Technische Universität Berlin | Faculty IV | Electrical engineering and computer science |
Advanced Projects at the Quality and Usability Lab

Project Report

Project 7:
Subjective Listening Only Test for Speech Quality Database Creation

Aylin Gabrielle Arslan,
Ben Berger,
Beyda Rojin Boran,
Chandimal Galmangoda Guruge,
David Sebastian Grohs,
Jabora Speder

Berlin, 27. March 2024
Project Management: Wafaa Wardah

Contents

1. Intro

1.1. Background and Motivation

Due to the growing use of voice processing algorithms in multimedia and telecommunications applications, voice quality evaluation has become crucial in modern times [1]. The precise and reliable Evaluation of voice quality is crucial for the satisfaction of the end-users or customers of the voice processing systems that are used, like mobile phones, speech synthesis systems, etc. Voice quality is a common assessment object for telecommunications service providers in order to optimize the quality experienced by their customers. Speech quality can be assessed using subjective hearing tests or using objective quality measures [2]. Accordingly, it is necessary to design a reliable method for evaluating speech quality [1]. Creating a comprehensive speech quality database, labeling the data into subcategories and developing a survey platform for those taking the test, provide a good method for conducting these listening tests [3]. The goal is not only to collect data but also to develop an understanding of how people perceive and evaluate voice quality. This is key to developing more advanced voice quality assessment models that can ultimately be integrated into, for example, telecommunications services. By integrating such models, telecommunications companies could more accurately assess and improve the quality of the transmitted audio, which in turn leads to a better overall user experience.

1.2. Goals of the Project

The primary goal of this project is to create a comprehensive and meticulously labeled database of speech quality for training and refining machine learning models for speech processing. To facilitate the collection of this critical data, we have developed an essential tool - a web-based platform tailored for conducting subjective listening tests. This platform is critical for collecting a wide range of data under different speech degradation conditions, including background noise, signal compression, and distortion, to closely replicate the challenges of real-world telecommunications and multimedia contexts.

The range of speech quality in our study is intentionally broad, ranging from high-fidelity, clear samples to severely degraded samples. This breadth is key to capturing a full range of possible speech-quality scenarios. Once gathered, the data is subjected to stringent analysis and validation using advanced data science tools. To ensure the accuracy and reliability of the dataset, this step is critical. The analysis will include categorizing and labeling the data based on the type and degree of degradation, making it a structured and valuable resource for training algorithms.

2. Method und Development

2.1. Related work

In recent years, several research projects have appeared on the topic of subjective listening tests for speech and audio quality. Most tests and research on this subject are based on the International Telecommunications Union (ITU)-T Recommendation P.800, published in 1996, which defines methods and procedures for subjective assessment of transmission quality [4]. The most important assessment metrics are the Absolute Category Rating (ACR), Degradation Category Rating (DCR) and Comparison Category Rating (CCR). The ACR method is for obtaining the overall quality ratings directly from the participants without reference to a comparison. In the process, the listeners are asked to rate the overall quality of a particular stimulus, like speech or audio, on a categorical scale without any reference to another stimulus. The scale typically ranges from excellent to bad with a set of discrete categories for the listeners to choose from. The second method DCR is for obtaining an assessment of quality degradation by comparing the quality of a received signal with that of an original signal. The Listeners are presented with both a reference or an original signal and a degraded version of the sound. Next, they are asked to rate the perceived degradation in quality. Similar to ACR, DCR also uses a categorical scale for listeners to express the degree of degradation. The third method CCR is obtaining an assessment of quality by comparing the quality of two or more stimuli. In the testing phase, the participants are presented with two or more stimuli, like speech samples, and are asked to compare their quality. The listeners then assign categorical ratings based on their perceived differences in quality between the stimuli. Moreover, the P.800 is internationally recognized and widely used in the telecommunications industry. The recommendations provide flexible testing because the testing method and scoring scales can be chosen based on the specific experiment [4].

