemented: some basic classrooms with different seating plans, a chemistry room, a media lab, and a foreign language classroom (see figure 1.1). Those rooms have theme-specific features like computers, whiteboards, and chemistry equipment. A faculty room is implemented as well, which serves as a tutorial and is currently reworked.

The Coach for Classroom Control

To control the general setting in a classroom, a desktop application has been created, named Coach (see figure 1.2). In the coach, the behavior of the students can be set, either through selecting a behavior and applying it to students or by creating a scenario. Furthermore, options like weather, volume, and lesson schedule can be changed through the Coach in the VR application.

Learning Analytics for Teach-R

In the summer term 2023 a group of students added learning analytics to the application using OmiLAXR [GHS23], an ecosystem for the integration of learning analytics in virtual reality. With the tracked data, three visualizations to reflect and evaluate a session were integrated. The visualizations are a heat map for position data, one visualization for showing the position over time and rotation of the user, and one representing the focus of the eyes on students via eye-tracking.

As Heinemann and Schroeder already mentioned in their paper [HS23], integrating learning analytics with different visualizations might be helpful to teachers-to-be to review and reflect on their training. Using the data from the learning analytics combined with the feedback from others, either the coach or other participants watching, the user can benefit from many different aspects (c.f [HS23]). Additionally, the more diverse data the learning analytics represent, the more precise an evaluation can be done and might reveal unconsidered actions. To link up with the integrated learning analytics, the goal of this thesis is to extend the range of learning analytics in Teach-R by adding an analysis of voice to it. More on this can be found in section 1.4.

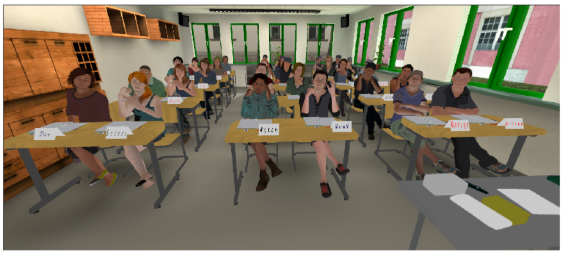


Figure 1.1.: A screenshot from Teach-R. The frontal-lecture classroom. The view is from the teacher's desk onto the pupils at their desks.

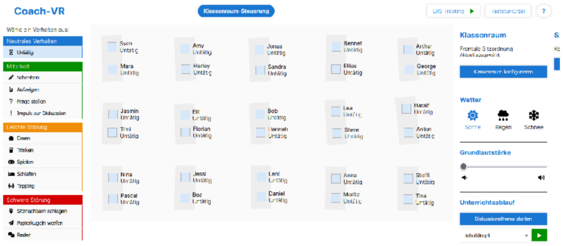


Figure 1.2.: A screenshot from the coach. The frontal-lecture classroom. On the left are the different behaviors in categories. On the right different options to manipulate the active classroom simulation, like volume, weather, schedule, and scenarios.

1.4. Goal and Structure of this Thesis

The goal of this thesis is to implement a foundation for speech analysis into Teach-R [WHLS21](former VR-Classroom), a simulation of an interactive classroom for teachers-to-be to work on their classroom management. This analysis should build upon existing voice screenings, for example, the one offered at the RWTH Aachen University. In contrast to a professional voice screening, the analyses should not only serve as a possibility to reflect on the proper usage of the voice but also investigate the usage of the teacher's voice in a realistic environment. In more detail, the possibility of reflecting and evaluating the session during and afterward should be given. Not only aspects like the vocal range, but adapting the volume to the classroom, speaking comprehensively, and what is spoken can be of interest.
However, since the users are not supposed to actively participate in the analysis, only freely spoken text will be used for the analysis. A further limitation is the microphone of the head-mounted display used for Teach-R, as well as the given possibilities from Unity, in which Teach-R is built, and the scope restricted by the time of this thesis. This cuts the aspired overall implementation goal to the following parts:

- Managing recording in Unity via the Coach or in Teach-R
- Some basic analysis on recorded audio in VR and in the Coach
  - Volume
  - Speech rate
  - Pronunciation
  - Pitch
  - Neutral speaking pitch
  - Vocal range
- Speech to text
- 3D visualizations in Teach-R
- 2D visualizations in the Coach (an additional page in the coach)

The final product in Teach-R should provide the users with some visualizations showing the speech behavior of their session, in a two- and three-dimensional manner. Furthermore, the user should have the option to *revisit* their school lesson, i.e. be able to create a clone with the recorded audio of the user, to experience their speech behavior. At last, speech-to-text can be used to implement a change of behavior through a user's verbal instructions by using keyword recognition.

Structure of this Thesis

To do so, in the related work chapter 2, the very basis of phonetics and the emergence of the voice is discussed. This is not only intended to serve as a foundation of this thesis but also for the further development of speech analysis in Teach-R. The full spectrum of speech analysis cannot be dealt with in one master thesis, especially with the starting conditions given by Unity.

In the chapter Conceptualization 3, first the restrictions and challenges are presented. Additionally, the added features are planned for both, Teach-R and the coach.

Which of the intended features were implemented and how they are realized is discussed in chapter 4.

To get feedback on the implementation result concerning functionality, usability, and helpfulness, an evaluation took place. A detailed methodology and the results are written in chapter 5. What can be concluded from the evaluation is also critically weighed up in this section.

Chapter 6 contains an overall review of the thesis. Next to a discussion and conclusion, it explains the limitations and future work.

This thesis is done partly in cooperation with Jasmin Hartanto and her master's thesis "Evaluating Classroom Management in VR using Multi-Modal Learning Analytics". The cooperation is based on the joint participation in the research focus class in 2023. Together, the visualization menu was redesigned and general mutual consultation and research was done.

Chapter 2 Related Work

Structure of this Chapter

To implement voice screening into Teach-R, basic knowledge is gathered to establish a foundation backed by research. This is done by firstly finding parallels and differences between voice analysis and speech analysis. Speech analysis was added to this project because both areas partly use the same data source and speech analysis can further enhance the feedback by looking at prosodic aspects. Subsequently, both types of analysis are discussed in the following sections. A short excursion to the topic of mood or emotion analysis is made, as it is an active topic in research. Lastly, existing approaches and further aspects that ought to be taken into account are debated in the last section.

2.1. Speech Science Versus Speech Analysis

Speech science and the analysis of speech are two distinct, but related fields within the study of verbal communication. While both disciplines have several similarities in understanding how humans convey meaning through sound, they differ in their specific focuses, methodologies, and objectives, as discussed later.

Speech/Voice Science

Speech science, on the one hand, is the study of how speech is produced, transmitted, and received or perceived. One part of it is the analysis of the voice in the area of phonetics, which is the study of what happens physically, i.e. when humans talk and how sounds are produced in the human body (cf. [Mac23]). Among others, it focuses on examining the qualities and characteristics of an individual's voice, such as pitch, tone, intonation, rhythm, resonance, and timbre. Furthermore, it can be used to investigate changes in the voice that can be caused by disorders (cf. [MSD14], [BH97]). It is often applied in areas, such as voice coaching, acting, and singing as well as scientific approaches like acoustic analysis and spectrography (cf. [BDB15]). The primary objective of voice analysis varies depending on the context. Voice analysis may be utilized in fields such as acting to convey a character's emotions convincingly, singing to enhance vocal performance, in therapy to address vocal disorders or in investigations for voice identification. For teachers, the medical aspect is important, as their voice is the foundation of their profession. Also, emotions seem to play their part as teachers are contact persons for their students as discussed in section 1.1.

Speech Analysis

Speech science and speech analysis overlap, as they involve understanding how vocal qualities influence persuasion, emotion and rhetorical effectiveness. The analysis of speech, on the other hand, focuses on linguistic elements, including phonetics, phonology, syntax, semantics, and pragmatics. Aristotle created the area of rhetoric and defined it as "the faculty of observing the available means of persuasion, in any given situation."[AFA06]. It evolved to explore more than speech communication, but speech, representation, and power used to persuade (cf. [Hal22]). The objectives of speech analysis typically revolve around comprehending language structure and usage, such as studying language acquisition, determining dialectal variations, analyzing discourse patterns, investigating linguistic phenomena or developing speech recognition systems. Although both fields have a certain overlap, they employ different methodologies for their specific focus and objectives. In summary, speech science and speech analysis are complementary disciplines within the broader study of verbal communication, each contributing different insights to how humans communicate through sound. For the sake of simpler distinction, speech science is called voice analysis in this thesis.

2.2. Creation of Sounds and Speech

To first understand which types of acoustic measures exist, a further introduction to the emergence of the voice ought to be. This subchapter is based on the textbook by Seidner and Wendler [SW97]. If not stated otherwise, the information is taken from this book.

Sound Transmitted through Sonic Waves

The voice is a sonic phenomenon that is based on oscillations of particles. These oscillations can be perceived by the sense of hearing over the eardrum through a series of pressure differences. Those pressure differences need to happen between 20 to 20,000 times a second to be audible for a human. Additionally, the speed of sonic waves in the air at a temperature of 20° Celsius and normal pressure of 344 m/s can vary in length between 2cm to 5m.

Important characteristics of Sound

Four important characteristics of the sonic phenomenon are **frequency** (subjective pitch), **sound pressure** (subjective volume), **spectrum** (subjective timbre), and **duration**. The frequency is the number of oscillations per second and is measured in Hertz (1Hz = 1 oscillation per second). As a physical measure of sound pressure, the relative scale of Decibel (tenth of a Bel) is used. It is an auxiliary measurement unit of the common logarithm to better compare the subjective volume. Finally, the spectrum describes the range of frequencies that compose a sound. A pure tone is a periodic oscillation pattern, e.g. a sine wave, but the most common sonic phenomenons are combined and form complex oscillation patterns.

Creating of Sound in the Human Body

To create a tone, an interaction at the glottis, between the muscles in the larynx and breath pressure, or airflow during exhalation, takes effect. First, the vocal folds are closed. The pressure and the muscles open the vocal folds and lead to an oscillation of the vocal folds determined by the mass, length, and pressure of the vocal folds. The pressure is then released by opening the glottis, resulting in a pressure drop below and, consequently, a closing of the vocal folds.

This is a very brief and simplified description of the complex process but should suffice to get a basic understanding for this thesis. For visual insight, figure 2.1 and 2.2 show the larynx in some detail. Figure 2.1 displays a picture taken near the epiglottis showing the vocal cords and further parts of the larynx. For a better overview, figure 2.2 shows an illustration with more exact locations as a vertical cut-through of the larynx. Not only the larynx is relevant for creating tones, but also the articulators and the lungs, for creating air pressure. The articulators lie above the larynx and consist of the tongue, cheek, lips, et cetera. Its purpose is to articulate and filter the resulting sounds from the vocal cords.

Changing Pitch

Changing Volume

The pitch of the voice can be changed by varying the tension of the vocal cords. To raise the pitch, the tension must be increased and vice versa but also the air pressure is involved. To modify the tension of the vocal cords, the thyroid cartilage is modified by a group of muscles, for example, the vocalis muscle (cf. figure 2.1, 2.2). A change in the volume is archived mainly by the amount and speed of the air, i.e. the exhalation pressure. When increasing the volume, the vocal cords are brought closer together, leading to a higher frequency. To decrease, the opposite happens: the tension is reduced and

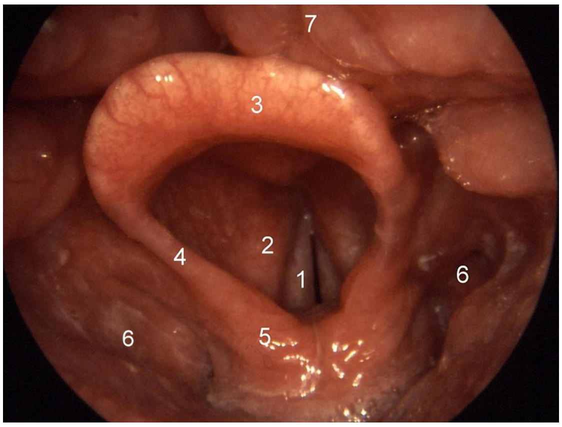


Figure 2.1.: A foto of the larynx with numbers on each part, namely 1 = true vocal cords, 2 = false vocal cords, 3 = epiglottis, 4 = aryepiglottic folds, 5 = arytenoid cartilage, 6 = pyriform sinus, and 7 = base of the tongue. - I, Welleschik, CC BY-SA 3.0 via Wikimedia Commons [1]

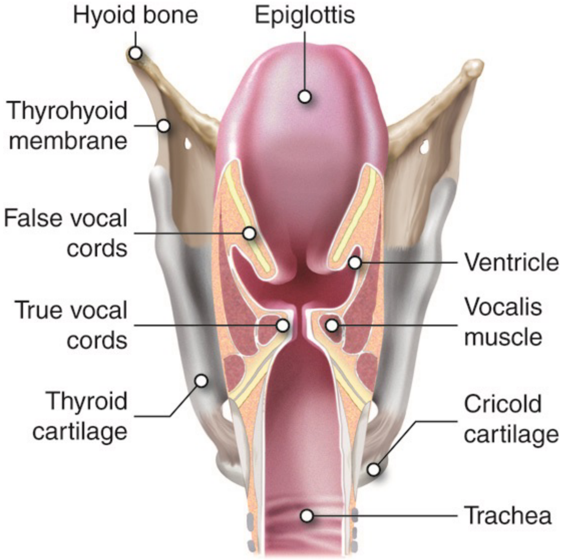


Figure 2.2.: An illustration of the larynx and the vocal folds with descriptions of each part, namely the epiglottis, the hyoid bone, thyrohyoid membrane, false vocal cords, true vocal cords, thyroid cartilage, trachea, cricoid cartilage, vocalis muscle, and ventricle (beginning at the top and counterclockwise). - Cenveo, CC BY 4.0 via Wikimedia Commons [2]

Changing Timbre

the vocal cords may not be fully closed. A lot of different factors influence the timbre of the voice. Some of the potential factors are the state around the glottis, the length, and thickness of the vocal cords, and the condition of the mucous membrane. Without the influence of the vocal tract, the difference between the sounds directly from the glottis compared to what sounds leave through the mouth, the audible result is not comprehensible because they have no space to spread. The vocal tract alters barely during the phonation and, therefore, impacts the individuality of the voice. It acts as a resonator to amplify certain frequencies by directing the sound's sonic waves through different chambers. The size and shape of these chambers can be modified using movements of the tongue, lips, and jaw.

2.3. Analysis of the Voice

Overview of Methods

There are a lot of different methods that are meant to evaluate various areas, i.e. visual analysis, perceptual evaluation, aerodynamic measures, acoustic analysis and self-evaluation. Barsties and De Bodt [BDB15] conducted a meta-analysis on the state-of-the-art assessment of voice quality in 2015. Objective acoustic analysis describes the assessment of the sound via automized methods, usually on sustained vowels, focused on in this thesis. They found by investigating the individual measurements that reliability and validity are rather poor. However, they claim that a combination of parameters could

[1] https://commons.wikimedia.org/wiki/File:Larynx_normal.jpg (accessed 14.11.2023)

[2] https://commons.wikimedia.org/wiki/File:Anatomytool_larynx_and_vocal_cords_English.jpg (accessed 14.11.2023)

improve these. In order to diagnose and check on the voice for medical purposes, several methods are recommended in Europe and the USA, namely the voice range profile, irregularity measurements, sonograms or spectrograms, and combined parameters for classifying voice disorders (cf. [BB14]).

Voice Range Profile

A voice range profile describes the range of volume and pitch when speaking or singing. This can be measured by a phonetogram which maps the result to a graph with the fundamental frequency in semitones at the x-axis and the sound pressure level in decibels at the y-axis. It serves as a common frame of reference for various voice measurements and can be used in clinical, pedagogical, and research settings to analyze and understand voice production (cf. [TPS16]). To do so, the participant produces sustained phonations at different pitch and loudness levels [BB14]. As presented in [TPS16], the voice range profile involves several limitations. The variability is an important factor according to Ternstöm et al., as voice production is based on large individual variations and summarizing all aspects of voice production into a single image is very difficult. It focuses mainly on fundamental frequency and sound pressure level which do not capture all aspects. Therefore, further measurements might be necessary for a comprehensive assessment. Finally, the interpretation of the visual results requires prior knowledge in the field and can depend on the model used to interpret the results, for example, different statistical models or voice pathology models.

Irregularity Measurements

Irregularity measurements investigate mainly the jitter and the shimmer. The jitter reflects the fluctuations of pith between two acoustic oscillations, while shimmer assesses the deviation between the volume between two acoustic oscillations (cf. [BB14]). A common way to measure the irregularities, is by requesting patients to hold the vowel /a/. Fourcin [Fou09] discussed the possibility of assessing voice irregularities, e.g. jitter, by studying healthy voices and examining how vocal cords move during speech. From this, abnormal voices can be better identified and clinicians can objectively assess voice irregularity and quality. This measure is limited by the dependency on the correctly detected fundamental frequency and the volume. To rightfully identify those parameters, a healthy or slightly hoarse voice can be used [BB14].

Sonograms and Spectrograms

Sonograms or spectrograms are used to represent the complete sound spectrum of a voice, i.e. the fundamental frequencies and the corresponding overtones. Information about the resonance function can be derived from this, as the overtone spectrum is influenced by the size and shape of the vocal tract [BB14]. To be more precise, sonograms or spectrograms provide a visual depiction of the various components of a voice, including fundamental frequencies (pitch) and harmonics (overtones), allowing for detailed analysis of vocal characteristics, such as pitch range, timbre, and resonance. Spectrograms represent sound in a detailed manner, which allows the creation of sound based on a spectrogram (cf. [Ana]).

Combined Parameters

For combined parameters, Brockmanmn-Bauser and Bohlender mentioned the Dysphonia Severity Index. It considers the highest frequency, lowest intensity, maximum phonation time, and jitter (cf. [WBM+00]). With a weighted combination of these parameters, the quality of the voice can be assessed, i.e. having a low score of the DSI can be regarded as having a dysphonic voice. The parameters are calculated from acoustic analyses of sustained vowel phonations.

Examples of Tools

To conduct the introduced methods, several software tools exist, both for a fee and free of charge. The Chair of Contemporary Art uses lingWAVES by WEVOSYS [WEV]. Besides acoustic assessment, they offer software for endoscopy, swallowing diagnostics, high-speed video endoscopy, nasality analysis, velopharyngeal opening, and electroglottography. For the acoustic assessment, methods such as voice protocol, standard voice analysis, vospector, or the Goettingen Hoarseness diagram are available. The aforementioned methods can assess parameters like the s/z ratio, maximum phonation time, fundamental frequency, loudness, voice quality, jitter, shimmer, and breathiness. For an example of a free product, Praat [Pra] an open-source software developed by Paul Boersma and David Weenink at the Institute of Phonetic Sciences at the Univerity of Amsterdam. It offers analyses like spectrograms, pitch, format, irregularities, or intensity, but can synthesize speech as well, for example from pitch, formant, and intensity.

2.3.1. Parameters Used in Voice Screening

After comparing which parameters are used by different methods, this section investigates into what these parameters are and how they are measured. For this, the Master Handbook of Acoustics [EP22] gives an overview of physics. If not stated otherwise, the information presented is from that book.

Propagation of Sound

Sound needs an elastic medium to propagate, for example, through gases, liquids or solids. If a sound is produced in a vacuum, nothing can be heard. It can travel through obstacles like solids, but will lose intensity or is entirely absorbed [Ash21]. The waves spread out evenly in all directions from the source until they decay if they are not interrupted by obstacles. Without obstacles, the sound intensity decreases as the square of the radius, while the sound-pressure level decreases by 6 dB whenever the distance is doubled. Sound can be represented as amplitudes or energy, e.g. when the distance between the peak of the upper half of the wave and the trough separating the waves increases, loudness increases (cf. [Ash21]). Furthermore, the amplitude is described by frequency and the number of oscillations, at which the waveform repetitions per second, whereas the intensity describes the length of the amplitude.

Compared to a simple sine wave, speech and music build complex waveforms. These waveforms are comprised of several combined simple sine waves, which can be separated with Fourier transformation. Two similar measures of sound are the sound pressure level and the sound intensity. While the sound pressure level focuses on the pressure variants caused by the sound waves (expressed in dB), the sound intensity reflects the actual energy transferred (expressed in W/m^2).

Fundamental Frequency

Going over the parameters often used in acoustic analysis, a short overview of what those parameters are and what they state is presented. Firstly, the fundamental frequency or F0, describing the first harmonic is widely used to characterize the human voice. Christina Oliveira et al.[COGM21] state that, physiologically, the fundamental frequency is determined by the number of glottal cycles performed by the vocal folds per second, which is influenced by their length. Longer vocal folds result in faster glottic cycles and a higher fundamental frequency. The analysis requires defining the recording task, as vocal signals exhibit differences in sound energy between the initial and final portions of vowel stretches. Sustained vowels are preferred for analysis due to easier control and standardization of phonation production. The fundamental frequency can be extracted from the list of frequencies by applying a Fast Fourier Transformation and magnitude spectrum analysis (cf. [DKI23]).

Jitter and Shimmer

Jitter and shimmer are values of irregularities in the sound. Jitter is a measure of the variation in the period length of the fundamental frequency (cf. [BRZ+22]). Shimmer, on the other hand, represents the variation in the amplitude of the fundamental frequency (cf. [BRZ+22]).

Harmonic to Noise Ratio Harmonic to noise ratio (HNR), specifies the relationship between the harmonic part vs. noise, the periodic component, and the aperiodic component (cf. [FTG+18]). It is expressed in decibels (dB) and reflects the general frequency content of the speech. Different vocal tract configurations result in variations in HNR values, with vowels containing higher frequency content having lower HNR values compared to those with lower frequency content. HNR is sensitive to differences in vocal tract resonance, making it useful for assessing vocal quality and speech clarity in various applications such as speech pathology and voice recognition technology.

Maximum Phonation Time Maximum phonation time measures the longest duration a person can sustain a vowel sound (cf. [SBP+10]). From this, information can be derived about the overall vocal quality, pitch, steadiness of sustained tone, and loudness [SM77]. It is considered an objective measure by measuring it on a sustained vowel after a maximal inspiration (Neiman and Edson via [SBP+10]). As this measurement is criticized in validity, [SBP+10] found a significant increase in the validity of the method was done repeatedly and taking the mean result on the same day (0.836), or higher on two different days (0.911).