2.2. Our Development of the Web Platform

2.2.1. Selection of Tools

For the technical implementation of this project, we decided to use a web platform that can be run in the browser. The reason for this is that web-based software has many advantages compared to local desktop apps or mobile apps. One of the reasons is that multiple users can use the same version of the platform. Another reason is that users do not have to have installed the app beforehand, but can use it directly. In addition, web-based software can be used on various devices, such as laptops, cell phones or computers, but it also works on various operating systems or browsers [5]. Accordingly, a web-based platform is ideal for the in 1.2. mentioned goals for this project. In order to create this web software, we used Hypertext Markup Language, or HTML in short, which creates a basic framework for the website. Moreover Cascading Style Sheets or CSS in short is used to format the appearance of the website by specifying the layout, typography, colors and general design. The dynamic functions, such as form validation and page navigation, are programmed in JavaScript which is a programming language that is primarily used for client-side web development. The advantage of JavaScript in contrast to alternatives such as CoffeeScript or Dart is that it is widely supported by web browsers and has a large developer community. There are also numerous tools and frameworks specifically for JavaScript that make working with it easier

[6]. One of these tools we used is the open-source JavaScript runtime cross-platform named Node.js [7]. We also used the TheFragebogen framework for this project, which allows you to create a digital survey using pre-created functions and classes in JavaScript [8].

2.2.2. The used Data

One of the most important aspects of our study is to evaluate various audio files. For this purpose, we have 216 audio files of two female and two male speakers in German. What is later important to the participant is not the content of what was said, but rather how clear, distorted, loud and quiet the audio file is and whether any noise can be heard in the background. The exact degrading conditions can be found in the appendix under 7.2.

2.2.3. Technical Development

We used 214 of the 216 audio files to ensure an even distribution among five batches. Each batch consists of 42 audio files with one of the degrading conditions C02 to C63 and one audio file with excellent audio quality (condition C01), referred to as a gold question. These gold questions were implemented to test whether the participants answered the questions attentively. If, for example, a participant rates one of these files as very poor, it suggests that they may not be providing an honest evaluation but rather clicking on random answers. Test subjects were permitted to participate in the experiment up to two times, evaluating two batches per attempt. For that purpose, an ID system was developed in JavaScript in order to avoid participants rating the same batches twice. A total of 100 unique IDs, along with their respective batches and usage counts, were stored in a list. Each user is assigned one of these IDs before the experiment, which must be inputted into the software to initiate the test. If the same ID is entered again, it is recognized by the program and two different batches are randomly selected. Furthermore, the order of the audio files, except for the two gold questions, is randomized for each test. The individual questions were built using TheFragebogen [8] code blocks and afterward formatted with CSS. The final software, which integrates the ID system and the individual questions, is hosted on an HTML page. Upon completion of the test, a CSV file containing all the answers is automatically downloaded. The whole project can be found on GitHub (see appendix 7.1).

3. User-Testing

3.1. Preparation of tests

To carry out the experiments, we were provided with two rooms that inhibit sound. The total room volume was 30 DB, which is P.800 compliant [4].

The equipment we used included

- Windows laptops,
- USB audio interface/headfyer: Roland Quad-Capture 24bit 192khz
- Headphones: byerdynamic DT 797

To ensure that all participants are subject to the same test conditions, we tested the sound volume of our audio files in the respective rooms. We wanted to achieve a volume of around

73 DB, as this corresponds to the recommendations of P.863.2 [9]. For the sound test, the audio files were measured with the artificial head “HEAD acoustics II.3”, which essentially acts as a sound pressure meter. With the help of the measurement and analysis software “ACQUA” [10], the output volume for the experiment was derived. We concluded that a Windows volume of 90 and a setting of “2 p.m.” on the Roland Output Button produced an average volume of 72.9 DB.