Speech Rate Speech rate or speaking rate is the speed of speech referring to produced words or syllables. It often is measured in words per minute or syllables per minute, but also considers the pauses between words (cf. [SBCWI18]). The average speech rate strongly varies due to different factors, for example, language, dialect, and gender. Matsuura et al (via [SBCWI18]) provide some values for these factors, for example, an American English male speaker has on average a speech rate of 245 syllables per minute, while a female speaker has on average 254 syllables per minute.

Voice Range As the name suggests, the voice range represents the range of the tones a voice can produce or in other words the maximal voice capacity in terms of sound pressure level and fundamental frequency. It is often measured with a voice range profile as presented in section 2.3. The participant is asked to hold phonations at different pitch and loudness levels, for example, as it was tested i.e. in a study by Sanchez et al [SODH14].

Volume The volume of the voice can be derived from the intensity of the sound wave. It is usually measured from the lowest to the highest volume a person can produce (cf. [BB14]).

Pitch As a subjective measure, pitch is not a linearly related function of frequency [EP22]. For example, a 1-kHz signal can have a different pitch if the level is increased depending on the sound pressure. It is dependent on the frequency and intensity, i.e. when the intensity rises, a low tone has a low frequency, while a high tone comes with a high frequency.

Timbre Timbre is the perception of the tonal quality of complex sounds (cf. [EP22]). It describes the difference in perception, for example, when two different instruments play the same pitch. The number and relative strength of the partials of a wave determines the timbre. Similar to frequency and pitch, timbre is the subjective counterpart of the spectrum.

2.3.2. Voice Analysis on Connected Speech

As the goal of this thesis is to implement voice and speech analysis into Teach-R, the feasibility of it on free or connected speech solely needs to be evaluated. In research, several different phrases are used for freely spoken speech, for example, running or connected speech. All refer to speech that utilizes words for utterance but can differ by using standardized texts or just a topic. Barsties and De Bodt [BDB15] claim that objective-acoustic parameters can barely be utilized on connected speech. They state, that the application of common parameters must be separated into voice and non-voice segments before they can be used. Still, some approaches are based on free speech for voice analyses.

Airflow Patterns

The differences between airflow patterns during either sustained vowels, running speech, or phonation in running speech were investigated by Gilman et al. [GMH19]. They used a professional setup to assess the airflow patterns during sustained vowels and running speech and calculated the flow rate during phonation in running speech by only looking at the mean flow rate during phonation. For the sustained vowels on the one hand, the participants were asked to hold an /a/ for five to ten seconds at a comfortable pitch and volume. On the other hand, to get free speech samples, all participants should describe the same picture. Through this, they found significant differences between the overall mean airflow rates (P = 0.0152) and between the sustained vowel and phonation in running speech (P = 0.021). No significant difference in airflow rate was detected between the sustained vowel and running speech (P = 0.051) and between running speech and phonation in running speech (P = 0.94). They conclude that the variance should be considered, even if some correlation is present. This is an interesting finding, as they entirely focused on free speech and no standardized text, which is the same requirement for this thesis. Moreover, it can be used as a foundation for analyzing the impact of parameters on connected speech, which usually is done on sustained vowels. The results might differ, but the results could be given relative manner instead of absolute values and be used for giving a first impression to the user. For an overview of different approaches done in the last years, some research is presented in the following.

Frequencies, Vocal Range, and Speech Rate

In a study by Borrego et al. ([BGB07]) they investigated the development of the voice of a radio announcing course before and after the course. Based on those recordings, the experimenters analyzed among other things the mean, minimum, and maximum fundamental frequency, frequency range (Hertz and semitones) [above mentioned as vocal range], and sentence and pause duration [above mentioned as speech rate]. To accomplish this, they used Visi-PitchIII/Sono-Speech (Model 3900/3600) from Kay Elemetrics (Lincoln Park, NJ) that extracts values of time (seconds), energy (decibels), and pitch (Hertz) in real-time.

Intonation, Frequencies, Vocal Range

Another study that used the same tool to analyze connected speech was done by Medrado et al. ([MFB05]). They asked voice-overs to read some short sentences like “Brazil is worth this emotion!”. They examined similar parameters of speech, namely the total text length, the total pause length, the length of the three emphatic pauses selected for this analysis, values of the mean, minimum, and maximum fundamental frequency, and the semitones range. By taking these two findings together, possible parameters to investigate on running speech can be the **fundamental frequency, the vocal range or pitch, speech rate**.

Vocal Fold Frequency

Taking the capabilities of human sound processing into account, Fourcin [Fou09] investigated voice irregularity in connected speech. It is noteworthy that he not only mentions acoustic measurements in his paper but also discusses deriving the possibility of objective measuring irregularity in connected speech from the ability of the human to do so. More precisely, the vocal fold frequency can be determined by looking at normal versus “unaccaptable” irregularity in voice pitch. The vocal fold frequency provides insight into voice quality and laryngeal function. He concludes that connected speech material can be effectively used for the objective assessment of primary physical aspects of voice quality.

Frequencies

For a last approach, Fraile et al. [FGS$^+$13] compared spectral power distribution on sustained vowels and connected speech. Spectral analysis in this context refers to the process of examining the frequency content of a voice signal. One approach for this is to represent the frequencies in a spectrogram, as discussed before. This analysis was done on a database containing recordings of each participant holding a vowel and connected speech, here the first three sentences of the standardized text *Rainbow Passage*[Rai].

Their results match with previous findings that the relevance of spectral tilt better performs on running speech. Furthermore, dysphonic voices showed a higher increase in high-frequency power compared to normal voices during running speech, making high-frequency power a potential indicator of dysphonia in this context. On the contrary, the stability of spectral power was not as strong of an indicator of dysphonia in running speech compared to sustained vowels.

Summarizing these findings, analysis of connected speech might be more suitable for certain aspects but ought to be compared with existing analysis on sustained vowels. Another approach would be to investigate new methods for analyzing specifically connected speech or create synthetic data from connected speech for existing analysis on sustained vowels.

2.4. Speech Analysis

Prosody for Speech Analysis

As mentioned before, not only the analysis of voice can assist a teacher in their workday, but analyzing aspects of speech can further improve their teaching skills. For example, looking at the prosodic aspects of a teacher's presentation could help to convey information. McCabe states that prosody investigates how voice dynamics, pitch, speech rate, and timbre influence the meaning of what was said (cf. [McC17]). He expands that only looking at word meanings makes it difficult to understand the expression of emotion, social values, and semantics. One part of the study of prosody concentrates on the analysis of acoustic properties, like the fundamental frequency and duration patterns (cf. [Pro12]) similar to voice analysis discussed before.

Parameters for Teachers

Not only should the parameters be considered that are possible to realize, but also those that are interesting for the target audience of Teach-R, i.e. teachers-to-be. Additionally, next to the parameters examining the health and proper usage of the voice, parameters, and techniques to analyze the lesson from other perspectives might be interesting. The role of a teacher could be described as versatile and involves much more than just spoken words. However, reviewing voice and speech can support teachers in conducting lessons.

Reviewing different talk strategies, Sharpe [Sha08] investigated a held history lesson and combined with related work, specified some techniques teachers could use to convey content to students. She emphasizes that teachers can better promote learning through joint negotiation instead of merely transmitting information. Techniques discussed by her are repeating, recasting, recontextualizing, cued elicitation, questioning, contextualization of experiences, and recycling language. Repeating, recasting, and recontextualizing can be grouped as taking what a student has said and reformulating it. To enhance the statement, the teacher can change it to use more technical terms or set it in a technical context while the key message remains the same. Thus, teachers can acknowledge the statement and involve students more in the problem-solving process. This could be analyzed by examining the speech-to-text results and comparing them with statements from the students. Furthermore, an emphasis could be placed on the differences in words used with the same message, i.e. whether technical terms were used. As another technique mentioned, cued elicitation is applied to provide the students with enough information to develop a discourse but leave gaps to be filled by this discourse. With a similar aim as the techniques before, the completeness of the information can be checked, and whether it is automatically assessable. Questions as a technique can be an essential tool for teachers. This is underlined by the rate of questions posed by teachers of two questions a minute (Edwards and Mercer via [Sha08]). The frequency of questions could be analyzed by the pronunciation and intonation or pauses of the teachers' speech. Moreover,

Repeating, Recasting, Recontextualizing

Cued Elictation

IRF Pattern

an initiation-response-feedback pattern could be examined by a combination of semantic and syntactic analyses, i.e. if the teacher gives feedback to the response of a student. Automatically assessing for contextualization of the experiences of the students is probably not feasible in Teach-R, as the teacher cannot draw on the memories of the students. As the final technique mentioned in [Sha08], recycling words or phrases can help to bring structure to the lesson. For example, in the lesson discussed in the paper, some words are used throughout the lesson while others start appearing in later episodes or are used less often or not at all. This could be automatically analyzed by looking at the appearances of frequently used words and organize in a timeline. Consequently, approaches for analyzing techniques for teachers' talk can be developed and could give the users feedback on the transfer of information of the lesson.

Types of Questions

The technique of the initiation-response-feedback pattern is discussed in more detail by Nassaji and Wells [NW00]. First, different types of questions are presented. These include known information questions, negotatory questions, opinion and conjecture questions, clarification questions, and evaluation questions. As every type has a different aim, the structure of the question and the semantics could be analyzed to create a count of question types used.

Pronunciation

Looking closer at the pronunciation of teachers, the intelligibility of the spoken words can be evaluated. Analyzing and assessing pronunciation are often used in the area of learning languages, for example, in the paper by Tejedor-Garcia et al. ([TEC+20]), in which they aim to test the effectiveness of a computer-assisted pronunciation training tool in improving the pronunciation skills of adult native Spanish speakers in the production of a set of difficult English sounds. One idea to assess pronunciation accuracy is to compare the actual spoken text to the speech-to-text result, as done in this study. A difficulty with this is that the coach, controlling the Teach-R environment, would have to constantly listen and understand what the user is saying and mark incorrect words interpreted by the tool to have the original spoken text for comparison. Software that offers autonomous pronunciation assessment already exists, for example, the Microsoft Azure SDK [eri23], but software already applicable in Unity and open-source could not be found. Another approach would be to look at the tolerance range of the speech-to-text tool, and whether it is possible to derive statements from the size of the range and the probability of the word being correct.

Considerations for Speech Analysis

As mentioned before, the goal of these analyses is to help the users reflect on their lessons, not to rate them, as it is not possible to decide on what is good and what is bad in an automatic manner. Furthermore, some aspects should be considered when looking at the visualizations of the analyses. Speakers have different vocabularies, experiences, and personal backgrounds. The analyses can be influenced by this, for example, a dialect can lead to some words being used more often. However, in a purely automatic assessment, detecting dialects is more difficult, as a tool needs to be trained on different data sets with dialects or accents. The influence on the automatic assessment of gender, social background, and native language along with others ought to be observed and incorporated into the analyses if necessary.

2.5. Moodanalysis

Another factor that influences the voice can be the mood or emotions of a person in certain situations or social positions, for example, a press conference or call center. Emotions can be conveyed with the voice by the content of what is said, as well as linguistic and paralinguistic signals, so how it is said (cf. [AGDM20]). This is a current topic of research and takes place in many analyses, for example, [IHC+23]. On the other hand, Sutton and Wheatley [SW03] found a relationship between the emotional state of a teacher and their teaching. To be more precise, they did a meta-analysis on this topic and found that teachers' emotions influence the cognition, motivation, and behavior of both, teachers and students.

audEERING

"Die Stimme der Autorität ist nicht die Stimme der Hilflosigkeit."[The voice of the authority is not the voice of the Helplessness] ([SW97], p. 157). However, the complexity of how the voice changes due to certain circumstances is difficult to analyze on a psychological, physiological, and physical level. It is ongoing research to analyze voice in a conversation to automatically assess the mood of a person. The company audEERING [Inn21] works on technology to detect emotions and health information. For example, they have a product called devAIce to analyze emotions that can differentiate between emotion classes like neutral, happy, angry, and sad, or emotion dimensions, i.e. activation, valence, and dominance. Furthermore, it can differentiate between background noises like music and human voices and detect the speaker's age and gender. Another example of a development with a different focus is the COVID-19 Test. They investigate the development of voice biomarkers, i.e. a change in the voice through affected respiratory system and muscles, and use over 6,000 parameters to detect subtle differences. In an ongoing study (date 14.12.2023) they collect voice recordings to further improve their AI model and to develop an app afterward.

Analysis on Fundamental Frequency

Focusing on the fundamental frequency, Dimitrova-Grekow et al. [DKI23] did a speech emotion recognition on a database of artists and non-artists. They state, that many different parameters are used in research, for example, basic tones, formats, intensities, energy, speech rate, voice quality, and vibrations. To investigate the relationship of the fundamental frequency with emotions, they applied different AI models to features of the fundamental frequency. With this approach, they archive an accuracy of 89.74% when differentiating between two emotions, 76.14% between three, and 62.99% for four emotions. Finally, they state that fundamental frequency is a solid basis for emotional analysis, but further research needs to be done.

Analysis on Voice Quality

Patel et al. [PSSB10] analyzed a sustained /a/ of actors with different emotional states, similar to before. Instead of using the fundamental frequency, intensity, and duration, they focused on voice quality. They further state, that depending on the emotional state, the prosodic patterns change. For the analysis, they focused on the emotions of relief, sadness, joy, panic fear, and hot anger, as they have strong differences in the context of arousal, power, and valence. Based on their analysis of sustained vowels, they conclude that voice quality is an important aspect of the analysis of emotions.

Validity and Reliability

Additionally, they state, that acoustic markers have not been found for every emotion to properly differentiate between them. Different cues based on power and valence dimensions are found but are inconsistent. The analyses of the app are further limited to the quality of the audio, i.e. whether noise is present or the speaker increases their speaking loudness.

A critical perspective is taken by Arana et al. [AGDM20]. They evaluated the efficiency and reliability of a mobile application for emotion detection. On one hand, they agree that the app can identify dimensions of emotions, i.e. positive or negative emotions. On the other, the application is not able to properly identify categories of emotions, like anger or sadness. In more detail, in scenes of happiness or sadness, the application recognized, for example, the emotion of anger.

Taking this into account, mood and emotional analysis based on speech can give an intuition on the emotional state of the speaker, but it would require refinement for detecting individual emotions. Further research should focus on the potential impact of emotions on teaching and how accurately they can be assessed automatically. At this point, emotion and mood detection have comparatively less informative value than aspects of voice and speech analysis discussed in this thesis. Therefore, it will not be included in the analysis in this thesis.

2.6. Voice and Speech Analysis in Teach-R

To conclude this chapter, the aspects that can be implemented in Teach-R are discussed, as well as their advantages and disadvantages. Integrating this topic into existing research, some approaches for voice and speech analyses are reviewed. From current knowledge, no approach for voice and speech analysis discussed before has been implemented in a virtual environment. Thus, this is an attempt to close this research gap. An evaluation of the implemented features is done, but the added value should be further investigated by improving the features and adding more analyses similar to a professional voice screening. This is discussed in more detail in the final chapter 6.

No Similar Application

To look into existing tools that combine the topic of VR and speech, some publications are examined. The majority of those tools address anxiety for public speaking. For this, a meta-analysis was conducted by Lim et al. [LAE23], with the topic of treating the fear of public speaking in VR. This study does not include the analysis of voice, but the usage of a virtual environment as a method for training or treatment. They state that the usage of a virtual environment can offer advantages for a protected environment, the possibility to adapt the environment to the required needs, and that the users can experience it at their own pace (Botella et al. via [LAE23]). One of the results of this meta-analysis is that VR is an appropriate method to treat public speaking anxiety.

Simulations for Public Speaking

Among the analyzed papers of the meta-analysis is the article by El-Yamri et al. [ERGM19], who developed a virtual reality game for public speaking. One of the features is the reaction and feedback of the audience based on the speaker's voice tone, speech content, and gaze. Based on the voice tone, the predominant group of emotions is determined and combining this with the content of the speech, the behavior of the audience is generated.

Analysis based on Voice Tone, Content, Gaze

Another paper on overcoming speech anxiety using VR is presented by Aljabri et al. [ARQ+20]. It is similar to the approach from El-Yamri et al., but it also takes the severity of stuttering into account. This is done by looking at unexpected pauses in speech and repetition of a word or character.

Analysis based on Stuttering

Even more parameters are introduced by Igras-Cybulska et al. in [IHC+23] for analysis in a public speech simulation. Next to emotions and speech rate, they consider stress, intonation, pauses, speech types, gestures, eye-tracking, and psychophysiology analysis. The intonation is analyzed by looking at the speech melody based on pitch contours. In contrast to the intonation, the pauses focus on parts in the speech when either nothing is said or no statement is made because fillers are used. The type of speech refers to the context where the speech is held, for example, a TED talk or vlog. Particularly interesting for the future work of this thesis is the gesture analysis. Igras-Cybulska et al. classify the gestures as low, medium, and high and into the categories of dynamics, range, and distance between the hands.

Analysis based on Speech Rate, Stress, Intonation, Speech Types, Eye-Tracking

Concluding, the combination of virtual environments and voice already exists, but the focus of these analyses is not on the voice or speech itself. However, the usefulness of VR as an environment for simulations is undisputed, and the issue of voice problems by teachers is at hand. Therefore, the implementation of voice and speech analysis into Teach-R is an attempt to investigate its usefulness. Limitations and restrictions discussed later suggest that this cannot replace a professional voice screening, but can provide further assistance to teachers-to-be.

Consid-ereations for Teach-R Setup

The setup of virtual systems has further advantages and disadvantages, which should be discussed. The major advantage is that the analyses do not take place in a laboratory, but in a simulation, which can increase the validity of the data. Using an integrated microphone in the headset (for this thesis, an HTC Vive was used) is an advantage and disadvantage at the same time. The advantage is that the distance of the microphone to the mouth stays constant for each user. This improves the validity of the recording, similar to a professional voice screening. Furthermore, it induces a lowering in environmental noise according to [BDB15]. On the other hand, they state that the proximity effect leads to boost lower frequencies if not compensated by the device. The quality of the integrated microphone is not as good as the professional equipment from a voice screening, which reduces the validity and reliability of the recording. Barsties and De Bodt support this with the statement that microphones and digitized mediums influence voice recordings. Additionally, they state two further factors for the reliability, the environmental noises and the voice production of the user. The environmental noise should be as low as possible, i.e. a threshold lower than 50dB (Titze via [Bar13]). For the user's voice, factors such as loudness, pitch, gender, produced vowels, and the interaction effect need to be taken into account. As mentioned before, the analysis should be carried out incidentally without the user engaging in tasks. Thus, produced vowels are not of interest for this thesis, and, consequently, are excluded.

Chapter 3 Conceptualization of the Implementation Goal

The goal was to implement speech analysis in Teach-R and additionally create visualizations in both, Teach-R as and in the coach. Therefore, visualizations for two and three dimensions needed to be planned and designed. After discussing concepts of voice and speech analysis in subsection 2.3, general guidelines for visualizations are presented in section 3.2, and potential visualizations are presented here. However, several restrictions came up during the implementation phase, which will be discussed first.

3.1. Restictions

As the goal of this thesis was originally to implement a voice screening similar to the one offered by the Chair for Contemporary German Language [Sti], several restrictions have occurred and need to be discussed. First of all, a professional voice screening uses a setup to properly record the generated sounds by the participants. The setup used for this master thesis does not include a professional voice screening setup because it is not available in the according lab. Instead, the microphone of the HTC Vive has been used. This is not only a restriction, but also a limitation of the analysis, as the results are less precise and may be error-prone compared to professional equipment.

No Professional Sound Equipment

Another restriction previously mentioned is the passive nature of the planned voice screening in Teach-R. The user should not actively conduct a voice screening, but get some analysis on the fly while training in the application. This means that holding a tone or changing the volume on command cannot be done for the analysis, but it should be based on connected speech. As presented in sections 2.3 or 2.3.1 several analyses require the participant to carry out those active tasks. An idea to handle the problem would be to edit snippets of a voice recording together and duplicate excerpts to simulate a held tone. This topic and its implementation do not fit into the scope of this thesis but will be added to the future work chapter.

Passive Manner of the Planned Voice Analysis

When planning possible software solutions, in the beginning, some important considerations were not taken into account by the author. Teach-R is an open-source project in the sense of open educational material. This work intends to maintain the licensing based on its existing level. Moreover, the existing software solutions need to be either adjusted to Unity or already adjusted for it. Limited by the scope of this master's thesis, only packages/ libraries made for the usage in Unity were included.

Continuing Open-Source Project License

An initially unexpected challenge posed the Audio Management of Unity since it is quite rudimentary. Unity itself provides just a few methods to work with audio recordings and hence, functionality to base analyses on those. A consequence of that is the shift from the ideal goal of creating a complete voice screening in Teach-R to the goal of building the

Rudimentary Functionality for Audio from Unity basis for voice analysis. Most analyses implemented are based on calculations from the amplitude values originating from the audio recordings. Likewise, Unity only supports recording on one channel, i.e. only mono sound can be recorded.

Performance of Teach-R

The final restriction discussed is the performance in Teach-R. As Teach-R is already a quite complex simulation and is set in VR, the performance needs to be optimized. Performance is always an important aspect when implementing, this should be not too much of a discussion. Unfortunately, when preparing for the evaluation of the implementation, it has been noticed that the performance in VR only passes the threshold in some parts. Most of the implementation was done in a two-dimensional manner. VR was only used to test the functionality prior to the evaluation because the equipment was in a lab at the chair. This has several reasons. On the one hand, as mentioned before, Unity provides only a rudimentary repertoire, and every analysis is based on the amplitude values. New amplitude values are drawn from the active audio recording every second, depending on the set frequency of the audio recording. To reduce the computational cost, the frequency is set to 16kHz, meaning an array with 16,000 floats needs to be handled every second. On the other hand, the data needs to be transmitted, preferably synchronized, to have visualizations in the coach as well. The connection or communication between Teach-R and the coach operates via websockets in both directions. Sending packages every second with an array of floats with a length of 16,000 requires optimization. Lastly, speech-to-text is included in the course of this master thesis for the basis of speech analysis. Again, the synchronized operation is preferable, therefore a stream is started to translate the recognized speech. Those operations combined with synchronized visualizations in Teach-R are compute-intensive and reduce the frame rate significantly when utilized in VR.

3.2. Visualization Guidelines

Despite the restrictions, data is collected and used in Teach-R. However, how visualizations for the three-dimensional space should be designed is a rarely discussed topic so far. Thus, the principles are reviewed for the two-dimensional use case and are then transferred to the three-dimensional space. The section is inspired by a paper from 2023 that was set to analyze designing visualizations and a dashboard for VR (cf. [Rec23])

3.2.1. Data Visualization

After gathering the data, the interest shifts to interpreting and analyzing it to gain insights. Taking a look at tabular data can be a first start to get a general overview of what has happened, but it is more difficult to see relations and developments in it. For a better understanding of the data collected by the audio recording, visualizations can help to look into different aspects of this data.

Existing Voice Analysis Visualizations

For voice analysis, some visualizations have become established, for example, a spectrogram or a waveform, but common plots like line graphs are commonly used as well. An example of a spectrogram can is given in figure 3.2. A spectrogram is a visual representation of the spectrum of frequencies in a sound or other signals, as they vary with time. It is one approach to analyze the frequency content of a signal over time, for example, when analyzing speech. To do so, the continuous audio signal is broken down into its component frequencies. For each segment, the amplitude of each frequency component is calculated and represented on the horizontal axis of the spectrogram. The intensity of a frequency is depicted by the intensity of the color, usually done by scaling from low (blue or green) to high frequencies (yellow, orange, or red). The resulting spectrogram

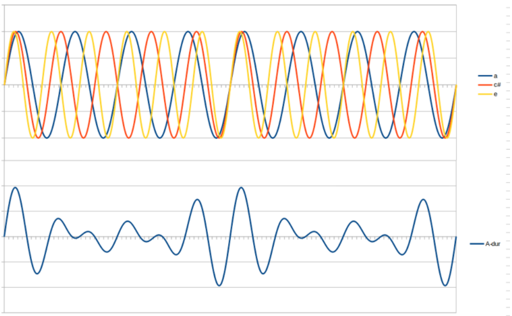


Figure 3.1.: Josef Pavlik, CC BY-SA 4.0, via Wikimedia Commons - A major chord in a waveform. In the top picture are the three constituent notes' waveforms. Below, is the combination of these waveforms as the complex waveform.

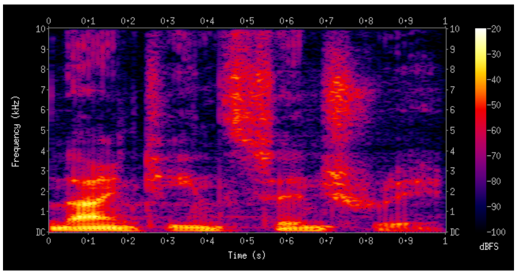


Figure 3.2.: Aquegg, Public domain, via Wikimedia Commons - A spectrogram visualizing the results of an STFT of the words "nineteenth century". Frequencies are shown increasing up the vertical axis, and time on the horizontal axis. The legend to the right demonstrates that the color intensity increases with higher density.

can be used to search for patterns in the audio, for instance, a held vowel can be distinguished from a consonant as vowels usually are stable sounds, whereas consonants are less stable.

Figure 3.1 shows a type of waveform, consisting of three notes. When combined, they result in a complex waveform. It provides a visual representation of the audio's intensity and amplitude changes, making it easy to spot patterns, loudness variations, and pauses. Simpler variations of waveforms represent only the combined form by illustrating it with small bars from the center to the relative position of the magnitude.

Principles presented from Midway

For a more general approach, the paper by [Mid20] introduces principles of effective data visualization. To give a brief introduction, some principles are presented here. Principle three states *use an effective geometry and show data*, which means that depending on the desired message of the plot, different types are appropriate. The author states, that most data can be sorted into four categories, amounts or comparisons, compositions or proportions, distributions, and relationships. Depending on the category, some types of plots are close at hand, for example, for amounts or comparison bar plots are used frequently. The fourth principle is *colors always mean something*, as they can add information and make visualizations memorable. With this Midway gives some recommendations, namely to use colors that work in black-and-white, as well as in color formats and are effective for colorblind people. It is important to note that the visualizations created within this thesis are only meant to give different perspectives on speech without a rating. Therefore, colors that are not stigmatized should be used where possible. More significant for adding visualizations for the coach is the principle of *panel, when possible*. This indicates that using the same figure again with a single feature changed for an easy comparison. An obvious approach is, since the data source is always the same, to use time on one axis and keep the plot and scaling the same. This way, comparing and recognizing patterns can be enhanced. The last principle mentioned here is *get an opinion*. For this, an evaluation takes place after the implementation phase to get feedback on the functionality, the visualizations, and the general usefulness of the results.

Usage of Colors Discussed based on Wilke

A more in-depth analysis of how to visualize data has been presented by Wilke [Wil19]. Similar to Midway, an important point is the use of color. One especially relevant point is encoding too much or irrelevant information by using color. For example, trying to differentiate a lot of categories by color might lead to a loss in readability. Wilke states that three to five categories are best when using colors to mark the categories. This can be interesting when analyzing what is spoken, for example, the separation of the type of lesson, i.e. giving instructions or holding a presentation, or the usage of filler words compared to technical terms and more simple explanations. Another guideline highlighted, similar to Midway, is not using color without purpose, as it can mislead the presented information.

When to Use which Types of Plots Discussed based on Wilke

Wilke goes into more detail about the use of plots depending on the type of data. For this, he categorizes visualizations into five categories from which amounts, distributions, proportions and x-y relationships can be of interest for this thesis. Amounts can be used when analyzing speech, for instance, depicting the number of words spoken and the time between speaking parts, or the amount of time the fundamental frequency is used or deviated from. To illustrate amounts, simple or grouped and stacked bar plots can be used, as well as dot plots and heatmaps. When the complete number of words is represented, a stacked bar plot could be used to show the overall amount split. The boundaries of the categories are less strong, as similar plots are useful for several cases. Thus, a stacked bar plot can also be used to show the proportions in the data, as the respective parts should add up to the whole data set. Other possible visualizations focusing on the proportion are pie charts, side-by-side or stacked bar charts. As a comment, pie charts fell out of favor for many researchers, as they are often criticized for being less easy to compare (cf. [Mid20, page 2] vs. [Wil19, page 96]). For distributions, Wilke recommends using histograms and density plots. A histogram could be used to compare the speech between different sessions or sections of a session. Finally, the x-y relationship could be valuable for this thesis when the course of a session should be evaluated. A line graph or a connected scatter plot can be used to look at the volume over time or the speech rate.

3.2.2. Data Visualization in VR

In a paper ([Rec23]) from 2023, visualizations in three-dimensional or VR are already discussed in more detail. The advantages of visualizations in three-dimensional or virtual reality are, for example, that the human brain is already accustomed to three-dimensional environments and that three-dimensional displays lead to better performance regarding tasks requiring undivided attention, information integration, or mental model development, compared to common displays ([WOH$^+$06] through [Rec23]). Another aspect is that interaction with a three-dimensional scene increases the sense of presence, this is in the sense of immersion into the environment. Additionally, in VR the user has more interaction possibilities with the data set which further enhances the exploration (cf. [vFL$^+$00]). On the other hand, there are several disadvantages. For instance, if three-dimensional environments are depicted on a two-dimensional display, namely perceptual ambiguities of depth, size, and distance, or the fact that three-dimensional displays are often ineffective visualizations for focused attention tasks ([MCFB91] through [Rec23]).

Transfering Guidelines to a Three-Dimensional Manner

Two guidelines for VR cannot be transferred easily from the guidelines introduced before, namely *use an effective geometry and show data* and *panel, when possible*. Typical plots like the ones mentioned before are less effective in a three-dimensional manner for the given reasons. On the other hand, the third dimension and the egocentric perspective enable other visualizations. Lee et al. [LBL$^+$21] provide some more advantages on how visualizations in a virtual environment can enhance the data evaluation. It offers a unique opportunity to create engaging and impactful experiences for users and can lead to a deeper level of engagement and understanding compared with two-dimensional visualizations. For example, sound waves can be made visible or the range of the used volume can be indicated. This can improve users' perception of data by providing a sense of scale, depth, and spatial relationships that are not easily conveyed in traditional data visualizations. Furthermore, the context of the data is far better established compared to merely looking at visualizations from a known data source. In hindsight, this would only represent the data for this specific moment, as it can be generally more difficult to represent a data history. For the data source of audio recording, playback could be implemented to revisit data and the visualizations representing the data according to it.

In conclusion, in the three-dimensional environment, the typical plots are less efficient to implement. Instead, visualizations that use the additional dimension and let the user explore the data by navigating through it should be used. An appropriate way of inspecting the data history needs to be found.

3.3. Conceptualization for Teach-R

As discussed in the section before, it should be made use of the advantages of visualizations in three-dimensional space. The virtual reality environment makes it possible to make the sound visible by giving real-time visualizations in the environment and the speech can influence the students. Combined with the knowledge from the related work chapter 2, many ideas for visualizing the voice through sound and speech can be developed. In this section, concepts for three-dimensional visualizations are presented and the two-dimensional menu for the visualization will be revised.

Considering the current state of Teach-R, there is no general concept for two-dimensional interfaces for interacting with features. In the first tutorial, a blackboard and speech bubbles are used and can be interacted with. However, this is currently undergoing an update at this point. As already mentioned, in a course in 2023 first visualizations for learning analytics were added, which have a two-dimensional menu to toggle them.

3.3.1. Visulization Menu

Old Visualization Menu

After discussing restrictions for the implementation, the conceptualization and planning for the visual elements are presented, beginning with the visualization menu. In a course in 2023, where the basis for learning analytics was created in Teach-R, a visualization menu was created to toggle visualizations based on the tracked data (see figure 3.3). This visualization menu was attached to hover over a book (see figure 3.4) placed on the teacher's desk and was activated by grabbing the book. The idea was to create the metaphor of a class logbook, in which the user can review notable situations of that session via learning analytics. The active state of the visualizations depended on the active state of the menu, i.e. when the menu was closed, the visualizations were turned off.

The drawbacks of that realization are that the menu might cover up the current visualization and it was a static translation from the book. This means, that if the book was rotated, the menu would be below the book and, therefore, not be interacted with easily.

Improvements for the Visualization Menu

In order to address these issues, the visualization menu should be reworked, as it needed more buttons for further visualizations as well. In cooperation with Jasmin Haranto, already mentioned in section 1.4, some specifications for the new menu were set up. The menu should be detached from the book, moved to a wall, and be more easy to extend.

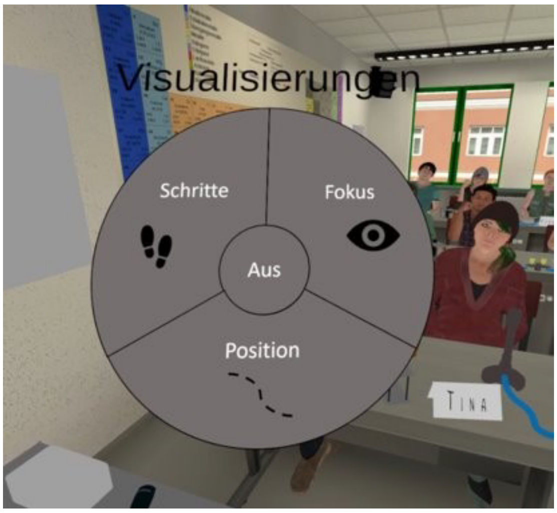


Figure 3.3.: Screenshot from Teach-R. The previous visualization menu from the course of 2023.

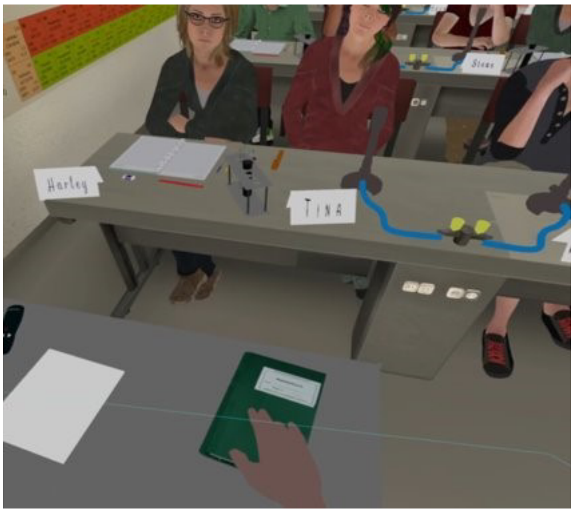


Figure 3.4.: Screenshot from Teach-R. The book to which the visualization menu was attached.

Instead of being attached to the book, the menu should be on a place at the wall to not block parts of the view of the user. Another advantage of using the book as an attachment point was that it already exists in every classroom, which made it a dynamic approach to locate the menu. Hence, another attachment point that already exists in every classroom would be needed, preferably on a wall. For extendability, a radial menu is already a reasonable choice, as further categories in circles around the center can be added.
Following this discussion, a menu was developed that included those aspects and can be seen in figure 3.5. The poster shows two generic diagrams, has the title *Learning Analytics - Analyse von Lernendendaten - Für Lehrende, Lernende und Forschung* [Anaylsis learner data - for teachers, learners, and research] and designed to have a play icon. The objective was to blend in with its surroundings while still standing out enough to be noticed. If the poster is touched, the visualization menu will appear. Still, it bears a resemblance to the previous menu design, but the text was removed and a fourth button was added to make it look less overloaded. In the middle, the off button is the same, while the inner ring of buttons now presents the categories, namely eye-tracking, movement, teaching management, and speech. The newly taken icons are from Font Awesome [Fon], for the reason of having free icons with the same design as used in the coach. The only icons created independently are the dashed line and the circle with two arrows. For the category of speech analysis, a microphone is used, while for the individual visualizations, a typical fog icon is used. A circle with a double-headed arrow within it is used to display the circular volume visualization and an icon showing two persons with a double-headed arrow representing the teacher clone. When one of the categories is touched, the individual visualizations are shown. At the moment, every category has three buttons for visualizations, but it can be further extended in the same manner with further categories in further extensions.

3.3.2. Volume Visulization

Several ideas came up through research and discussions with Jasmin Haranto and my supervisors to visualize volume. First, a surface map shows how sonic waves develop over distance and with other objects. It was discarded quite fast, as the physics in Teach-R

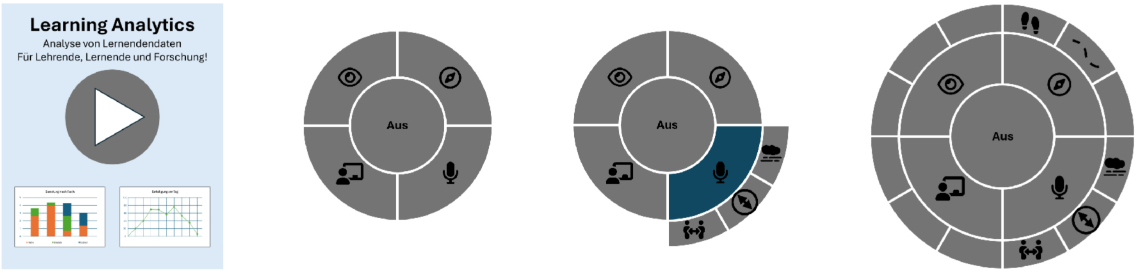

Learning Analytics
Analyse von Lernendendaten
Für Lehrende, Lernende und Forschung!
Aus
Aus
Aus

Figure 3.5.: Design of the activation poster and the visualization menu. From left to right: The activation poster. It is shown when the visualization menu is inactive. The new visualization menu without anything extended. The new visualization menu with toggled speech analysis visualizations. The new visualization menu with everything extended. The blank tiles are not yet in use.

are not that precise, and obtaining that information from Unity would be raather difficult. Nonetheless, with that idea and inspiration from an existing visualization in Teach-R, a heat map came up. As again the restrictions are too hard to handle, the basic concept was taken and a round sprite was created and scaled according to the current volume. The sprite that represents the circle has a color gradient and its position is updated to that of the player except on the y-axis. The range of the sprite can be calculated by taking existing estimations for the decrease of volume over distance. Another approach would be to place audio recorders at the farthest away position in the room and determine the point at which the volume drops to zero for that distance.

Circular Volume Visualization

After discussing the limitations of planned volume visualizations and how to handle them, another idea came up: a fog clearing up as the volume rises. It should work similarly to the scaling sprite, but instead of growing with increasing volume, it clears up. Unity already provides a fog, which needs to be tested whether it can be used for this purpose.

Fog Volume Visualization

3.3.3. Replay on a Teacher Clone

As the evaluation of the session in Teach-R should not only be based on current actions, the possibility of watching a replay of the session came up. In order to realize that, the model of the teacher is placed in front of the class. Initially, the position and potentially the hand position should be reproduced as well. It should be an easy extension, as the controller and headset positions and rotations are already tracked and a method to fetch this data is already implemented. This would further improve the precision of the reproduced session, as the teacher has an audio source on it and certain aspects, such as the volume can be experienced from different positions in the room as the teacher walks through the classroom.

Extensions for the Teacher Clone

To facilitate a better immersion into the simulation, options and sound effects should be added to the replay. Possible features would be to add a reverb effect to simulate speaking in a larger room. To some extent, disturbing background noises also could be added to enhance the immersion, as mentioned in section 1.1. To control the individual features, buttons can be added either to the menu or the clone itself. Both have their advantages, as adding buttons to the menu is comparatively easy. The buttons on the coach could be two-dimensional as well, but that might complicate the control as they are in a fixed place and do not rotate. To better reach the buttons, they could be converted into a billboard to always adjust their rotation to the view direction of the user. Even further, the

buttons could change direction and rotation depending on the view direction, i.e. circle around the coach to always be closest to the user. Away from two-dimensional buttons to an interaction via metaphor, the interaction could be connected to the mesh of the clone. Depending on where the coach is touched, different controls could be triggered. For example, when touching the left hand, the recording could be rewound, and when touching the right hand, it could be fast-forwarded. This might be less intuitive for the user though, as there are no hints to control the recording in this way.
As already discussed in section 3.2.2, other visualizations could be toggled together with the active teacher clone. For instance, when watching the replay and the clone represents the movement in the classroom, the circular volume visualization could be applied to it.

Combination of Visualizations with the Teacher Clone

3.3.4. Pronunciation Visulization

One possibility of visualizing the intelligibility in regards to pronunciation or even content could be via the students' behavior state. Presuming a way of tracking those measures, individual students could change their facial expression and their gestures depending on the comprehensibility as realized by El-Yamri et al. [ERGM19]. For example, if the user speaks unclearly, some randomly selected students could tilt their heads, draw their eyebrows together, or ask their neighbor to repeat what was said. The reactions need to be clear and noticeable for the user to be able to interpret the behavior of the students.
When replaying the session with the teacher clone, the behavior of the students could be recreated. It needs to be tested, whether it is more performant to save behaviors of the students or whether the students should react again on the recorded audio. For a new reaction to the recording, using the previously recognized speech could lead to the same behavior. Creating completely new reactions with the recording by rerunning speech-to-text might result in a more accurate recognition. This could be the case because an asynchronous larger model could be used but that might differ the reactions of the students.

Behavioral Reactions on Replay

3.3.5. Neutral Vocal Position Visulization

As the neutral vocal position can change in two directions, i.e. a higher or lower pitch, three states need to be visualized. One approach to represent those states could be achieved through the controllers. Should the pitch get too low, the left controller could emit a vibrating signal and vice versa. When the user is using the neutral vocal position, no controller would be vibrating. However, this could be quite perceived by some as rather confusing, as the metaphor of controller and voice might not be intuitive and the advantages of the virtual environment, as presented in section 3.2 would not be used.

Feedback via Controllers

Moreover, color filters could be added to the camera that ought to change the overall color of the view depending on the state of the pitch. In more detail, the whole view should not be bathed in one color, but one color should be more emphasized. This is inspired by various video games where the entire screen or the edges are dyed if the health status of the player drops critically low.

Feedback via Coloring

3.4. Conceptualization for the Coach

The design of the coach is quite straightforward, as seen in figure 1.2. On the left is a list of different behaviors, in the middle is a plan of the classroom the user is currently in, and on the right are some options, for example, to control the actions of students. In the header, the name of the coach is written on the left, the current page is marked in the middle, and options for tracking, keyboard shortcuts, and a button for help are positioned on the right. The current page functions already as a button itself, as more pages with features are in progress, but not running yet.

3.4.1. Visulization Page

Imitating the Design from the Control Page

Expanding on the conclusions of previous chapters , a page for visualizations in the coach should be added and accessed through a button next to the existing Klassenraum-Steuerung [Classroom Control]-Button. In congruence with the classroom control page, a list with possible visualization, categorized based on the tracking source is added on the left. For this master thesis, the category is speech analysis.

Differences to the Control Page

In contrast to the behavior selection, all visualizations should be able to be activated at once, i.e. multi-selection, not single-selection. The visualizations can be toggled to just appear below active visualizations or can be drag-and-dropped to a desired position. This can be expanded to a grid, so visualizations can be placed next to each other for a better comparison.

The controls on the right side of the page can be reduced, as scenarios and lesson programs are not that relevant for the visualizations. Instead, a simple recording control is added with buttons for starting and stopping the recording. The toggling of the recording should be separate from the existing tracking configuration, as the existing tracking is completely managed with OmiLAXR [GHS23] and is connected to a learning record store.

Extendability of the Design

This design allows an easier extension of the visualization page. On the left, categories can be added, and, if necessary, the groups can be made retractable. The same can be done on the right for further options. As mentioned before, most of the tracking done with OmiLAXR can be controlled via a button in the header. If another data source is added which is not tracked with OmiLAXR, it can be added to the right-hand side or another menu where that tracking is controlled can be added.

3.4.2. Visualizations

Speech Analyses

First of all, the **transcriptions of speech-to-text** should be simple visualizations. Each time a segment is recognized, it should be printed with the corresponding timestamp. Linked with this, an **analysis of the spoken text and the pronounciation** should be done. The analysis of spoken text could mark words that are used significantly often. Additionally, words could be categorized into appropriateness for a classroom with color coding, but that might be heavily depending on the context of the lesson. To analyze the pronunciation, the confidence of the speech-to-text tool could be displayed for words or segments. This, in turn, would be highly dependent on the quality of the results and the model used. Another method would be to let the person controlling the coach mark words that are recognized incorrectly by the speech-to-text tool. However, this would mean an increase in the workload. To visualize both attempts, the detected words and sentences could be marked by color coding, font size, or other marks like an underlining. For the color coding, the choice would need to be done carefully, as no rating should be done. Furthermore, the distribution of the teacher speaking versus the students speaking versus silence could be shown with, for example, a stacked bar chart or a pie chart. Another possibility for speech analysis would be to investigate the words used, for example, the amount of technical terms versus informal terms versus filler words as already mentioned in section 3.2.

A computation of the **speech rate** could also be based on the results of the speech-to-text as well. Both the overall average speech rate as well as the current speech rate can be of interest to that. For the overall speech rate, a numerical value might suffice, while for the current value development over time is more meaningful. One possibility for depicting the current speech rate is a line plot with the rate on the y-axis and the time on the x-axis.