3.2. Procedure of tests

The participant's task was to listen to short speech sequences and rate the various quality characteristics of the audio on a scale. The aim was to identify the audio characteristics and possible distortions and noise in the samples. The evaluation was subjective and anonymous. You can find the evaluation categories in the appendix under 7.4. Furthermore, the participating people each received an expense allowance of 30 euros for taking part in the study, except employees at the Technical University of Berlin (short: TUB). On average an experiment session lasted between 60 and 120 minutes, depending on how quickly the questions were answered. The prerequisite for participation was a very good knowledge and understanding of the German language, as the audio files and questions were in German, and no hearing impairment, as an excellent hearing ability is very important for an audio test. Also they shouldn’t work in the field of audio quality either or attend a hearing test up to six months before our test [4]. In order to find test subjects, we advertised our experiment on social media platforms such as Discord, Instagram, Whatsapp and Telegram. We also uploaded an advertisement on the TU’s Sona portal. The tests were carried out over a total of 12 days. The participants were allowed to take part up to twice, but they were given a different data set to evaluate for each attempt. Throughout the experiment, the participants first had to answer general questions about themselves followed by a tutorial to get familiar with the audio file evaluation process. Afterward, a total of 86 audio files had to be rated. Following the evaluation of 43 audio files, a short 10-minute break was provided to help ensure participants remained focused. In the final stages of the experiment, more questions regarding personal factors such as current mood or time pressure were presented to the participants (see appendix 7.3). We used these questions for the final evaluations in order to estimate if these factors might influence the ratings of participants. Furthermore, snacks and water were provided during the experiment to minimize the influence of hunger and thirst on the ratings. A minimum of 15 ratings per batch was needed.

4. Analysis

4.1. Procedure of the analysis

We saved the output from each participant as an .csv file which can be found in the GitHub. For the analysis, we wrote a code to analyze the average rating of each file, dimension and condition. Google Collaboratory based on Jupyter Notebook was used as the environment and the code was written in Python. The code can be found on GitHub called “QU_DataAnalysis.ipynb”. We did one merged excel-table called “overview merged.xls” just to have an overview, but for the data analysis we’ve used the .csv files. We created one database sorted by the files and their dimensions “SQL_Test_file.csv”, on sorted by the IDs

“SQL_Test_listening_test_ratings.csv, one for the conditions “SQL_Test_con.csv” and one only with the questionnaire answers “pre and postquestionnaire.xlsx”.

4.2. Results

Demographics

The total number of participants was 80. 24 of them participated twice. 28 of them were female, 47 were male, 3 were divers/intersexuals and 2 had no information. This can be explained by the fact that we mainly advertise on TUB Discord servers and Telegram groups and the number of women studying at the TUB is significantly lower compared to the number of men [11]. The average age was in all gender categories and in general 25/ 26, of which the youngest was 18 and the oldest 55 years old. The majority, 65 out of 80, were mother tongue speakers, the rest had a C1/C2 german level. Other mother tongues were Turkish, Hindi, Spanish, French, Bulgarian, English and one time Greek. There was no difference between mother-tongue speakers and participants who speak German as a foreign language.

Ratings

The best conditions were C23, C01, C04, C37, C59, C20, C42, C41 (from best to worst). From these conditions, the files with male speaker 2 and female speaker 1 were rated best, even if it wasn’t a huge difference. The worst conditions were C08, C58, C57, C34 and C45 (from worst to best). You can find the average ratings in the appendix 7.5. There is no clear correlation between the overall rating and the dimensions as you can see in the appendix. For example, we could not find one dimension which is always similar to the overall rating.

Questionnaire

Most people were used to telecommunication techniques and used them several hours a week. The romantic attraction scale showed a very differentiated attraction to gender. Also, we had mostly students as participants or people who finished their studies probably because of our advertisement in university groups. Most people agreed that electrical devices are helpful, but fewer people agreed that electrical devices are secure. Also 21 out of 80 left a comment, saying that their mood shifted during the test, mostly they were annoyed by loud noises.