Voice Analyses

Similar to the current speech rate, the **volume** can be visualized well with a line graph. On the y-axis is the volume in dB and on the x-axis are the according time stamps, as suggested for x-y relationships in section 3.2.

To gain a deeper understanding of the voice, a spectrogram can be added for interpreting **frequency of the voice and the intensity of the amplitude** and a waveform for the amplitudes as mentioned in 3.2. A spectrogram depicts how the intensity of different frequencies in a signal changes over time. These frequencies are shown on the y-axis in contrast to the timestamps on the x-axis. The intensity of the frequencies is represented by a color gradient having two extremes and a neutral state.

Another visualization that could be added is the **waveform** for voice. A waveform shows sound waves over time. It is similar to a scatter plot, but the individual points have a vertical line to the x-axis. The x-axis represents the time, while the y-axis represents the intensity of the wave sound. The intensity of the wave sound is analog to the volume. Different sounds create different shapes and patterns in the waveform. For example, a simple tone might create a smooth, regular waveform, while a complex sound like human speech will generate a more irregular and varied waveform.

Chapter 4 Implementation

After discussing the conceptualizations for new visualizations and improvements of the visualization menu, the implementation based on those is presented. Due to the restrictions and the scope of this thesis, several features are not implemented and are deferred as future work. Added features are volume analysis with visualizations, replaying of audio recordings, and speech-to-text. From speech to text, the speech rate and some numerical values are calculated.

However, as the foundation for speech analysis is implemented, building upon that and creating more analyses should be easier. This is discussed in more detail in section 6.2.

As a small remark, this master thesis is written in British English, while the implementation is done in American English. The reason for this is that the already existing implementation was done in American English and was continued for consistency.

4.1. Implementation in Teach-R

Teach-R is a simulation in VR to give teachers-to-be the possibility to train their classroom management, as already presented in section 1.3. It is realized in the game engine Unity [Uni] and the functionality is extended in C#.

Introducing Whipser.Unity

As the projection is open-source, the goal was to only add packages/libraries that do not change the license of the project. The only package added is whisper.unity [Gitb] created by Alex Evgrashin. It is an adaptation of Unity of the whisper.cpp [Gita] inference of OpenAI's Whisper. It runs under the MIT license and thus does not violate the open-source status of the project.

Dependencies of the new Module

Another requirement was to write the code to be as independent as possible of other modules. This has the advantage, that the added code can be used without other components of Teach-R and does not need fine-tuning depending on the usage. The requirements to use this speech analysis are

- The package whisper.unity needs to be included in the project,
- The visualization prefab, i.e. the visualization menu is needed to toggle the visualizations,
- The recording is dependent on the prefab of the teacher and is dynamically set to the prefab via the name.
- The changed visualization menu is placed on the location of a game object *posters*, which is in every classroom.

The teacher clone is using the Learning Record Store and the according fetcher but has a hard-coded alternative if the fetcher is not available. In the form of xAPI statements, the learning record store is a database for learning statements. In a simplified manner, xAPI statements consist of a subject, verb, and object, for instance, *User opens Visualization Menu*.

Structure of the new Module

Figure 4.1 shows the structure of the new and adapted classes used for the speech analysis. The class diagram does not represent the exact interplay of the classes but is slightly changed for an easier understanding. A short discussion of the interfaces between the newly written features and the already existing implementation, as well as an added package is described in the following. The `WhisperSTTManager` uses the package whisper.unity, whose results are processed for the `STTBehaviourChange` and are sent to the Coach. `SpeechCommunicationCoach` is the mediator class for the communication from Teach-R to the Coach. For the functionality of the visualization menu, the class `SpeechVisualisastionManager` is added. In connection with this, the `ToggleVisualisationScript` and the `VisualisationMenu` are extended for the new visualization and adapted according to the improvements presented in 3.3.1. A more detailed tour through the added functionality can be found in the subsequent sections.

To have the presented features in a scene, the *Visualisations* prefab needs to be added to the scene. The prefab contains the activation poster, the visualization menu and other objects needed for visualizations. Optionally, the prefab *LRS Fetcher* can be utilized as well for the initial position of the teacher clone.

4.1.1. AudioManager

Managing the Recording

The AudioManager class is the backbone of the entire functionality. As Unity does only provide rudimentary functionality for audio recording and processing, this has to be carried out manually. From here, audio recordings can be started and stopped, as well as replayed. The recording is set to a frequency of 16,000Hz which is a bit higher as the minimal frequency for voice recordings determined by Monson and Caravello [MC19]. Increasing the frequency can enhance the speech-to-text results and the precision of the analyses, but increases the required performance. When a recording is started, the speech-to-text stream and a repeating call to send the resulting amplitude values to the coach are triggered as well. Conversely, if the recording is stopped, the call to send the amplitude values is stopped, the resulting AudioClip is trimmed to the actual length, and all visualizations depending on an active recording are stopped.

Audio Replay

To replay an audio recording, the teacher clone needs to be active. During the replay, all other audio sources are muted for better comprehension. The functionality exists to either play the first recording from the beginning or resume the progress of the last replay after pausing.

Trimming Recordings

Trimming the resulting audio clips are necessary, as the recording requires a given length at the beginning, which is set to 30 minutes. In order to trim the AudioClip, the number of amplitude values is taken and a new AudioClip is created with the excerpt from the recording.

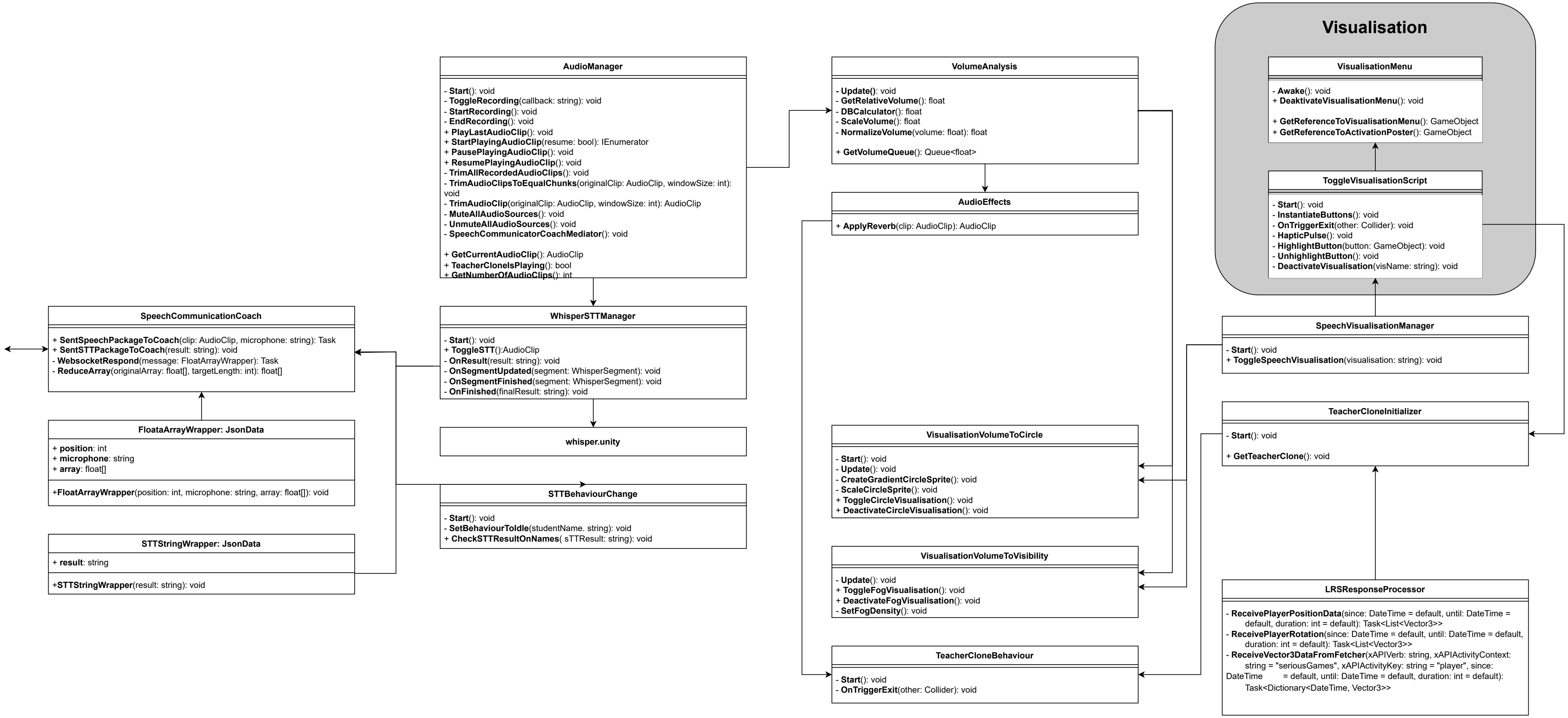

Visualisation
VisualisationMenu
- Awake(): void
+ DeaktivateVisualisationMenu(): void
+ GetReferenceToVisualisationMenu(): GameObject
+ GetReferenceToActivationPoster(): GameObject
ToggleVisualisationScript
- Start(): void
- InstantiateButtons(): void
- OnTriggerExit(other: Collider): void
- HapticPulse(): void
- HighlightButton(button: GameObject): void
- UnhighlightButton(): void
- DeactivateVisualisation(visName: string): void
AudioManager
- Start(): void
- ToggleRecording(callback: string): void
- StartRecording(): void
- EndRecording(): void
+ PlayLastAudioClip(): void
+ StartPlayingAudioClip(resume: bool): IEnumerator
+ PausePlayingAudioClip(): void
+ ResumePlayingAudioClip(): void
- TrimAllRecordedAudioClips(): void
- TrimAudioClipsToEqualChunks(originalClip: AudioClip, windowSize: int): void
- TrimAudioClip(originalClip: AudioClip, windowSize: int): AudioClip
- MuteAllAudioSources(): void
- UnmuteAllAudioSources(): void
- SpeechCommunicatorCoachMediator(): void
+ GetCurrentAudioClip(): AudioClip
+ TeacherCloneIsPlaying(): bool
+ GetNumberOfAudioClips(): int
VolumeAnalysis
- Update(): void
- GetRelativeVolume(): float
- DBCalculator(): float
- ScaleVolume(): float
- NormalizeVolume(volume: float): float
+ GetVolumeQueue(): Queue<float>
AudioEffects
+ ApplyReverb(clip: AudioClip): AudioClip
SpeechCommunicationCoach
+ SentSpeechPackageToCoach(clip: AudioClip, microphone: string): Task
+ SentSTTPackageToCoach(result: string): void
- WebsocketRespond(message: FloatArrayWrapper): Task
- ReduceArray(originalArray: float[], targetLength: int): float[]
WhisperSTTManager
- Start(): void
+ ToggleSTT():AudioClip
- OnResult(result: string): void
- OnSegmentUpdated(segment: WhisperSegment): void
- OnSegmentFinished(segment: WhisperSegment): void
- OnFinished(finalResult: string): void
SpeechVisualisationManager
- Start(): void
+ ToggleSpeechVisualisation(visualisation: string): void
TeacherCloneInitializer
- Start(): void
+ GetTeacherClone(): void
FloataArrayWrapper: JsonData
+ position: int
+ microphone: string
+ array: float[]
+FloatArrayWrapper(position: int, microphone: string, array: float[]): void
whisper.unity
VisualisationVolumeToCircle
- Start(): void
- Update(): void
- CreateGradientCircleSprite(): void
- ScaleCircleSprite(): void
+ ToggleCircleVisualisation(): void
+ DeactivateCircleVisualisation(): void
STTBehaviourChange
- Start(): void
- SetBehaviourToIdle(studentName. string): void
+ CheckSTTResultOnNames(sTTResult: string): void
STTStringWrapper: JsonData
+ result: string
+STTStringWrapper(result: string): void
VisualisationVolumeToVisibility
- Update(): void
+ ToggleFogVisualisation(): void
+ DeactivateFogVisualisation(): void
- SetFogDensity(): void
LRSResponseProcessor
- ReceivePlayerPositionData(since: DateTime = default, until: DateTime = default, duration: int = default): Task<List<Vector3>>
- ReceivePlayerRotation(since: DateTime = default, until: DateTime = default, duration: int = default): Task<List<Vector3>>
- ReceiveVector3DataFromFetcher(xAPIVerb: string, xAPIActivityContext: string = "seriousGames", xAPIActivityKey: string = "player", since: DateTime = default, until: DateTime = default, duration: int = default): Task<Dictionary<DateTime, Vector3>>
TeacherCloneBehaviour
- Start(): void
- OnTriggerExit(other: Collider): void

Figure 4.1.: A class diagram displaying the structure of the newly implemented foundation for speech analysis in Teach-R. Dashed lines represent transitions between modules. Only methods are given, no variables. For better readability use the digital version.

4.1.2. VolumeAnalysis

Calculating the Volume

The analysis of volume is based on the amplitude values from the active audio recording. From the current position in the audio recording a window is taken with a size of 128 amplitude values and the maximal value of the squared amplitude from this window is taken. This procedure is done on every frame to get live data and have smooth visualizations. The results are saved in a queue with a static length of 200, to roughly represent three seconds. Since an application for a head-mounted display should have a frame rate of 90 frames per second, a length of 270 would be needed, but Teach-R does not achieve this rate. Thus, a smaller window is utilized. The queue works on the principle of FIFO, i.e. the first item added is the first that is deleted when the queue is at its limit.

Processing the Volume Queue

The added volume values to the queue are further processed, as the amplitude values from the audio recording are very small. As discussed in section 3.3.2, the volume should be presented in decibels, as this is the standard unit to measure volume. This did not work, as the calculation always ends up with unrealistic results as they are done from scratch based on the amplitude values. Unfortunately, Unity does not provide an easier and more accurate option to get the decibels. The method used to calculate the volume in decibel is still in the class, but commented out, as it is not used for the given reason.

Normalizing the Volume Queues

As an alternative, the values for volume are normalized for an easier understanding. The values are normalized by the formula $x = \frac{x-min}{max-min}$, where min and max are updated each time a new minimal or maximum value is found. Additionally, the maximal value is reduced a bit each time no new value is found since otherwise one loud sound would squeeze all following values. This approach could be similarly useful as using decibels, as Lee et al. [LBL$^+$21] found that it is often about the experience rather than numbers.

4.1.3. Teacher Clone

Initialization of the Teacher Clone

The teacher clone is realized through two scripts: *TeacherCloneInitializer* nad *TeacherCloneBehaviour*. In the *TeacherCloneInitializer* the prefab for the clone is loaded and initialized. If possible, the first position of the user is fetched from the learning record store, otherwise a hard-coded position is taken. The rotation is set to have the blackboard in the back of the clone.

Appearance of the Teacher Clone

The prefab is a simplified copy of the model of the teacher the user controls. As the model could not be seen before, the look might not have been important. Now if the clone is placed in the room, it is noticeable that the model does not have eyes but just empty eye-sockets. Furthermore, as the hands are positioned at the relative positions of the controllers, the model has no own hands. The Unity components RigidBody and AudioListener are added to the prefab, to be able to interact physically and play AudioClips.

Functionality of the Teacher Clone

To now add functionality to the clone, the *TeacherCloneBehaviour*-script is applied. When the clone is touched, the audio recording is started, resumed, or paused. When less than two seconds have passed and the model is touched again, the recordings are started anew from the beginning.

Audio Effects for the Replay

Through the class *AudioEffects* effects to apply to the replay should be added. A draft for a reverb effect is written but does not work properly. No further work was done on this due to time reasons, but also the design of proper interaction techniques to control the individual effects did not come to an adequate result. The easiest way would be to add two-dimensional buttons to the visualization menu, but it turned out to be already quite big and further buttons would not fit near the menu.

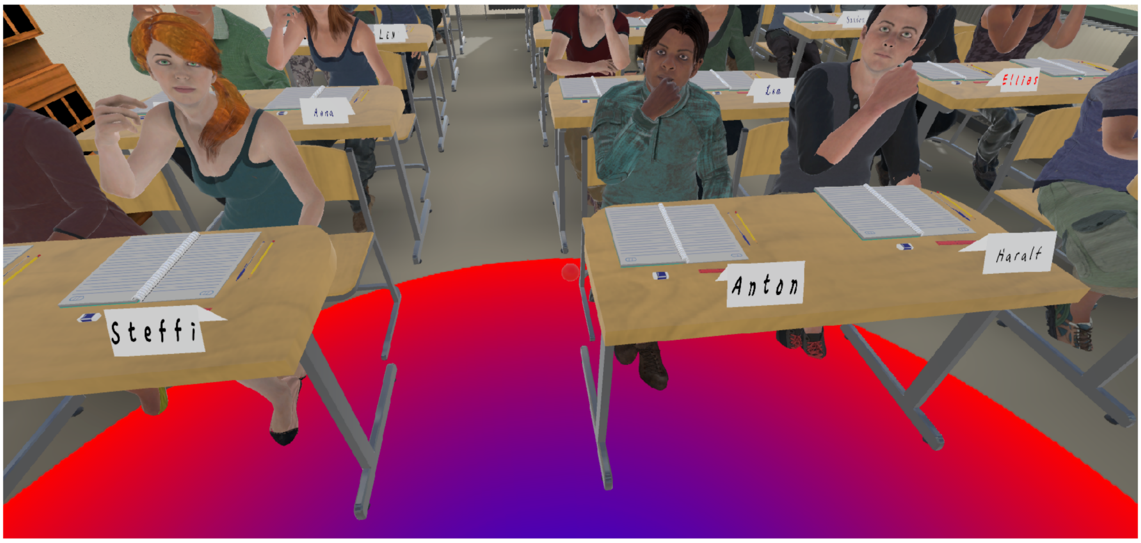


Figure 4.2.: Screenshot from Teach-R. The circular-volume visualization. A circular sprite positioned fixed below the user scaling relative with volume.

4.1.4. Volume Visualizations

Circular Volume Visualization

As already discussed in section 3.3.2, two visualizations are implemented for volume. The circle visualization is managed in class *VisualisationVolumeToCircle* and can be seen in figure 4.2. For this, a sprite is generated via scripting with a gradient from blue in the center to red at the edge. The color is chosen despite the discussion about colors in section 3.2 and the neutrality of the analyses. Blue and red are often used in heatmaps to represent two sides of a spectrum, but red is, depending on the culture, a warning or negatively connotated color. However, talking too quietly is bad in most scenarios for a teacher in a classroom which is why red is chosen. On the other hand, green is deliberately not used because speaking at a certain volume makes it difficult to rate positive without context. Instead, blue is chosen to represent a neutral area independent of the volume. When the visualization is activated, the sprite is scaled according to the volume calculated in *VolumeAnalysis* and the position is set to the position of the user on the x- and z-axes. Originally, the circle should be scaled to show the range to which a person at the edge can understand what is said. This can be done by a simplified formula that describes the drop of volume over distance, but that is dependent on having the volume in decibels. Since the calculation of decibels does not work properly, the circle is scaled to the normalized values, whereas a value of one corresponds to a circle radius of three meters in the virtual world.

Fog Volume Visualization

The second attempt for volume visualization by means of a fog is implemented as well in the class *VisualisationVolumeToVisibility* (see figure 4.3). Unity itself already provides a fog, which is utilized for this visualization. It adds an overlaying color to objects depending on the distance to the user. To use it for the visualization, the mode of the fog is set exponentially squared to work similarly to the logarithmic behavior of decibel. The mode of the fog represents how the color overlayed on an object is calculated and can be changed on runtime, also described as the density of the fog. Therefore, the density is changed if the visualization is activated by taking the minimum value of 0.9 and an initial density set minus the average of the current volume queue calculated from *VolumeAnalysis*. The minimum of these two values is taken to not get a value above 0.9, so the user can still maneuver through the classroom.

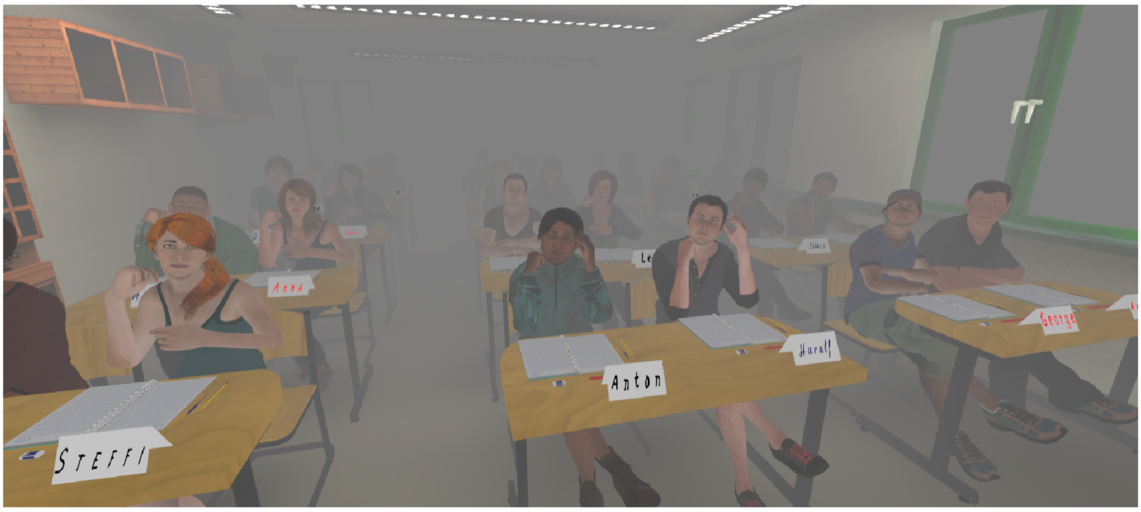


Figure 4.3.: Screenshot from Teach-R. The fog-volume visualization. The fog lies around the user and the density is scaled relative to the volume.

4.1.5. Visualization Menu

Activation Poster

For toggling every visualization described, the existing visualization menu is expanded as described in section 3.3.1. The menu is attached to a game object that already exists in Teach-R in every classroom, namely posters. This game object was originally used to pin actual posters teachers-to-be can use to hold a lesson. Therefore, the position of the posters added for a lesson might need to be adjusted. This was chosen because it already exists in every classroom at an appropriate position on the wall in terms of visibility and the metaphor of a poster is used for the activation of the menu.

Triggering the Menu

The existing functionality of the previous menu is reused, i.e. the navigation through the menu is triggered by touch, leading to a haptic feedback on collision and the last pressed button is highlighted. To fire events on collision, the Unity method `OnTriggerExit` is used. `OnTriggerEnter` is not used, as it leads to unexpected or even no behavior without displaying an explanation for it.

Improved Functionality of the Visualization Menu

The class *ToggleVisualisationScript* was refactored to improve readability and new functionality was added. In more detail, the visualization menu now handles the categories and extends the menu when one is selected as discussed before. The visualizations exclude each other, so only one can be active at a time. Additionally, a minor bug was fixed regarding the highlighting of buttons. Moreover, a bug in the existing eye-tracking visualization caused errors and broke the other visualizations as a consequence. Thus, the eye-tracking visualization is deactivated.

Gimbal Lock

VisualisationMenu was enhanced to handle the activation poster and with it the toggle for the visualization menu. A difficulty in initializing the game objects into the scene is the rotation. When copying the rotation of the poster object, the menu and the activation poster are off in the rotation. Therefore, the rotation is customized for both game objects. The result was tested in every classroom to work. Furthermore, the visualization menu seems to be affected by a gimbal lock. This means, that through the rotation of the menu itself, copying the rotation of the poster object, and adjusting the rotation afterward, two of the possible rotation axes are parallel to each other. This is fixed by giving the visualization menu an empty parent object, but it was not achieved to fix it properly.

4.1.6. Speech-To-Text

Added Speech-To-Text Models

Another added feature is speech-to-text using the whisper.unity package [Gitb] created by Alex Evgrashin. It uses a model that needs to be placed at the path *Assets/StreamingAssets/Whisper*. Two models are added to Teach-R, namely *ggml-tiny.bin* and *ggml-small.bin*. As the name suggests, the size of the model correlates with the accuracy of the results but also with the performance. Therefore, the tiny model is used in Teach-R as a stream is required for the implemented features. Tested on a desktop without the VR mode running, the performance with the tiny model did not influence the frame rate significantly. On the setup with the VR mode running, the speech-to-text stream leads to a drop in the frame rate.

Using Coroutines to Improve Performance

To handle this issue, the first attempt was to run the stream on coroutines. Usually, Unity executes everything on one main thread and for each frame. When a function is treated as a coroutine, the function does not need to finish in the period given by a frame but can operate over several frames. Thereby, other functions can continue in the cycle without waiting for that function to finish but still be in a synchronous manner. In the end, this approach did not solve the performance issue satisfactorily.

Using Async Tasks to Improve Performance

The next attempt was to use Unity's tasks, await, and async functionality. For async tasks, async and await keywords are used for handling asynchronous operations in Unity. The advantage is that those tasks are asynchronous but require a more elaborate handling. Nonetheless, when trying to adapt the stream from the whisper.unity package to be async, it either did not work properly or did not improve the performance to satisfaction.

Next possible Attempt

A last step would be to use other packages for asynchronous coding in Unity, but that exceeded the time scope. Since speech-to-text worked well without running VR and the implementation was completely done in a two-dimensional manner, the performance issue was discovered when preparing the evaluation. Next to other preparation tasks and fixing bugs that appeared through the usage of VR, no reasonable solution to this issue was found. How the problem was handled in the evaluation is described in 5.1.2.

Keyword Recognition

The package can return the results from the computation in four steps, i.e. when a segment is updated when a segment is finished, when the result is updated, and when the result is finished. The difference between a segment and a result, and update and finish is the weight added to the context and, thus, to the final choice of word probability. When a segment is updated, the recognized text is sent to the coach for further analysis. Furthermore, when the result is updated it is used in *STTBehaviourChange* to check for keyword recognition. At the moment, the result is examined on the names of the students in Teach-R. If a student is currently not idle and the name is recognized through speech-to-text, the student's behavior is set to idle. This can be extended to any keywords and change to other behaviors but is limited in the accuracy of the speech-to-text result. An unexpected problem is that not all student-objects' names match the names written on the nameplates. For keyword recognition with names, the names of the students are gathered through the nameplates, but the nameplates are changed on compiling Teach-R. In summary, some students react to names different from the names written on their nameplates.

The package was a bit adapted to the extent, that the recording was started automatically when speech-to-text was activated, as a reference to the active audio recording is necessary. Since a reference to the active audio recording is necessary for the analysis as well, the package was moved to the package folder instead of the library package cache to be able to change the implementation. Now, only a reference to the recording was given to the appropriate class when speech-to-text is activated.

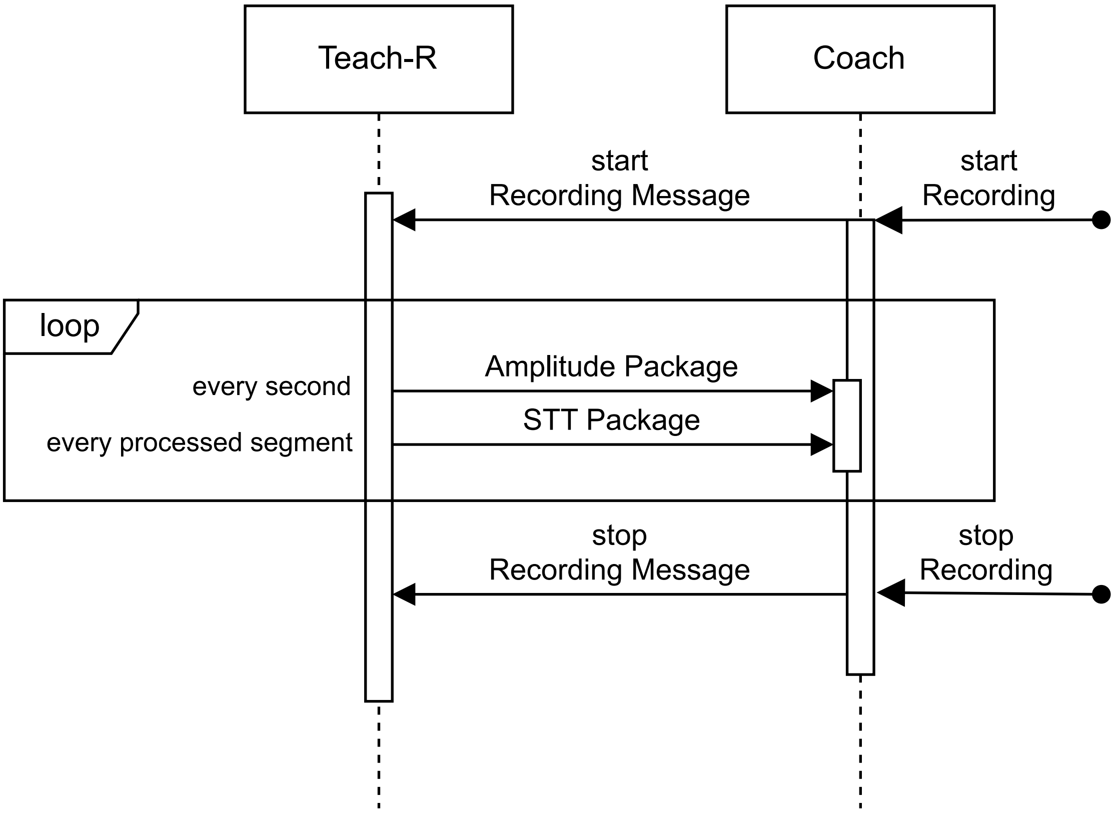


Figure 4.4.: Simplified sequence diagram representing the communication between Teach-R and the coach. If the recording is started, the amplitude and speech-to-text packages are sent in a loop. The amplitude package is sent every second, whereas the speech-to-text package is sent whenever a segment is processed.

4.1.7. Sending Packages to the Coach

Receiving Messages

To receive the message for activating or deactivating audio recording from the coach, a listener is created in the *AudioManager* class. This might be not so clean concerning the separation of functionality, but it was easier as the recording is toggled in that class. A simplified sequence diagram represents the communication in figure 4.4.

Wrapping Packages

Sending packages is managed in *SpeechCommunicatorCoach*. To send a package, the data needs to be wrapped in a JSON data type. For this, the two classes *FloatArrayWrapper* and *STTStringWrapper* are created. Next to the data, the amplitude package contains the name of the used microphone and the position of the first sample in the audio recording.

Sending Packages

An amplitude package is sent every second while a recording is active. In contrast to that, a speech-to-text package is sent whenever a segment is updated by the whisper.unity package. The transfer is then done using websockets, as mentioned earlier.

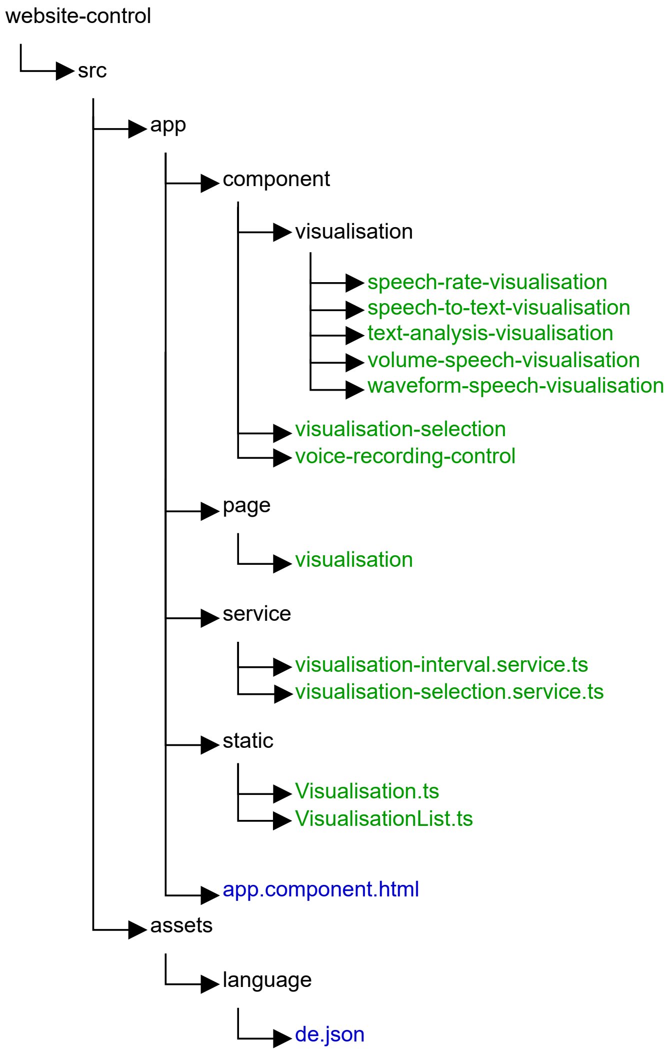
4.2. Implementation in the Coach

The coach is an Angular [Angb] project for controlling student behavior and further classroom features. Through websockets, bidirectional communication is established between the coach and Teach-R.

In section 3.2 it was already discussed that two-dimensional visualizations are better shown in a two-dimensional manner. To do so, some analyses and visualizations are computed in the Coach. This has the further advantage of having live visualizations for the person controlling the classroom via the coach.

Structure of the Implementation in the Coach

Figure 4.5 shows an excerpt from the new structure of the coach. This structure is not complete, as it mostly shows the files or folders that are added in the course of this thesis or actively changed. As angular reduces workload by creating components, services, pages, etc. by a simple command, additions and changes, for example in the routing, are not mentioned here.



website-control
src
app
component
visualisation
speech-rate-visualisation
speech-to-text-visualisation
text-analysis-visualisation
volume-speech-visualisation
waveform-speech-visualisation
visualisation-selection
voice-recording-control
page
visualisation
service
visualisation-interval.service.ts
visualisation-selection.service.ts
static
Visualisation.ts
VisualisationList.ts
app.component.html
assets
language
de.json

Figure 4.5.: Tree graph showing excerpt from new Coach structure. The structure is based on the Angular structure. Newly added classes are marked in green. Adapted classes are marked in blue.

4.2.1. Visualizations Page

Realization of the Visualization Page

Earlier mentioned in section 3.4.1, the new page is strongly imitated from the classroom control page (see figure 1.2 and figure 4.6). On the left is a list of buttons categorized by the data source, e.g. audio recording. Since speech or audio recording is the first data source to have visualizations in the coach, it is the only category at the moment. The icons for each button are again taken from Font Awesome. In contrast to the behavior list in the classroom control, any number of visualizations can be activated. After selection, the according visualization component is shown in the middle of the page. On the right, some options were removed compared to the classroom control page, but a component for controlling audio recordings was added. The header was expanded by a button to switch to the visualization page.

Visualization List

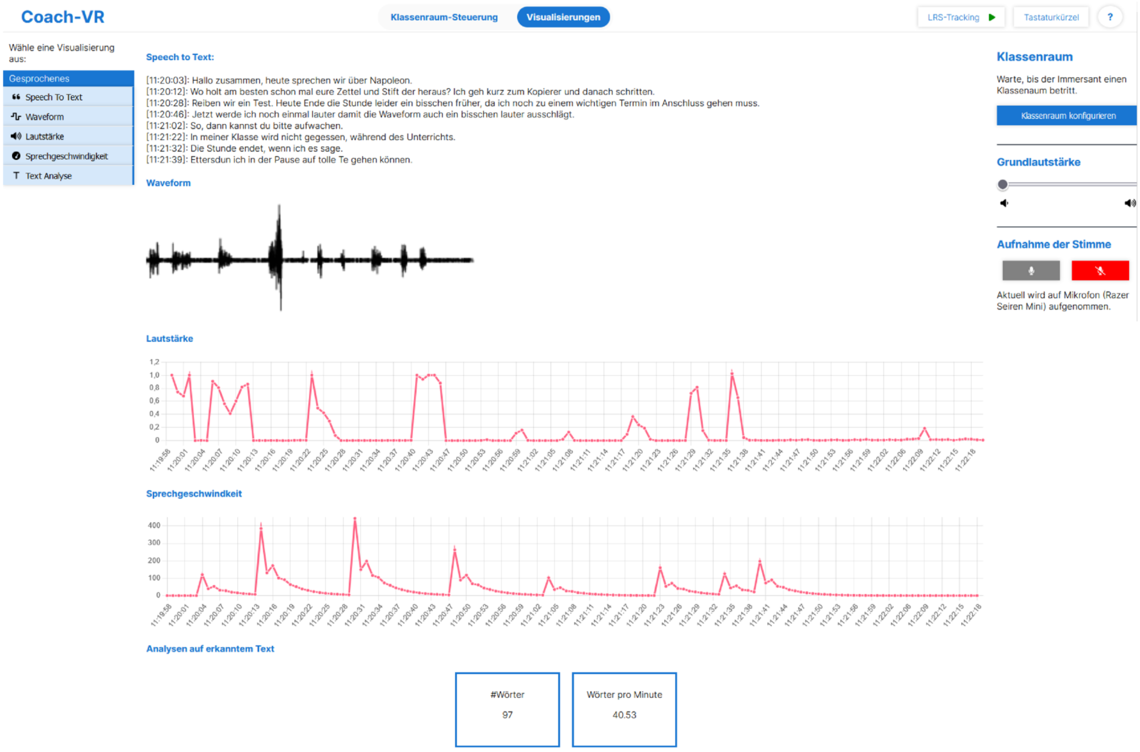
The realized visualizations are the speech-to-text results with timestamps in a list format, a waveform, a volume, and a speech rate visualization, as well as text analysis in a numerical manner. In order to dynamically create the visualization list, the two static classes *Visualisation* and *VisualisationList* were created. In *VisualisationList* the visualizations are created including the IDs, names, and icons. The *Visualisation* class is used for getters regarding the visualizations.

Requirements for Using Visualizations

When the coach is started and the visualization page is active, the same connection request is given as on the control page. Furthermore, when a connection is established, another prompt is shown requesting to activate an audio recording with a button similar to the prompt to connect. The button has the same functionality as the button in the audio recording control mentioned earlier.

Audio Recording Control

The audio recording control consists of two buttons, one for activating and one for deactivating the audio recording. If a recording is active, the name of the microphone used is displayed below the buttons.



Coach-VR

Wähle eine Visualisierung aus:

Speech to Text:

[11:20:03]: Hallo zusammen, heute sprechen wir über Napoleon.
[11:20:12]: Wo holt am besten schon mal eure Zettel und Stift der heraus? Ich geh kurz zum Kopierer und danach schritten.
[11:20:28]: Reiben wir ein Test. Heute Ende die Stunde leider ein bisschen früher, da ich noch zu einem wichtigen Termin im Anschluss gehen muss.
[11:20:46]: Jetzt werde ich noch einmal lauter damit die Waveform auch ein bisschen lauter ausschlägt.
[11:21:02]: So, dann kannst du bitte aufwachen.
[11:21:22]: In meiner Klasse wird nicht gegessen, während des Unterrichts.
[11:21:32]: Die Stunde endet, wenn ich es sage.
[11:21:39]: Ettersdun ich in der Pause auf tolle Te gehen können.

Klassenraum

Warte, bis der Immersant einen Klassenaum betritt.

Aktuell wird auf Mikrofon (Razer Seiren Mini) aufgenommen.

#Wörter: 97

Wörter pro Minute: 40.53

Figure 4.6.: Screenshot from the added visualization page in the Coach. On the left is a listing of the available visualizations for selection. In the middle are the currently activated visualizations. On the right are the classroom configuration, the volume control, and newly added the recording control. The header was extended by a button to switch to the visualization page.

4.2.2. Services

The analyses are mainly handled in two services, namely *visualisation-selection.service* and *visualisation-interval.service*. Subsequently, the visualizations from these analyses are then created in the individual components.

Visualization Selection Service

visualisation-selection.service is mostly responsible for the communication with Teach-R, i.e. receiving the amplitude and speech-to-text packages and sending messages to toggle the audio recording. Additionally, it provides the functionality to add timestamps to each received speech-to-text package. The name of the service is according to the functionality of toggling visualizations.

Visualization Interval Service

As the name suggests, the *Visualisation-interval.service* manages the computation for the analyses in intervals. Every second, the volume is calculated, the speech-to-text result is updated if something is received, the number of words is counted, and the average words per minute, as well as the current speech rate, is calculated. The calculation of the volume is similar to the one carried out in Teach-R and the results are saved in a list. Compared to calculating the average word counts per minute for the active recording time, the speech rate is calculated by the average word count of received speech-to-text packages of the last three seconds.

Sending one complete Package

One could argue to just send the completed analyses from Teach-R to the Coach instead of amplitude values. But this would shift the computational effort to Teach-R where a performance issue is already at hand. On the other hand, it would be possible to create a WAV file from the audio clip in Unity and send it to the coach. This might make some analyses easier but would increase the size of packages to be sent, which might become another bottleneck regarding performance.

4.2.3. Visualizations

Generating Visualizations

Most visualizations from the lists are generated by HTML and SCSS, only the plots are initialized and, if necessary, manually updated in the Typescript files. The volume and speech rate visualizations are generated using the chart.js library [Cha]. Both are line graphs with a grid in the background and timestamps on the x-axis. On the y-axis, the according interval is given, but the volume is again normalized to the range 0 to 1.

One exception is the waveform visualization, which creates a line for given amplitude data using the HTML5 canvas.

All data in One Plot

The visualizations are paused if no audio recording is active. If resumed, the visualizations are continued, whereas the volume and speech rate visualizations display the pause by showing the timestamps. The waveform does not depict pauses, as it does not provide timestamps. Pauses could be visualized by blank spaces in between the existing data or by filling the data list with zeros.

Chapter 5 Evaluation

One of the principles from Midway [Mid20] states *get an opinion*. Following this, after the implementation phase, an evaluation of the result was applied. As a proof of concept, about ten people should test the added features in VR as well as the visualizations on the coach. The objective was to find remaining bugs and get feedback on the usefulness and usability of the features. Due to the fact that only a few features were implemented that are used in a professional voice screening, the focus of the evaluation was not to compare the validity of the results to existing software. Instead, the functionality, perceived helpfulness, and usability were evaluated. Comparing the results with a professional voice screening should take place after the voice analysis is further developed and optionally an audio input with a better quality incorporated.

5.1. Method

This section serves as a description of how the evaluation was conducted. The sample of participants is introduced, and the procedure and questionnaire are explained. The limitations and validity of this evaluation are discussed in section 6.1.1 in the final chapter.

5.1.1. Participants

Recruiting Participants

The participants were recruited from a pool of friends and acquaintances along with employees of the chair LufgI9. An appointment inquiry was sent in which people could register for appointments spread over one week. Participation was completely voluntary and no compensations were offered. Some snacks were provided during the evaluation.

Representativeness of the Sample

Ten people participated in the evaluation, three of them were female, and seven were male. The age range was between 20 and 34, while most of the participants were between the ages of 25 to 29 years. Half of them were students and the other half were employees, of which two had completed a teaching degree. The degrees achieved to date were in line with the employment status, i.e. half of the participants have a bachelor's degree and the other half a master's degree.

Prior Knowledge of Participants

In regards to experience with virtual reality, two participants had never used an immersive system before, one had used one once, another participant had used one up to five times, while the other six participants had used one more than five times. The use of immersive systems varied greatly, from entertainment, over holiday planning, to development, research, and teaching.

The second aspect participants were asked to answer was learning analytics. On a scale from one (not at all) to five (very much), the participants should rate their knowledge in the area. As a result, the mean answer was 2.9 with a median of 3. The participants used learning analytics as different stakeholders, namely one as a studying person, five as a teaching person, and six as a researching person.

5.1.2. Procedure

The evaluation took place in the lab of the LufgI9 chair, at the computer science center of the RWTH University Aachen. It was completely held in German, which was also a requirement for the participants to speak and understand German fluently. One complete run-through took about 35 minutes on average.

Tasks in Teach-R

After a short briefing of the procedure and the topic of the thesis, the participants entered Teach-R, i.e. the frontal lecture classroom (see figure 1.1). In Teach-R, the participants have been given tasks verbally to first open the visualization menu and try out each speech visualization including the teacher clone. After that, the behavior of the student *Sven* in Teach-R was set to sleeping, whereupon the participant was asked to wake him up verbally.

Tasks in the Coach

To further evaluate the result in the coach, the participant was placed in front of the computer on which Teach-R and the coach were running. Here, the task was to open the visualization page, stop and restart an audio recording, open the volume visualization first and afterward all remaining visualization. Teach-R was continuously running during the usage of the coach, to get live data.

Questionnaire

Finally, a survey was done including demographic data, prior knowledge, the UEQ, and some open questions regarding the experience from the result in Teach-R and the coach, as well as suggestions for improvements. Next to the specific questions regarding the result of the implementation phase, the User Experience Questionnaire (UEQ) [LHS08] is a recognized and frequently used questionnaire. It measures user experience in six factors, i.e. attractiveness, perspicuity, efficiency, dependability, stimulation, and novelty. For this, a 26-item, Likert-scale rating is used with statements like *annoying - enjoyable* or *attractive - unattractive*. This questionnaire was selected, as it is fast and reliable (cf. [LHS08, text]). The questionnaire was split into seven parts as mentioned before, of which the UEQ and the questions about state and improvements were each randomized but not the order of the parts. The complete survey is included in the digital appendix and the protocol and tasks are included in the attachments D.