5. Conclusion

5.1. Discussion and Conclusion

In summary, it can be said that a database was successfully created in accordance with the P.800 specifications. The evaluation of 80 participants made it possible to evaluate each condition several times with each male and female speaker. Both the technique and the procedure from the questionnaire to dimensions and advertisement could be adapted to international standards. The decision to build a website paid off, as there were hardly any technical problems. The data evaluation proved to be more difficult, as the high randomness of all evaluations made subsequent organization more difficult. Nevertheless, we were able to summarize all ratings and trace the demographic distribution. In retrospect, it must be criticized that we did not find a balanced age spectrum and also a gender imbalance in our data. Similarly, some files were rated more frequently than others, although all met the minimum rating.

5.2. Future Outlook and further Developments

This project laid the foundation for a database that can be used to implement machine learning techniques for speech quality assessment. The actual implementation of these algorithms falls beyond the project's scope. However, as our website and the questionnaire comply with international standards, it can be used as a prototype for German-language audio research. Also, the correlation between the questionnaires and the ratings can be further evaluated for improving these kinds of tests. For example, if there is a correlation between the rating of female or male speakers and the romantic attraction, it is important to consider this in future testing. In addition, we took the opportunity to implement some questions such as stress level, romantic affection and hunger in our questionnaire at the request of other projects at the TU Berlin. This anonymized data can provide clues for research into user engineering.

6. Bibliography

[1] Loizou, P.C. (2011). Speech Quality Assessment. In: Lin, W., Tao, D., Kacprzyk, J., Li, Z., Izquierdo, E., Wang, H. (eds) Multimedia Analysis, Processing and Communications. Studies in Computational Intelligence, vol 346. Springer, Berlin, Heidelberg

[2] Naderi, B., Möller, S., & Cutler, R. (2021, June). Speech quality assessment in crowdsourcing: Comparison category rating method. In 2021 13th International Conference on Quality of Multimedia Experience (QoMEX) (pp. 31-36). IEEE.

[3] Barry, D., Zhang, Q., Sun, P. W., & Hines, A. (2021). Go Listen: An End-to-End Online Listening Test Platform. Journal of Open Research Software, 9(1).

[4] Rec, ITU-T(1996). P. 800, Series P: Telephone transmission quality, Methods for subjective determination of transmission quality. International Telecommunication Union

[5] Contributor, T. (2023). web application (web app). Software Quality. https://www.techtarget.com/searchsoftwarequality/definition/Web-application-Web-app, last accessed 19.12.2023

[6] John, J. (2023). JavaScript vs Alternatives: Comprehensive Guide for Web Developers. Egochi. https://www.egochi.com/javascript-alternatives/, last accessed 19.12.2023

[7] Node.js. (2023). Node.js. https://nodejs.org/en, last accessed 19.12.2023

[8] Orefice R.H., Almeida L.A., Spur M. (2023). TheFragebogen. https://thefragebogen.de/, last accessed 19.12.2023

[9] ITU-T (2022). P. 863.2, Series P: Telephone transmission quality, telephone installations, local line networks. International Telecommunication Union

[10] HEAD acoustics (2024). ACQUA ist die leistungsfähige Mess- und Analyse-Software, um Sprachqualität und Audioqualität zu testen. https://www.head-acoustics.com/de/produkte/analyse-software/acqua, last accessed 12.03.2024

[11] Technische Universität Berlin (2024). Zahlen & Fakten. https://www.tu.berlin/ueber-die-tu-berlin/profil/tu-berlin-in-zahlen, last accessed 21.03.2024