Changes in Procedure

Due to the performance issue already mentioned in section 3.1, the procedure was changed after the second run-through. When speech-to-text was active, the frame rate dropped and both participants became cybersick. The first participant wanted to go through despite the low frame rate but mentioned the disturbance caused by this. For the second participant, the feature was just activated when the task was given to wake up the student, but the low frame rate was mentioned again. After that, the speech-to-text feature was not activated when entering Teach-R in VR, but it was tested in a two-dimensional way. Another device with Teach-R and the coach running was provided to switch to after testing the visualizations in Teach-R.

5.2. Result

As the goal of the evaluation was to assess the usability and usefulness of the implementation, this section is split into three subsections: Code bugs, the UEQ, and the open questions from the survey.

5.2.1. Code Bugs

Interaction via Touch Gestures

A bug was found that visualizations were only deactivated when another one was activated. In more detail, when the visualization menu was turned off or after the button of

UEQ Scales (Mean and Variance)		
Attractiveness	1.680	0.46
Perspicuity	0.725	1.65
Efficiency	1.200	1.19
Dependability	0.975	0.44
Stimulation	2.175	0.04
Novelty	2.175	0.46

Table 5.1.: Mean values and variance of the UEQ results in a table.

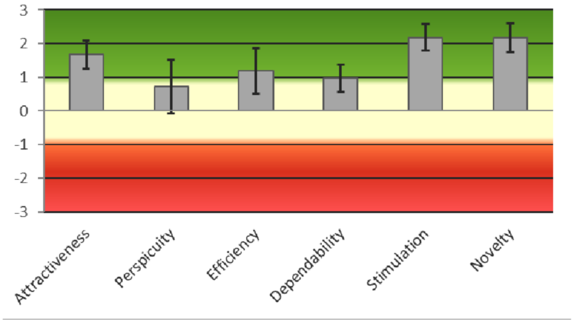


Figure 5.1.: Mean values of the UEQ results as a bar plot. The black lines on top of the bars represent the confidence intervals in which the true value of the scale mean should be located with a probability of 95%.

the active visualization was touched, the visualization remained active. For the evaluation, the functionality of the other categories in the menu, namely eye-tracking, position, and teaching management, was deactivated. This was done because the eye-tracking visualization led to errors, and the teaching management visualizations were not implemented yet. Still, after touching the eye-tracking button, the other buttons did not respond anymore. Lastly, the check for not triggering the off-button of the visualization menu when just touching the activation poster was revised, as the visualization menu was sometimes instantly turned off again.

Highlighting in the Visualization Menu

Not a bug, but it could be considered undesired behavior that the button to deactivate the visualization menu is highlighted when the visualization menu is reopened. This behavior is according to the implementation. However, highlighting the off-button might be confusing or at least unnecessary.

5.2.2. Usability Experience Questionnaire

A mistake was made when preparing the questionnaire for the evaluation, as an item of the UEQ was not present in the final questionnaire, i.e. *attractive* - *unattractive*. The results are calculated with the Microsoft Excel sheet provided on the UEQ website [Use]. The calculations are done with the missing item left out which has the effect of replacing it with the mean value.

How to Interpret the UEQ Results

Table 5.1 and figure 5.1 show the mean values of the six factors. As the UEQ does not provide one overall score, each item belongs to one factor, and for each factor, the mean is calculated. Which item belongs to which factor is shown in table C.2 in the appendix. The scale ranges from -3 to +3, but due to calculating the means from different people's perceptions and answer tendencies it is unlikely to get results below -2 or +2. As the sample for this evaluation only amounts to ten, the effect is less significant according to the analysis tool provided for the UEQ. Usually, a value between -0.8 and 0.8 is considered neutral, below a negative evaluation, and above 0.8 a positive evaluation.

Confidence of Results

To get an assessment of the validity of the results, the 5% confidence intervals are shown in table 5.2. A small confidence interval represents a precise estimation, as a small interval represents similar ratings of the items between the participants.

Confidence intervals (p=0.05) per scale						
Scale	Mean	Std. Dev.	N	Confidence	Confidence Interval	
Attractiveness	1.680	0.675	10	0.418	1.262	2.098
Perspicuity	0.725	1.283	10	0.795	-0.070	1,520
Efficiency	1.200	1.092	10	0.677	0.523	1.877
Dependability	0.975	0.661	10	0.410	0.565	1.385
Stimulation	2.175	0.635	10	0.394	1.781	2.569
Novelty	2.175	0.678	10	0.420	1.755	2.595

Table 5.2.: Confidence intervals (p=0.05) per scale. Including mean, standard deviation, sample size, confidence, and confidence interval.

5.2.3. Open Questions Regarding the Implementation

The open questions were structured in such a way that questions regarding the state of the implementation were evaluated first and afterward potential improvements were suggested. This is done once for the features in VR and once for the new visualization page in the coach. The complete answers can be found in the digital appendix, but some of the demographic data is left out on purpose, as it might harm the participants' anonymity. The answers are in German, as the evaluation was only done in German.

State of Teach-R

Visualization Menu

The first question was about the **intuitiveness of activating the visualization menu** via the poster. Half of the participants perceived the poster as an intuitive metaphor, with the most critique regarding the position on the wall, the interaction technique (touching), or that it blends too much into the environment. To evaluate the visualization menu further, the **understandability of the icon selection** was questioned. Most of the answers are positive, but it was mentioned that the context of the other buttons as well as the given task helped to interpret the icons. Adding a text, i.e. a name or short description, and depicting speech analysis with a combination of a microphone plus an icon for analysis, was suggested.

Visualizations

The feedback for the **usefulness of the circular and fog-volume visualizations** was mostly positive. It was experienced less well because visualizations were a bit delayed and jerky. Furthermore, the scaling was not perceived as accurate with both problems related to performance issues. Overall, the circular visualization was preferred in this regard, partly because it did not affect the view that strongly. The perceived **usefulness of the teacher clone** was mixed. In general, the interaction seemed to be not very intuitive, and hearing one's voice felt strange according to the participants' comments. Again, the quality and additionally the volume was mentioned to be not satisfactory.

Improvements for Teach-R

Visualization Menu

To improve the intuitiveness of activating the visualization menu, the participants were asked to give **suggestions different from the poster**. Several ideas were contributed like using something that looks more interactive like a whiteboard, laptop, or tablet. Furthermore, a book about learning analytics was suggested similar to the previous approach. As a more abstract improvement, the possibility of toggling the menu independent of the current position was brought up. More in detail, some answers propose using a button on the controller or a two-dimensional interface attached to the view. Likewise, more hints for the interactiveness of the visualization menu were desired. In addition, when selecting a button in one of the categories the highlighting of the category button should remain. One of the suggested improvements regards the size of the buttons, as some participants had difficulties selecting one particular button because they also accidentally hit another button.

Visualizations

Regarding the **circular-volume visualization**, the most named argument was the color of the sprite. A change to less striking colors or making the circle transparent was suggested. Additionally, to better estimate the impact of the volume, a grid or an actual color change is recommended, i.e. a green circle when talking in an appropriate voice, and red if too loud but here people with colorblindness need to be considered. Another idea was to dye the students or change their transparency depending on the residual volume at their position. The performance was mentioned for both, the circular- and **fog-volume visualization**. Linked to the performance, the sensitivity or scaling of the fog should be improved. Moreover, the density of the fog should be less high or the fog could be just to the height of the tables. The majority of **improvements for the teacher clone** referred to the model. Adding hands and eyes, and a better posture, as well as a female or a gender-neutral model, were mentioned several times. Besides that, the added value of having a game object for the replay was questioned. Improvements for the interaction, e.g. UI buttons, and further features like the spectrum of the volume were stated.

State of the Coach

Visualization List

To start with the coach, the **list of visualizations** to toggle each was shortly evaluated. The only point of criticism was that the buttons appeared to be as if they were mutually exclusive, as the same selection is used on the classroom control with mutually exclusive buttons.

Visualizations

For the visualizations themself, the **waveform** was perceived as useful, but the exact meaning was not clear, i.e. what does a peak or fall of the bar mean, and what source is used to measure it? Also mainly positive was the feedback for the **volume visualization**, with comments regarding the missing unit on the y-axis and lack of connection to the room, as the acoustics vary depending on the room. Several aspects to discuss were given for the **speech rate visualization**. One comment states, that with such an analysis a teacher might try to explain something with fewer words to get a better *score*, hence possubily limititing the comprehensibility. Another point concerned is the diversity of speech or language, as, for example, some dialects use more words for the same message. Therefore, the words per minute cannot be taken as a simple classification for the quality of speech. This makes it more difficult to give a reference for a desirable interval of words per minute, which was another suggestion. Lastly, the meaningfulness was questioned, as the teacher might stop speaking in the middle of a sentence because a student interposes something. The participant suggested creating a metric for the pauses in speech caused by students and incorporating that into a score for speech rate. The **numerical visualizations** were perceived as a good addition to the other graphs, but not that informative by themselves. Furthermore, the average speech rate has received similar feedback as the current speech rate.

Recording Control

Different Recordings Combined

Helpfulness

As some general feedback for the state of the coach regarding the added visualization page, the design of the voice recording control was mentioned. The participant perceived the icons as not very intuitive, as the crossed-out microphone was interpreted as muting instead of stopping. Additionally, adding the data from different recordings in the same plots was remarked as unexpected. One overarching comment stated that each visualization on its own is not very helpful, but the combination can be used for a proper analysis. For using every visualization together, a common refresh rate would be beneficial.

Improvements for the Coach

Further Visualizations

Suggestions by the participants for **further speech analysis visualizations** for the coach are a bar plot (perhaps a kind of spectrogram), numerical visualization of word occurrence, specifically fillers, and a sound propagation with its radius. Another analysis was mentioned regarding the pitch of the voice or whether a user is mostly speaking with a monotonous voice. For the visualization menu, changing the buttons was mentioned again by adding checkboxes or other toggle options. Continuing, the visualizations should be better separated and should not have a fixed place, i.e. newly toggled visualizations should appear on top or below already toggled visualizations.

Existing Visualizations

To improve the **waveform visualization**, a bigger scaling, a unit on the y-axis, and a numerical scale in the form of a grid in the background should be added for an easier estimation. The only improvement suggestion for the **volume visualization** is to add colors depending on the interval the volume is in. Approval for the **speech rate visualization** is that the visualizations look neutral or non-judgemental. This contradicts the improvement suggestion to add colors, which has already been made before as well. Furthermore, using a different kind of plot for speech rate and volume was proposed. For the numerical visualizations, a reference value and an explanation are desired and it was suggested to add the possibility to save and create a new one for different phases.

Replay in Coach

Finally, a generally suggested improvement for the visualization page in the coach is to add the possibility of replaying the audio recording there and add time filters to inspect parts of the session more accurately.

5.3. Discussion

To interpret the results from the evaluation, in this section, the findings are discussed and conclusions are drawn. Additionally, some observations from the evaluation by the conductor of the evaluation are added for a more detailed insight. First, the quantitative results from the User Experience Questionnaire are interpreted, followed by the qualitative ones from the open questions. This is used as a basis to discuss improvements for the implementation.

5.3.1. UEQ

In this section, the results from the User Experience Questionnaire are discussed and interpreted. How the feedback can be realized is further discussed in section 5.3.2.

Factors of the UEQ

Six factors are used for the UEQ to evaluate the user experience, namely attractiveness, perspicuity, efficiency, dependability, stimulation, and novelty. The overarching factor is attractiveness, while the rest is split into pragmatic or goal-directed aspects and hedonic quality or not-goal-directed. Schrepp et al. provide benchmark intervals for each factor in [SHT14] from excellent to bad. It is excellent, if the result is among the 10% of the best results, good when it is in the top 75%, above average for over 50%, below average for over 25%, and bad for everything below 25%. This classification is generated by comparing the results from different studies applying the UEQ (cf. [SHT14]).

Attractiveness

Attractiveness describes the overall impression from the users, i.e. how they like it. In the evaluation, the perceived attractiveness is rated with a mean of 1.68 (std. dev. 0.675) and a confidence of 0.418. This is a good rating according to the benchmark intervals and the small confidence interval supports this rating.

Perspicuity, efficiency, and dependability are the pragmatic quality aspects. These aspects are concerned with how well the product fulfills its intended purpose and helps the users accomplish their tasks. In this group, perspicuity describes how easy it is to get familiar with the product. It was rated with a mean of 0.725 (std. dev. 1283) and a confidence of 0.795, which is the lowest achieved rating and the largest confidence interval, and considered below average. This was to be expected, as many participants did not know how to activate the visualization menu and experienced issues when interacting with it. The high confidence and standard deviation indicate that there are participants who still experienced it in a good way. The visualization menu and the menu on the visualization page should be reworked and improved to be more intuitive and robust. Efficiency has a mean value of 1.2 (std. dev. 1.092) and a confidence of 0.677. It is again in the category of above average and has a higher standard deviation and confidence. Therefore, the interaction was perceived as requiring less unnecessary effort and reacting fast. Last in the category is dependability which achieved a score of 0.975 (std. dev 0.661) and a confidence of 0.41 which is below average. Dependability in this context is how much the user feels in control of the interaction with the product. This score is too low and needs to be corrected by improving the menus in Teach-R and the coach. The standard deviation and confidence are relatively low, which indicates that the participants match their perception of it. One big impact on this perception might be the experienced bugs, for example, as the visualization menu in Teach-R sometimes immediately closed again after being activated or the robustness of the triggers for the buttons in the menu in Teach-R was not convincing. Overall the pragmatic quality aspects did not perform very well. An interesting factor would be how these aspects would have scored without bugs influencing them. This could be partly answered by the results with the results from the upcoming evaluation from Jasmin Haranto's evaluation.

Perspicuity

Efficiency

Dependability

In contrast to the results from the pragmatic quality aspects, the hedonic aspects, stimulation and novelty, performed far better. The hedonic aspects are related to pleasure, enjoyment, and emotional satisfaction derived from interacting with the product. Stimulation and novelty have both a mean value of 2.175, a similar standard deviation (0.635, 0.678), and a confidence interval size (0.394 and 0.42). Both aspects are categorized to be excellent according to the benchmark intervals for the UEQ scales. How exciting and motivating the product is to use is determined by stimulation, while novelty refers to the innovation and creativity of the product.

Stimulation and Novelty

To conclude the results of the UEQ, the implemented foundation for speech analysis in Teach-R and the coach are useful and might help reflecting the voice and speech, but the handling needs to be reworked.

5.3.2. Open Questions

The results from the open questions were overall positive but with plenty of suggestions for further improvement.

Visualization Menu

Activation

Poster

Toggling and interacting with the visualization menu was less easy than assumed, as participants did not find it intuitive. First of all, using a poster to toggle the menu might be changed. Some participants tried to open the menu by pressing buttons on the controllers and needed a hint on how to activate it. Extending the formulation of the task from "Please open the visualization menu." to "Please open the visualization menu for learning analytics" as a hint already helped the participants, as the poster has the title *Learning Analytics*, but does not improve the intuitiveness. Therefore, redesigning the poster might already be an improvement. One suggestion was to use a whiteboard instead of the poster because it appears more interactive. How to design the whiteboard for this imposes a new challenge, as a blank whiteboard might provide too few clues for a toggle. Adding a canvas to the whiteboard with handwriting shortly discussing learning analytics could be used, but the big play button similar to the poster might be confusing.

Different Metaphors

Exchanging the metaphor by something entirely different, for example, to a book, a tablet, or the whiteboard was suggested as well. As mentioned earlier, the previous metaphor was a book which was changed because the menu hovered above the book, thus, blocking the user's view upon activation. For this, a further evaluation would be needed to compare the results, as both have advantages and disadvantages. Another possibility would be to use a book or tablet and put it on the teacher's desk, but keep the menu on a wall. This might initially lead to confusion for the first use, as the menu might not be seen at first. Using the blackboard poses disadvantages as well, as other features might be disrupted by the menu, for example writing on it. One approach to have both would be to exchange all blackboards in Teach-R for digital whiteboards like the ones already used in a lot of German schools. Those whiteboards could have different tabs, one for writing, one for showing a presentation, and one for the visualization menu.

Detaching the Menu from a Metaphor

Being able to interact with the menu from everywhere in the classroom was also requested by the participants. In order to detach the menu from a static object to be able to interact with it regardless of position, it was proposed to use an UI attached to the camera. Possible drawbacks of this approach could be that it might hinder sight when active, it could reduce the level of immersion, and the interaction has to be done by clicking the buttons on the controller. Further advantages might be that no buttons are required in VR. The menu buttons were criticized for being too small, but increasing the size of the menu might be counter-productive, as some buttons might not be reachable anymore when touch is used and might lead to a visual overload.

Generally speaking, issues regarding the discovery of the metaphor to toggle the visualization menu could be bypassed by introducing learning analytics and the visualizations in the tutorial of Teach-R. Nonetheless, a more adequate solution would be to find a metaphor that is intuitive for the majority of users.

Different Interaction Techniques

Furthermore, the metaphor of touch for interaction was perceived as less intuitive, as mentioned by some participants during the interaction with the menu. In the beginning, the participants were informed about the two possibilities of moving around in Teach-R, which are actual movement and teleporting via the touchpad on the controllers. After testing the teleportation in VR, which features pre-travel visual feedback depicted by a ray, some users tried to use the menu by pointing at it, with and without activating the teleporting ray. The metaphor of interacting by pointing did not need to originate from the teleporting feature but might have biased the participants.

The suggestion to add a title or short description to the icons on the buttons is a relevant comment and this ought to be implemented as well as adding an icon representing the analysis to the microphone for speech analysis. The text on the buttons was removed during the implementation phase, as the icons were assumed to be self-explanatory. However, it turned out to be a misconception.

Fog- and Circular-Volume Visualization

Both, the fog- and circular-volume visualization, seem to be an appropriate way to get an impression of the current volume while in the virtual environment. To address the low frame rate and the delay, the performance should be improved by changing parts of the implementation, for example, by reducing the used arrays to smaller and simultaneously more representative ones. Furthermore, libraries already exist that offer a wider range of functionality and consequently might perform better, but a lot of them are not open-source. The idea of limiting the fog to a certain height or adapting the color of the circle depending on the volume should be further tested and evaluated.
When more visualizations are implemented, the added value of having two visualizations for volume should be considered. At the current state, having both has no notable drawback and both visualizations were expressed as favorites by different participants.

Teacher Clone

Hearing own Voice

The usefulness of the teacher clone was perceived the least, but when looking at possible extensions, potential can be seen. Hearing one's voice might be unfamiliar to most users, but it could certainly help to reflect the session. A lot of participants wondered if their voices sounded like what they had heard. They were told, that the quality of the recording was quite low to improve the performance. Thus, improving the frequency of the recording (currently at 16kHz) could be realized easily, when the performance allows this.

Interacting with the Coach

Various suggestions can be implemented quite easily and make sense, i.e. adding a female and gender-neutral model, hands, and eyes, and improving the posture to make it appear more natural. The interaction should be reworked, as most participants were unable to interact with the clone as intended without help. It was suggested to add (UI) buttons, similar to one possible concept discussed in section 3.3.3. When including new features, the interface needs to be extended anyways which is easier when buttons are used for interactions. Furthermore, the interaction by touching the mesh might be not reasonable, as there is no known metaphor and no cues for interaction can be added easily.

Combining Features with the Clone

Lastly, developing more features to improve the usefulness of the clone were noted. This is in line with the originally planned clone, for example, by adding audio effects to it. The idea to add a spectrum of the volume directly to the clone could prove to be quite challenging, but the possibility of applying visualizations to the clone might be intriguing. For example, when the clone is replaying a recording, the circular-volume visualization could be applied to that and originate from the clone.

Visualization Page in the Coach

Toggle Buttons

Order of Visualizations

To start with discussing the results from the evaluation of the coach, the suggested improvements for the visualization page should all be integrated. The list of visualizations to toggle has no visual cue that it is not mutually exclusive. For instance, a checkbox ought to be added to convey that cue. Currently, the visualizations are always in the same order as the list when activated. One statement was to add newly activated visualizations either above or below already active ones. This would make it easier to compare the results of visualizations when desired. However, if a grid is implemented for dragging-and-dropping visualizations, this might increase redundancy. In the process, a grid could be visualized and used for clearer separation of the respective visualization, as this was mentioned in the evaluation as well.

Further Features

Furthermore, the option to replay the recordings in the coach, especially selecting an interval to look at, might be a good extension. The icons for the audio recording were perceived as misleading, as they show a microphone each, one crossed out diagonally. Combined with the statement that it is unexpected to get an analysis of data in the same plot for different recordings, those two microphones make sense, as the recording could be perceived as muted, not stopped. Overall, changing the icons to the typical recording icons (dot + square) could be a valid change, but the analysis should still be kept in one plot with the option to save and clear the plots at any time.

Synchronizing Updates

Another point from a participant was to synchronize the refresh rates of the visualizations. Due to performance reasons, the refresh rate from the plots is set to once every second, but the interval to call the update is set in different classes. Together, the fact that they use different data packages for the analysis leads to an asynchronous update. This could be fixed by adding the initialization of the intervals in a service, but circular dependency should be regarded. Moreover, consideration should be given to whether the packages from Teach-R to the coach ought to be rather sent independently from each other, i.e. separately according to information, or one package for all information. If the performance is no problem anymore, a complete package with a .wav file snippet for the audio recording could be sent and merged in the coach. The advantage of sending separate packages is that they can be sent whenever new information is processed in Teach-R independently from the other sources.

Recommendations when adding more Visualization

Adding the suggested features, could certainly enhance the experience, as more visualizations are usually better, but it needs to be taken care of not overfitting the needs of an individual. This means, that more visualizations are useful if they add new information or another perspective, but not if they just depict the same information as another one. Additionally, the value of a visualization needs to be evaluated. For example, a spectrogram was already considered in the conceptualization chapter 3.4.2, but it is not easy to understand without any introduction to it. To accommodate this, a short description of the visualization could be included, either individually for every visualization or for the complete visualization page. This description should explain how a visualization is read and how the values are calculated. In a general description, the note should be written that these visualizations are not for rating the performance of the user, but to give a basis to evaluate the performance for themselves or in cooperation with the person operating the coach.

Descriptions for Visualizations

Visualizations

Some critiques for the visualizations would be already responded to by adding an axis label everywhere and giving a short description when hovering over it or when pressing an according button. Further to this, participants wished for some guiding values, for example, for the volume so that they could speak loud enough for students in the last row to hear everything. For volume, this could be achieved by placing an invisible object in the room on the position of the farthest student away and measuring the volume. When done dynamically, this would also give a reference depending on the room, as one statement mentioned. This should not be done for every visualization though, as the speech rate should not encourage the users to use more or fewer words. A lot has been contributed in general to the speech rate visualization regarding its expressiveness. Again, the analyses are not meant to rate a user, but to help them reflect on their lesson and work on characteristics they would like to change. The speech rate analysis could

Guiding Values

be enhanced by including more context in the calculation, for instance, when students speak or disrupt the speech of the teacher.

Design of Visualizations

In order to have a more appealing look, every non-numerical visualization could use the same grid in the background, as a grid was desired for the waveform visualization. If a guiding value is added to a visualization, adding colors to highlight values far below or above that value needs to be discussed. As it would be difficult to consider every exception, sometimes marking something in a flashy color may be a wrong attempt, as speaking more loudly or softly is intended. Instead, less striking colors without rating could be used to mark edge cases, but the added value might be less expressive. Finally, the usage of the same kind of plot for different visualizations was scrutinized. Using the same visualization might improve the understanding of the plots, while different visualizations might emphasize different values. In general, the decision on the type of plot should be based on which plot represents the values best.

5.3.3. Implementation

Further Development of the Visualization Menu

As all bugs found during the evaluation originated in the visualization menu, it is further improved by Jasmin Haranto. Since many participants expressed difficulties interacting with the menu and the reason that she will use the same menu for her master's thesis, this was the first part to be reworked. The design remained the same, but the implementation of the interaction was newly implemented. Originally, the interaction was created without further assistance from a library. Now, the XR UI toolkit is used, which makes the implementation more efficient and the usage more robust.

Interacting with the Visualization Menu

The most suitable way of activating the menu is still in discussion, but the menu is not fixed at a position anymore. Instead, when the technique to trigger the menu is used, the menu is attached to a hand. In more detail, when activated, one hand determines the position of the menu and the other one can be used to interact with it. Two possibilities are obvious and mentioned before, using an object as a metaphor in Teach-R or using a button on the controller. The advantage of the button on the controller would be that depending on which controller the button was clicked, the hand to interact with the menu can be chosen by the user. On the other hand, a button on the controller might interfere with immersion and using a metaphor in the room might be easier to remember for people not using a virtual system and its controller often. Either way, an introduction to learning analysis and the associated visualizations should be made in the tutorial in the future.

5.3.4. Limitations

The results of the evaluation contribute to the estimation of aspects like the usefulness, intuitiveness, and novelty of the implementation. From this, possible improvements can be derived and discussed to counteract the critique of the participants and enhance the experience. However, the validity and expressiveness of the quantitative results have several limitations which will be discussed in this section.

Sample

The first limitation regards the recruiting process and sample of participants. For the evaluation, a sample size of 10 was chosen, as it mainly serves as a proof of concept, i.e. whether the initial feedback leads to further development. Therefore, the qualitative open questions are the main factors to consider, as they are not influenced by the number of participants. In contrast to this, the quantitative measure, i.e. the UEQ, depends on a certain number of participants. With this number of participants, the precision is around 0.5 and the error probability is between 0.1 to 0.05 (values taken from the data analysis tool provided by the UEQ website). The precision in this context describes the deviation between the true scale mean in the population and the estimated scale mean from the sample, while the error probability is the chance that the true scale mean is outside the area described by the error bars. For both units, a small value is desirable. Furthermore, most participants are friends of the conductor and might have decorated the feedback. To avoid this, the data from the questionnaire was gathered anonymously and they were not observed while filling out. Additionally, in the introduction text, the participants were explicitly asked to answer the questions truthfully.

Target Group of Teach-R

The target group of Teach-R are teachers-to-be, but only two participants studied to be teachers. This limitation might seem worse than it actually is because the usefulness of the features does not depend on knowledge acquired in the studies and the two participants who did study to be teachers mentioned that voice and speech were not much of a topic during their studies. Finally, some participants already know their way around Teach-R, for example, from their own projects with it or through participation in the course where the basis for learning analytics was integrated. Those participants were chosen deliberately, as aspects such as consistency for the newly added features or usefulness in the context of Teach-R should be considered as well. In more detail, as the evaluation only considered a small part of Teach-R and how the added speech analysis could be used, the perspective of how these features contribute to Teach-R in general and whether the integration of such makes sense to develop further. In conclusion, those participants with prior knowledge of Teach-R could help to evaluate the implementation in context, but on the other hand, that might have biased the feedback to compare it to existing features.

Performance and Bugs

As already mentioned, the performance hindered the experience of the features, as it influences the synchronization of visualizations as well as the smoothness of the animations. Consequently, the execution of the evaluation was changed after the first and second run-through. This and the occurred bugs probably influenced the results of the participants but did not make them unusable because the basic functionality worked for all participants and valuable insights could be gathered nonetheless.

Chapter 6 Discussion & Conclusion

In this final chapter, this thesis is framed within the scientific context and further summarized. This is done by summarizing and concluding in section 6.1 and discussing the general limitations of the results in section 6.1.1. Finally, improvements, further features, and scientific topics to look at for future work are presented in section 6.2.

6.1. Conclusion

The original goal of this thesis was to implement voice analysis into the application Teach-R, a virtual system for teachers-to-be to train their classroom management. The idea and concept of voice screening was inspired by the existing program from the Chair for Contemporary German Language [Sti]. As teachers use their voice on a daily basis, even up to two-thirds of a lesson [Law77], many suffer from problems with their voices.

Related Work

To introduce the topic of voice analysis, the creation of sound was presented in chapter 2, and based on this the analysis of the voice was discussed. Furthermore, as some aspects of the analyses are similar and the way teachers convey information is important for learning (cf. [Ure12]), speech analysis was added to the topics of interest. Drawing from this, several possible parameters can be analyzed. Taking into consideration what is possible and what is interesting for teachers, possible analyses and visualizations are discussed in chapter 3. For this, some further research is done on visualizations, in more detail, what are some guidelines and potential plots for both, the two-dimensional and three-dimensional environment.

Conceptualization

Challenges

Moving forward to the implementation phase, several challenges appeared. Unity only provides rudimentary functionality for working with audio recordings. Thus, the analyses were created from scratch, which heavily influenced the schedule for this phase. This and other restrictions are presented in 3.1 and are included again in section 6.1.1. Despite these challenges, the added features are analyses and visualizations in both applications, Teach-R and the coach as described in chapter 4. In Teach-R, the previous menu for toggling visualizations has been reworked by moving it to another place, giving it another trigger, and expanding it. From this, three visualizations can be used, namely a visualization representing the current volume by a scaling circle sprite at the feet of the user's model. Similar to this, the volume can be investigated with a fog around the user that lifts if the volume increases. Finally, a clone of the teacher can be activated to replay the recorded audio of the user, to experience the session from the perspective of a student.

Implementation in Teach-R

Implementation in the Coach

The recorded data and the newly implemented speech-to-text results are sent to the coach for additional analyses to visualize them similar to a professional voice screening. For this, a new page was created in the coach for these and future visualizations. Currently, the results for speech-to-text are listed, and based on this, an average speech rate and the total word count are calculated. In addition, a course of speech rate for time segments of three seconds is shown to compare the rate over time. Similar to the visualizations in Teach-R, a line plot for volume over time is included. Finally, a waveform similar to professional voice screening is displayed.

Evaluation

To evaluate the implementation on aspects such as usability, functionality, and helpfulness, an evaluation was conducted, discussed and presented in chapter 5. A group of ten participants tested every implemented feature and were asked to give feedback through a questionnaire at the end, including a rating of the usability using the User Experience Questionnaire [Use]. The functionality was tested indirectly by using the features and comments for the state of the implementation as well as suggestions for improvement of the implementation were given by open questions in the questionnaire.

Through the evaluation, bugs in the visualization menu were detected. Moreover, the results of the questionnaire can be interpreted as the implemented features being helpful to reflect on the usage of the voice, but need some refinement for a pleasant experience. Many suggestions were given to improve visualizations or which visualizations could be added, with the critique, that some visualizations were perceived as less helpful. According to the participants' comments, this is the case because the single visualizations are less expressive but helpful when combined with others.

Comparing what was planned at the beginning and mentioned in both the related work and conceptualization, only a small part was implemented. This can be attributed to two facts. On the one hand, the goals were very ambitious and, on the other hand, the mentioned challenges further hindered the process. However, a foundation for voice and speech analysis was implemented and concepts for future work were presented, which will be discussed in section 6.2.

Placement in Scientific Context

This can be broadly sorted into the scientific context of two areas: virtual reality, classroom simulation, and voice and speech analysis. Considering the current state of knowledge, no combination of these domains is known. Examples of voice or speech analysis are done in the related work chapter, for example, the voice screening at the RWTH Aachen University. A domain where the analyses are carried out in a virtual environment is public speaking, which has several similarities to a classroom, but next to frontal lectures, teachers need to moderate discussions and engage with students in general. Therefore, the added functionality to Teach-R offers teachers-to-be a new possibility of training their classroom management in a simulation and reflecting on their usage of voice and speech.

Achieved Objective of this Thesis

Taking the working features and the results of the evaluation into account, the goal of implementing voice and speech analysis into Teach-R is achieved. Even though only the foundation is implemented, the features seem to help the users to reflect on their usage of the voice and the way they talk. Additionally, a person controlling the coach has real-time data to either make notes for later discussion or give feedback while the user is in the virtual environment.

6.1.1. Limitations

As the limitations are similar to the restrictions discussed in section 3.1 which provides a summary as well as further aspects that came up during the development. In contrast to the limitation section in the evaluation chapter, this section refers to the limitations of this complete thesis. Future work is based on the positive results from the evaluation, which reveal the helpfulness of the analyses in Teach-R and the coach. However, the limitations discussed in section 5.3.4 need to be kept in mind, especially the potentially biased and small sample.

Performance

The final product of this thesis poses a foundation for voice and speech analysis in Teach-R, but for using the product, some aspects should be considered. First of all, the new features affect the performance to such an extent that the frame rate is reduced when used in the virtual environment. When the synchronous speech-to-text model is deactivated, the frame rate is increased to the state of Teach-R before the features are added. Thus, this feature and the according analyses can only be used in a restricted manner.

Microphone Quality

Since the audio is planned to be recorded with the integrated microphone of the head-mounted display, the accuracy and hence the validity of the data source is a limitation. This problem relativizes itself to some extent, as the goal is to give the users a rough estimate of their voice which still is able on the given quality of data. A similar problem is the results of the speech-to-text results. Not only are they dependent on the quality of the audio recording, but the model used for the transcription is not optimized for German. With this at hand, further analysis of the recognized text is limited as well. Therefore, it should be considered in future work to either train a model or search for a better model that can be used in combination with the Whispher package for Unity.

Model for Speech-To-Text

Required Background Knowledge

As indicated by the evaluation of the waveform visualization, the visualizations can require certain background knowledge for interpretation. No professional voice coach will most likely control the coach for Teach-R. Hence, the usefulness of some visualizations used in professional voice screening needs to be considered. To some extent, descriptions of the visualizations can help the users interpret them, but in future work, a study on non-trained users working visualizations like a spectrogram should take place.

Interdependency of Visulizations

Finally, the comments the participants of the evaluation remarked on the usefulness of some visualizations if used individually. Either the individual visualizations should be adapted to be independently expressive, or they should only be shown together.

6.2. Future Work

This section discusses improvements that should be considered when continuing work on this research. Moreover, future applications and features that can extend the product are discussed. In the related work chapter (2) and the conceptualization chapter (3), several ideas were presented that were not implemented in the course of this thesis. Those can be taken as a good start for further developing voice and speech analysis in Teach-R.

6.2.1. Improvements of the Product

Potential sections to work on are already addressed in section 3.1, the results of the evaluation 5.2, and here in the limitations. Now, concrete approaches are discussed for some fundamental features. For example, the design of the visualization page in the coach could be revised following the suggestions of the participants of the evaluation, but is not further discussed to focus more on the pragmatic functionality.

Improving Performance

To improve the performance, some refinements for the speech-to-text feature can be made. At the moment, the transcription is done in real-time. On the cost of live results, a delay could be introduced for this, which could enhance the performance and the accuracy of the results. As already attempted, the model could be done asynchronously or in another thread by third-party solutions, because Unity works on one big main thread. Handing the model larger packages or using more complex or more specialized models could improve the results as well.

Visualization Menu

Based on the evaluation done the visualization menu is already reworked in the interaction, as mentioned before. Through this, a more robust way of touch is established and the poster as a trigger is removed. This adaption will be tested as well to get feedback on the intuitiveness.

Validity of Analyses

The analysis of the volume in Unity is done from scratch which influences the values as discussed before. To get more realistic values, the cause for the low values needs to be discovered. One workaround would be to use other software to measure the decibel of the recording and find out the scaling factor. This approach still has limited validity, but estimations on the volume can be improved. With this, the scaling of the volume visualizations can be made more accurate. Another approach mentioned before would be to use another recorder in the scene and calculate the distance of the volume. However, having the loudness in decibels ought to remain the preferred option, as it can be used in combination with empirical data to depict the range of acoustically understandable speech instead of the absolute range of volume.

Package Format

Sending individual packages for the different analyses also should be considered. One package in an uniform interval might increase the performance, but, for instance, the speech-to-text results depend on the speed of the model. Furthermore, the package for the voice analysis currently consists of the amplitude value. In connection with compression tools, using WAV files to send the audio might prove to be beneficial. This could be particularly advantageous when using a third-party analysis tool on audio in the angular project. Either way, a WAV file can be constructed from the amplitude data.

6.2.2. Future Research and Extensions

To discuss possible extensions, those that were introduced in the related work and conceptualization should be added. Additionally, some completely new aspects can be more interesting as well, which will be addressed here.

Analyses on Connected Speech

As outlined earlier in section 2.3.2, it is necessary to delve deeper into speech analysis on connected speech. In the current state of the added features, only the volume is analyzed and a waveform is given to look at irregularities in the voice. Analyses conducted like of airflow patterns in [GMH19], pitch in [Fou09], or frequencies by Fraile et al. in [FGS+13] are already possible on connected speech. The validity of those with the setup used for Teach-R needs to be evaluated in a future study. Additionally, research in existing literature as well as experiments could be done to add more analyses to Teach-R like a voice range profile on connected speech. Moreover, a longitudinal study should take place to investigate, whether using Teach-R in combination with voice and speech analysis improves the usage of the voice and prosodic aspects.

Designated Room for Analyses

Another intriguing approach for this is creating a virtual room, especially for conducting voice screening. Leaving the accuracy of the microphone and the environment of the virtual system aside, a room for this would have the advantage of actively performing the screening. Similar to the tutorial of Teach-R, a computer-controlled character could guide the participant through the analysis. Thus, the voice screening offered by the Chair of Contemporary German Language could be imitated. The room could be equipped with information material, such as posters, to inform the users on certain aspects of voice production. It ought to be considered, whether this should be integrated into Teach-R or if this should remain a stand-alone product, as this has clear separations from classroom management. This application can be inspired by the virtual simulations for training public speaking presented in section 2.6, for example, introduced by [ERGM19].

Analyzing Non-Verbal Communication

Tracking hand gestures and combining them with recognized meaning could be interesting for analyzing the conveyed message, as was similarly done in [IHC+23] or [Cie23]. More general, non-verbal communication, if practiced by teachers, has a positive and profound effect on students' mood (cf. [BS17]). Jackob et al. [JRP11] found that spoken content has a more substantial impact on speech. Nevertheless, emphasis and gestures can improve certain aspects of speech, such as liveliness and power. The basis for this analysis is already implemented, as the tracking of the hands was already used in the research focus class in 2023. From this, an attempt could be made to anticipate the hand and arm position.

Text-To-Speech

Using AI

At the current state of Teach-R, the coach offers a limited choice of verbal topics. To enhance the vividness of Teach-R and with this improve the basis for speech analysis, the communication of the students could be improved. In the master thesis by Mattiussi [Mat23] a package was implemented to integrate OpenTTS [Han24] into Unity. With this, the person controlling the coach could output questions or answers via text in Teach-R to simulate realistic teaching scenarios with more specific interactions. Having this at hand, for example, the initiation-response-feedback pattern could be trained and analyzed with speech analysis. Based on this, the simulation could be even further improved by artificial intelligence that gets a prompt from the person controlling the coach to ask a question regarding a topic and autonomously create a question. Ideally, the complete behavior could be controlled by the students, i.e. instead of giving the students a short-term behavior, they could get a behavior profile for that session.

Acoustics

An interesting combination of topics would be acoustics and computer science, like in this thesis of speech science and computer science. The acoustic properties of Teach-R are already adapted to be more realistic. Research in combination with acoustic could further simulate the experience of speaking in a crowded or large room, for example, a lecture hall. Currently, a recording of a classroom is played in a loop in Teach-R. To improve this, several single recordings could be played and mixed for a more diverse experience. This mix can be combined with disruptive factors as mentioned in the introduction 1.1, for instance, people talking with different loudness, people rummaging, or someone walking in the hallway. Furthermore, this research could improve the scaling circle visualization to the original idea to represent a realistic propagation of sound in the room.

Appendix A Bibliography

[AAA18] Ahmed A. Alrahim, Rawan A. Alanazi, and Mohammad H. Al-Bar. Hoarseness among school teachers: A cross-sectional study from Dammam. J Family Community Med, 25(3):205–210, 2018.

[AFA06] Aristoteles, John Henry Freese, and Aristoteles. Aristotle. 22: The "art" of rhetoric / with an English transl. by John Henry Freese. Number 193 in The Loeb classical library. Harvard Univ. Press, Cambridge, Mass, repr edition, 2006.

[AGDM20] José M. Arana, Fernando Gordillo, Jeannete Darias, and Lilia Mestas. Analysis of the efficacy and reliability of the Moodies app for detecting emotions through speech: Does it actually work? Computers in Human Behavior, 104:106156, March 2020.

[ALAS14] R. Akinbode, K. B. H. Lam, J. G. Ayres, and S. Sadhra. Voice disorders in Nigerian primary school teachers. Occupational Medicine, 64(5):382–386, July 2014.

[Ana] The Analysis & Resynthesis Sound Spectrograph. https://arss.sourceforge.net/.

[Anga] Angebote für Lehramtsstudierende - Universität Regensburg. https://www.uni-regensburg.de/zentrum-sprache-kommunikation/mkuse/kursangebot/angebote-fuer-lehramtsstudierende/index.html.

[Angb] Angular. https://angular.io/.

[Anw23] Anwendungsszenarien · Wiki · mm_vr / VR-Klassenzimmer · GitLab. https://gitup.uni-potsdam.de/mm_vr/vr-klassenzimmer/-/wikis/Anwendungsszenarien, August 2023.

[ARQ+20] Amal Aljabri, Daniah Rashwan, Rawan Qasem, Rola Fakeeh, Rehab Albeladi, and Najla Sassi. Overcoming Speech Anxiety Using Virtual Reality with Voice and Heart Rate Analysis. In 2020 13th International Conference on Developments in eSystems Engineering (DeSE), pages 311–316, Liverpool, United Kingdom, December 2020. IEEE.

[Ash21] Julian Ashbourn. The Physics of Sound. In Julian Ashbourn, editor, Audio Technology, Music, and Media: From Sound Wave to Reproduction, pages 15–19. Springer International Publishing, Cham, 2021.

[AWNN20] Antoinette Am Zehnhoff-Dinnesen, Bozena Wiskirska-Woznica, Katrin Neumann, and Tadeus Nawka, editors. Fundamentals - Voice Disorders - Disorders of Language an Hearing Development. Number 1 in Phoniatrics. Springer, Berlin [Heidelberg], 2020.

[Bar13] B. Barsties. Einfluss verschiedener Methoden zur Bestimmung der mittleren Sprechstimmlage. HNO, 61(7):609–616, July 2013.

[BB14] Meike Brockmann-Bauser and Jörg E. Bohlender. Praktische Stimmdiagnostik: theoretischer und praktischer Leitfaden ; [+online-Extras]. Forum Logopädie. Thieme, Stuttgart New York, 2014.

[BDB15] Ben Barsties and Marc De Bodt. Assessment of voice quality: Current state-of-the-art. Auris Nasus Larynx, 42(3):183–188, June 2015.

[BGB07] Maria Cristina Borrego, Gisele Gasparini, and Mara Behlau. The Effects of a Specific Speech and Language Training Program on Students of a Radio Announcing Course. Journal of Voice, 21(4):426–432, July 2007.

[BH97] B. Boyanov and S. Hadjitodorov. Acoustic analysis of pathological voices. A voice analysis system for the screening of laryngeal diseases. IEEE Engineering in Medicine and Biology Magazine, 16(4):74–82, July 1997.

[BRZ+22] Tom Bäckström, Okko Räsänen, Abraham Zewoudie, Pablo Pérez Zarazaga, Liisa Koivusalo, Sneha Das, Esteban Gómez Mellado, Mariem Bouafif Mansali, and Daniel Ramos. Introduction to Speech Processing: 2nd Edition. July 2022.

[BS17] FATEMEH BAMBAEEROO and NASRIN SHOKRPOUR. The impact of the teachers' non-verbal communication on success in teaching. J Adv Med Educ Prof, 5(2):51–59, April 2017.

[Cha] Chart.js. https://www.chartjs.org/.

[Cie23] Alan Cienki. Speakers' Gestures and Semantic Analysis. Cognitive Semantics, 9:167–191, June 2023.

[CM19] Johnny C.-H. Chui and Estella P.-M. Ma. The Impact of Dysphonic Voices on Children's Comprehension of Spoken Language. J. Voice, 33(5), September 2019.

[COGM21] Rafaella Cristina Oliveira, Ana C. C. Gama, and Max D. C. Magalhães. Fundamental Voice Frequency: Acoustic, Electroglottographic, and Accelerometer Measurement in Individuals With and Without Vocal Alteration. Journal of Voice, 35(2):174–180, March 2021.

[DCPRC12] Victor Da Costa, Elizabeth Prada, Andrew Roberts, and Seth Cohen. Voice Disorders in Primary School Teachers and Barriers to Care. Journal of Voice, 26(1):69–76, January 2012.

[dGF19] Ehrlson de Sousa, H. C. Goel, and Vinson Louis Gonzaga Fernandes. Study of Voice Disorders Among School Teachers in Goa. Indian J. Otolaryngol. Head Neck Surg., 71:679–683, October 2019.

[DKI23] Teodora Dimitrova-Grekow, Aneta Klis, and Magdalena Igras-Cybulska. Speech Emotion Recognition Based on Voice Fundamental Frequency. Archives of Acoustics, August 2023.

[EP22] F. Alton Everest and Ken C. Pohlmann. Master Handbook of Acoustics. McGraw-Hill Education, New York, 7th ed edition, 2022.

[ERGM19] Meriem El-Yamri, Alejandro Romero-Hernandez, Manuel Gonzalez-Riojo, and Borja Manero. Designing a VR game for public speaking based on speakers features: A case study. Smart Learning Environments, 6(1):12, November 2019.

[eri23] eric-urban. About the Speech SDK - Speech service - Azure AI services. https://learn.microsoft.com/en-us/azure/ai-services/speech-service/speech-sdk, July 2023.

[FGS+13] Rubén Fraile, Juan Ignacio Godino-Llorente, Nicolás Sáenz-Lechón, Víctor Osma-Ruiz, and Juana María Gutiérrez-Arriola. Characterization of Dysphonic Voices by Means of a Filterbank-Based Spectral Analysis: Sustained Vowels and Running Speech. Journal of Voice, 27(1):11–23, January 2013.

[Fon] Font Awesome. https://fontawesome.com.

[Fou09] Adrian Fourcin. Aspects of Voice Irregularity Measurement in Connected Speech. Folia Phoniatr Logop, 61(3):126–136, 2009.

[FTG+18] Joana Fernandes, Felipe Teixeira, Vitor Guedes, Arnaldo Junior, and João Paulo Teixeira. Harmonic to Noise Ratio Measurement - Selection of Window and Length. Procedia Computer Science, 138:280–285, 2018.

[GHS23] Sergej Görzen, Birte Heinemann, and Ulrik Schroeder. Ein Konzept zur Evaluierung eines Ökosystems für die Integration von Learning Analytics in Virtual Reality. 2023.

[Gita] GitHub - ggerganov/whisper.cpp: Port of OpenAI's Whisper model in C/C++. https://github.com/ggerganov/whisper.cpp.

[Gitb] GitHub - Macoron/whisper.unity: Running speech to text model (whisper.cpp) in Unity3d on your local machine. https://github.com/Macoron/whisper.unity.

[GMH19] Marina Gilman, Carissa Maira, and Edie R. Hapner. Airflow Patterns of Running Speech in Patients With Voice Disorders. Journal of Voice, 33(3):277–283, May 2019.

[Hal22] Atilla Hallsby. Reading Rhetorical Theory. University of Minnesota Libraries Publishing, August 2022.

[Han24] Michael Hansen. Synesthesiam/opentts, March 2024.

[HS23] Birte Heinemann and Ulrik Schroeder. Computer Science Teachers in VR Training: Improve Classroom Management (with Learning Analytics). In 2023 9th International Conference of the Immersive Learning Research Network (iLRN), page 8, San Luis Obispo, CA, USA, 2023.

[IHC+23] Magdalena Igras-Cybulska, Daniela Hekiert, Artur Cybulski, Slawomir Tadeja, Marcin Witkowski, Konrad Nakonieczny, Izabela Augustyn, Jan Jasinski, Daria Hemmerling, Tomasz Skrzek, Stanislaw Kacprzak, Magdalena Kaczorowska, Julia Juros, Marek Warzeszka, Ewa Migaczewska, Filip Malawski, Paulina Slomka, Rafal Salamon, Aleksandra Szumiec, Kinga Kornacka, Bartlomiej Blaszczyński, Katarzyna Blaszczyńska, Radoslaw Sterna, Marek Makowiec, and Magdalena Majdak. Towards Multimodal VR Trainer of Voice Emission and Public Speaking -Work-in-Progress. In 2023 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW), pages 355–359, March 2023.

[Inn21] Innovator in Voice AI audEERING Home, September 2021.

[JRP11] Nikolaus Jackob, Thomas Roessing, and Thomas Petersen. The effects of verbal and nonverbal elements in persuasive communication: Findings from two multi-method experiments. In Communications, volume 36, January 2011.

[KAAA15] Gustavo Polacow Korn, Antonio Augusto de Lima Pontes, Denise Abranches, and Paulo Augusto de Lima Pontes. Hoarseness and Risk Factors in University Teachers. Journal of Voice, 29(4):518.e21–518.e28, July 2015.

[LAE23] Mei Hui Lim, Vahid Aryadoust, and Gianluca Esposito. A meta-analysis of the effect of virtual reality on reducing public speaking anxiety. Curr Psychol, 42(15):12912–12928, May 2023.

[Law77] Peter O. Lawson. Review of Towards an Analysis of Discourse: The English Used by Teachers and Pupils. TESOL Quarterly, 11(2):203–206, 1977.

[LBL+21] Benjamin Lee, Dave Brown, Bongshin Lee, Christophe Hurter, Steven Drucker, and Tim Dwyer. Data Visceralization: Enabling Deeper Understanding of Data Using Virtual Reality. IEEE Trans. Visual. Comput. Graphics, 27(2):1095–1105, 2021.

[LHS08] Bettina Laugwitz, Theo Held, and Martin Schrepp. Construction and Evaluation of a User Experience Questionnaire. In USAB 2008, volume 5298, pages 63–76, November 2008.

[Mac23] Ian R. A. MacKay. Phonetics and Speech Science. Cambridge University Press, Cambridge New York, NY Port Melbourne New Delhi Singapore, first published edition, 2023.

[Mat23] Sam Luc Caroll Mattiussi. Enhancing learner assistance in RePiX VR. Master's thesis, RWTH Aachen University, October 2023.

[MC19] Brian B. Monson and Jacob Caravello. The maximum audible low-pass cutoff frequency for speech. J Acoust Soc Am, 146(6):EL496, December 2019.

[McC17] Daniel J McCabe. Prosody: An Overview and Applications to Voice Therapy. GJO, 7(4), May 2017.

[MCFB91] C. MELODY CARSWELL, SYLVIA FRANKENBERGER, and DONALD BERNHARD. Graphing in depth: Perspectives on the use of three-dimensional graphs to represent lower-dimensional data. Behaviour & Information Technology, 10(6):459–474, November 1991.

[MFB05] Reny Medrado, Leslie Piccolotto Ferreira, and Mara Behlau. Voice-over: Perceptual and Acoustic Analysis of Vocal Features. Journal of Voice, 19(3):340–349, September 2005.

[Mid20] Stephen R. Midway. Principles of Effective Data Visualization. Patterns, 1(9):100141, December 2020.

[MMB17] Eva Mailänder, Lea Mühre, and Ben Barsties. Lax Vox as a Voice Training Program for Teachers: A Pilot Study. Journal of Voice, 31(2):262.e13–262.e22, March 2017.

[Mmv23] mm_vr / VR-Klassenzimmer · GitLab. https://gitup.uni-potsdam.de/mm_vr/vr-klassenzimmer, November 2023.

[MSD14] Vikas Mittal, Yuvraj Sharma, and Dixit. Voice Parameter Analysis for the disease detection. IOSRJECE, 9(3):48–55, 2014.

[NSE+20] Manfred Nusseck, Claudia Spahn, Matthias Echternach, Anna Immerz, and Bernhard Richter. Vocal Health, Voice Self-concept and Quality of Life in German School Teachers. J. Voice, 34(3):488.e29, May 2020.

[NW00] H Nassaji and G Wells. What's the use of 'triadic dialogue'?: An investigation of teacher-student interaction. Applied Linguistics, 21(3):376–406, September 2000.

[Pra] Praat: Doing Phonetics by Computer. https://www.fon.hum.uva.nl/praat/.

[Pro12] Prosody and Meaning. In Prosody and Meaning. De Gruyter Mouton, December 2012.

[PSSB10] Sona Patel, Klaus R. Scherer, Johan Sundberg, and Eva Björkner. Acoustic markers of emotions based on voice physiology. In Proc. SpeechProsody 2010, page paper 865, 2010.

[Rai] Rainbow Passage | Speech and Language Professionals. https://www.rit.edu/ntid/slpros/media/rainbow.

[Rec23] Jona Recker. Visualisation of Data in a VR Dashboard. Paper, RWTH, June 2023.

[Roy03] Nelson Roy. Functional dysphonia. Current Opinion in Otolaryngology & Head and Neck Surgery, 11(3):144, June 2003.

[SBCWI18] Siti Salwa Salleh, Azlina Bujang, Ketty Chachil, and Wan Ana Zuriana Wan Ismail. Analysis of Prominent Malay Da'i Voices Frequency and Characteristics. In 2018 8th IEEE International Conference on Control System, Computing and Engineering (ICCSCE), pages 26–30, November 2018.

[SBP+10] Renée Speyer, Hans C.A. Bogaardt, Valéria Lima Passos, Nel P.H.D. Roodenburg, Anne Zumach, Mariëlle A.M. Heijnen, Laura W.J. Baijens, Stijn J.H.M. Fleskens, and Jan W. Brunings. Maximum Phonation Time: Variability and Reliability. Journal of Voice, 24(3):281–284, May 2010.

[Sha08] Tina Sharpe. How can teacher talk support learning? Linguistics and Education, 19(2):132–148, June 2008.

[SHT14] Martin Schrepp, Andreas Hinderks, and Jörg Thomaschewski. Applying the User Experience Questionnaire (UEQ) in Different Evaluation Scenarios. pages 383–392, June 2014.

[SM77] Susan J. Shanks and Donald Mast. Maximum Duration of Phonation: Objective Tool for Assessment of Voice. Percept Mot Skills, 45(3_suppl):1315–1322, December 1977.

[SODH14] Katherine Sanchez, Jennifer Oates, Georgia Dacakis, and Eva B. Holmberg. Speech and voice range profiles of adults with untrained normal voices: Methodological implications. Logopedics Phoniatrics Vocology, 39(2):62–71, July 2014.

[Sti] Stimmscreening - RWTH AACHEN UNIVERSITY DSG - Deutsch. https://www.dsg.rwth-aachen.de/cms/dsg/Studium/~ldpxs/Stimmscreening/.

[SW97] Wolfram Seidner and Jürgen Wendler. Die Sängerstimme: Phoniatrische Grundlagen Der Gesangsausbildung. Henschel Verlag, Berlin, 3., erw. aufl edition, 1997.

[SW03] Rosemary E. Sutton and Karl F. Wheatley. Teachers' Emotions and Teaching: A Review of the Literature and Directions for Future Research. Educational Psychology Review, 15(4):327–358, December 2003.

[TEC+20] Cristian Tejedor-Garcia, David Escudero-Mancebo, Enrique Camara-Arenas, Cesar Gonzalez-Ferreras, and Valentin Cardenoso-Payo. Assessing Pronunciation Improvement in Students of English Using a Controlled Computer-Assisted Pronunciation Tool. IEEE Trans. Learning Technol., 13(2):269–282, April 2020.

[TPS16] Sten Ternström, Peter Pabon, and Maria Södersten. The Voice Range Profile: Its Function, Applications, Pitfalls and Potential. Acta Acustica united with Acustica, 102(2):268–283, March 2016.

[Tri17] Baiba Trinite. Epidemiology of Voice Disorders in Latvian School Teachers. J. Voice, 31(4):508.e1, July 2017.

[Uni] Unity Real-Time Development Platform | 3D, 2D, VR & AR Engine. https://unity.com.

[Ure12] Roxana Urea. The Influence of the Teacher's Communication Style on the Pupil's Attitude Towards the Learning Process. Procedia - Social and Behavioral Sciences, 47:41–44, January 2012.

[Use] User Experience Questionnaire (UEQ). https://www.ueq-online.org/.

[vCWv12] Evelyne van Houtte, Sofie Claeys, Floris Wuyts, and Kristiane van Lierde. Voice disorders in teachers: Occupational risk factors and psycho-emotional factors. Logopedics Phoniatrics Vocology, 37(3):107–116, October 2012.

[vFL+00] A. van Dam, A.S. Forsberg, D.H. Laidlaw, J.J. LaViola, and R.M. Simpson. Immersive VR for scientific visualization: A progress report. IEEE Computer Graphics and Applications, 20(6):26–52, November 2000.

[WBM+00] Floris L. Wuyts, Marc S. De Bodt, Geert Molenberghs, Marc Remacle, Louis Heylen, Benoite Millet, Kristiane Van Lierde, Jan Raes, and Paul H. Van De Heyning. The Dysphonia Severity Index: An Objective Measure of Vocal Quality Based on a Multiparameter Approach. J Speech Lang Hear Res, 43(3):796–809, June 2000.

[Wei07] Johannes Weiß. Chronische Heiserkeit kann oft erfolgreich behandelt werden. Dtsch Med Wochenschr, 132(28/29):p13–p13, July 2007.

[WEV] WEVOSYS Home English. https://www.wevosys.com/index.html.

[WHLS21] Axel Wiepke, Birte Heinemann, Ulrike Lucke, and Ulrik Schroeder. Jenseits des eigenen Klassenzimmers: Perspektiven & Weiterentwicklungen des VR-Classroo
Gesellschaft für Informatik e.V., 2021.

[Wil19] Claus O. Wilke. Fundamentals of Data Visualization: A Primer on Making Informative and Compelling Figures. "O'Reilly Media, Inc.", March 2019.

[WOH+06] D.A. Wiegmann, T.J. Overbye, S.M. Hoppe, G.R. Essenberg, and Yan Sun. Human factors aspects of three-dimensional visualization of power system information. In 2006 IEEE Power Engineering Society General Meeting, page 7 pp., Montreal, Que., Canada, 2006. IEEE.

Appendix B Digital Appendix

Additionally, a digital appendix is included for this master thesis in the form of a CD on the last page. The digital appendix includes:

- The written thesis as PDF
- Implementation Teach-R
- Implementation coach
- A blank version of the questionnaire
- Results of the questionnaire (PDF, csv, xlsx)
- Screenshots and videos from the visualizations

Appendix C UEQ Results

Evaluation UEQ Results

unerfreulich | erfreulich	6	6	5	5	7	5	6	5	6	7
unverständlich | verständlich	5	2	6	4	6	5	6	3	7	6
kreativ | fantasielos	2	2	4	1	1	2	1	3	2	1
leicht zu lernen | schwer zu lernen	2	5	1	1	3	6	1	4	1	2
wertvoll | minderwertig	2	3	3	1	3	3	1	3	2	1
langweilig | spannend	7	6	6	7	7	5	7	7	5	7
uninteressant | interessant	7	7	5	6	7	5	6	6	6	7
unberechenbar | voraussagbar	6	3	4	5	2	5	5	4	5	5
schnell | langsam	3	2	2	4	1	4	3	6	6	3
originell | konventionell	2	4	2	1	2	2	1	2	1	1
behindernd | unterstützend	6	5	4	7	7	4	6	5	6	7
gut | schlecht	2	2	2	1	1	2	1	3	1	1
kompliziert | einfach	4	3	5	6	2	4	6	1	7	6
abstoßend | anziehend	6	5	5	7	5	5	6	4	6	6
herkömmlich | neuartig	7	6	6	6	6	5	7	6	6	7
unangenehm | angenehm	4	7	5	7	2	5	6	4	5	6
sicher | unsicher	2	1	2	6	3	4	3	4	4	2
aktivierend | einschläfernd	2	2	2	1	1	3	1	3	1	1
erwartungskonform | nicht erwartungskonform	2	5	4	2	3	3	3	5	2	2
ineffizient | effizient	5	4	5	6	6	2	6	4	4	6
übersichtlich | verwirrend	4	5	2	6	3	3	5	5	3	3
unpragmatisch | pragmatisch	5	6	6	7	7	4	7	3	5	6
aufgeräumt | überladen	1	1	1	4	3	2	1	5	2	2
attraktiv | unattraktiv										
sympathisch | unsympathisch	2	4	2	2	4	2	1	4	1	2
konservativ | innovativ	6	6	5	7	7	5	7	5	7	7

Table C.1.: Results from the individual items of the UEQ per participant.

Mean	Variance	Std. Dev.	Left	Right	Scale
1,8	0,6	0,8	unerfreulich	erfreulich	Attraktivität
1,0	2,4	1,6	unverständlich	verständlich	Durchschaubarkeit
2,1	1,0	1,0	kreativ	phantasielos	Originalität
1,4	3,4	1,8	leicht zu lernen	schwer zu lernen	Durchschaubarkeit
1,8	0,8	0,9	wertvoll	minderwertig	Stimulation
2,4	0,7	0,8	langweilig	spannend	Stimulation
2,2	0,6	0,8	uninteressant	interessant	Stimulation
0,4	1,4	1,2	unberechenbar	voraussagbar	Steuerbarkeit
0,6	2,7	1,6	schnell	langsam	Effizienz
2,2	0,8	0,9	originell	konventionell	Originalität
1,7	1,3	1,2	behindernd	unterstützend	Steuerbarkeit
2,4	0,5	0,7	gut	schlecht	Attraktivität
0,4	3,8	2,0	kompliziert	einfach	Durchschaubarkeit
1,5	0,7	0,8	abstoßend	anziehend	Attraktivität
2,2	0,4	0,6	herkömmlich	neuartig	Originalität
1,1	2,3	1,5	unangenehm	angenehm	Attraktivität
0,9	2,1	1,4	sicher	unsicher	Steuerbarkeit
2,3	0,7	0,8	aktivierend	einschläfernd	Stimulation
0,9	1,4	1,2	erwartungskonform	nicht erwartungskonform	Steuerbarkeit
0,8	1,7	1,3	ineffizient	effizient	Effizienz
0,1	1,7	1,3	übersichtlich	verwirrend	Durchschaubarkeit
1,6	1,8	1,3	unpragmatisch	pragmatisch	Effizienz
1,8	2,0	1,4	aufgeräumt	überladen	Effizienz
			attraktiv	unattraktiv	Attraktivität
1,6	1,4	1,2	sympathisch	unsympathisch	Attraktivität
2,2	0,8	0,9	konservativ	innovativ	Originalität

Table C.2.: Items of the UEQ with associated category, mean, variance, and standard deviation

Appendix D Evaluation Procedure

Evaluation Procedure

Protocol

Vorbereitung

- Neuste Version von Teach-R und Coach pullen (Branch: MA_Jona)
- Falls neue Version: Einmal alles Testen: Aufnahme, Wiedergabe, Visualisierungen, STT -> STT Behaviourchange
- Brille + Controller bereitlegen
- Konnektivität prüfen (Controller, Lighthouses, Brille)
- TinyBaby anschalten und VR View drauf öffnen
- Teach-R starten
- Coach starten
- Evtl Frontal-lecture-room öffnen und Visualisations und LRS Fetcher reinziehen
- Snacks bereitlegen

Evaluationsdurchlauf

- Snacks und Wasser anbieten
- Ablauf erklären
  - Evaluation meiner Masterarbeit -> Voice in VR
  - Erst VR
  - Coach während selbst in VR
  - Fragebogen
- Teach-R Run starten
- Mit Brille und Controllern helfen
- Teach-R Steuerung erklären
- Audioaufnahme in Coach starten
- Aufgaben vorlesen

- Mit Brille und Controllern helfen
- An Desktop setzen und Aufgaben geben
- Teilnehmer*in zu Fragebogen wechseln
- Für Teilnahme danken und noch Snacks anbieten

Nachbereitung

- Brille desinfizieren
- Teach-R Run beenden
- Coach neu laden
- Fragebogen neu öffnen
- Snacks nachräumen
- Wasser auffüllen
- Falls letzter Teilnehmer für Tag:
    - Tiny Baby ausschalten
    - Desktop ausschalten
    - Brille aufhängen
    - Controller laden

Tasks

In Teach-R/ VR

1. Umschauen so lange Teilnehmer*in möchte
2. Öffne bitte das Learning Analytics Visualisierungsmenü
3. Aktiviere und teste bitte die Lautstärke-Kreis Visualisierung
4. Aktiviere und teste die Lautstärke-Nebel Visualisierung
5. Aktiviere den Teacher-Clone
6. Spiele bitte die Aufnahme über den Teacher-Clone ab
7. Pausiere bitte die Aufnahme über den Teacher-Clone
8. Teilnehmer*in so lange in Teach-R bleiben lassen wie gewollt (im Raum bleiben)

Im Coach/ 2D

1. Gehe bitte zu der Visualisierungen-Ansicht
2. Stoppe bitte die aktuelle Aufnahme, die in Teach-R läuft
3. Starte bitte die Aufnahme vom Teacher-Clone von neu
4. Wecke bitte verbal Sven auf (auf 2D verlegt)

5. Öffne bitte die Sprechgeschwindigkeitsvisualisierung
6. Öffne bitte jede Visualisierung bis auf die Sprechgeschwindigkeitsvisualisierung
7. So lange rumprobieren lassen wie gewollt

Eidesstattliche Versicherung

Recker, Jona	367356
Name, Vorname	Matrikelnummer (freiwillige Angabe)

Ich versichere hiermit an Eides Statt, dass ich die vorliegende ~~Arbeit/Bachelorarbeit/~~ Masterarbeit* mit dem Titel

Voice in Teach-R

ohne unzulässige fremde Hilfe (insbes. akademisches Ghostwriting) erbracht habe. Ich habe keine anderen als die angegebenen Quellen und Hilfsmittel benutzt. Für den Fall, dass die Arbeit zusätzlich auf einem Datenträger eingereicht wird, erkläre ich, dass die schriftliche und die elektronische Form vollständig übereinstimmen. Die Arbeit hat in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegen.

Aachen , March 27, 2024	
Ort, Datum	Unterschrift

*Nichtzutreffendes bitte streichen

Belehrung:

§ 156 StGB: Falsche Versicherung an Eides Statt
Wer vor einer zur Abnahme einer Versicherung an Eides Statt zuständigen Behörde eine solche Versicherung falsch abgibt oder unter Berufung auf eine solche Versicherung falsch aussagt, wird mit Freiheitsstrafe bis zu drei Jahren oder mit Geldstrafe bestraft.

§ 161 StGB: Fahrlässiger Falscheid; fahrlässige falsche Versicherung an Eides Statt
(1) Wenn eine der in den §§ 154 bis 156 bezeichneten Handlungen aus Fahrlässigkeit begangen worden ist, so tritt Freiheitsstrafe bis zu einem Jahr oder Geldstrafe ein.
(2) Straflosigkeit tritt ein, wenn der Täter die falsche Angabe rechtzeitig berichtigt. Die Vorschriften des § 158 Abs. 2 und 3 gelten entsprechend.

Die vorstehende Belehrung habe ich zur Kenntnis genommen:

Aachen , March 27, 2024	
Ort, Datum	Unterschrift