7. Appendix

7.1. Github to full project

https://github.com/WafaaWardah/SLOTest_2023

7.2. Degraded conditions

C1	FB clean
C2	FB, stationary noise 12dB SNR
C3	FB, Noise 25dB (P50MNRU)
C4	FB Level -20dB
C5	FB 500-2500Hz
C6	FB 100-5000Hz , Level -10 dB
C7	FB, Time Clipping 2%
C8	FB, Time Clipping 20%
C9	NB G.729, Level -10dB
C10	NB G.711, Level +10dB
C11	NB AMR 12.2, street noise, 30 dB SNR
C12	NB AMR 12.2, car noise, 15 dB SNR
C13	NB G.711, babble noise, 35 dB SNR
C14	NB G.711, Time Clipping 5%
C19	WB G.722, Level -20dB
C20	WB G.722, Level -10dB
C21	WB G.722, Level +10dB
C22	WB AMR WB 23.85, car noise, 30 dB SNR
C23	WB AMR WB 12.65, street noise, 15 dB SNR
C24	WB AMR WB 23.85, babble noise, 25 dB SNR
C25	WB G.722, Time Clipping 5 %
C30	SWB EVS 13.2, Level -20dB

C31 SWB EVS 24.4, Level -20dB, street noise, 30 dB SNR

C32 SWB EVS 24.4, Level -10dB

C33 SWB EVS 24.4, Level -10dB, car noise, 30 dB SNR

C34 SWB EVS 13.2, street noise, 15 dB SNR, time clipping 2 %

C35 SWB EVS 24.4, babble noise, 25 dB SNR, time clipping 5 %

C36 SWB EVS 13.2, time clipping 10 %

C37 Reference condition

C38 Bandwidth limitation 20 Hz-8kHz

C39 Bandwidth limitation 50 Hz-7kHz

C40 Bandwidth limitation 120 Hz-8kHz

C41 Bandwidth limitation 20 Hz-12kHz

C42 Bandwidth limitation 20 Hz-14kHz

C43 Bandwidth limitation 50 Hz-3800Hz

C44 Bandwidth limitation 300 Hz-3400Hz

C45 Babble noise, SNR = 10 dB

C46 Babble noise, SNR = 20 dB

C47 Babble noise, SNR = 30 dB

C48 Car noise, SNR = 10 dB

C49 Car noise, SNR = 20 dB

C50 Car noise, SNR = 30 dB

C51 Street noise, SNR = 10 dB

C52 Street noise, SNR = 20 dB

C53 Street noise, SNR = 30 dB

C54 Signal clipping, 2 % loss rate

C55 Signal clipping, 5 % loss rate

C56 Signal clipping, 10 % loss rate

C57 Signal clipping, 15 % loss rate

C58 Signal clipping, 20 % loss rate

C59 Level enhancement + 10 dB

C60 Level enhancement + 20 dB

C61 Level attenuation - 10 dB

C62 Level attenuation - 20 dB

7.3. The used Questionnaires

We decided to do a post- and a pre-questionnaire, because some questions could have an influence on the rating. Therefore we asked questions which excluded the participant (like their work and their German skills) and the usage of audio technique (as these questions don't have an impact on the rating) in the pre-questionnaire. In the post-questionnaire we asked aspects which could impact the rating like education, romantic attraction, stress and hunger.

Pre-questionnaire

Bitte geben Sie hier Ihre PIN ein:	Please enter your PIN here:
Wie alt sind Sie?	How old are you?
Wie würden Sie Ihre Deutschkenntnisse einschätzen?	How would you rate your German language skills?
Wenn Deutsch nicht Ihre Muttersprache ist: Was ist Ihre Muttersprache?	If German is not your native language: What is your native language?
Wie viele Stunden pro Woche telefonieren Sie übers Festnetz?	How many hours per week do you use a landline?
Wie viele Stunden pro Woche telefonieren Sie mit dem Handy?	How many hours per week do you use your cell phone?
Wie viele Stunden pro Woche telefonieren über das Internet (z.B. Über Discord, Messenger-Apps wie Whatsapp)?	How many hours per week do you make calls via the Internet (e.g. via Discord, messenger apps such as Whatsapp)?
Wie viele Stunden pro Woche nutzen Sie Videoanrufe?	How many hours per week do you use video calls?
Wie sind Ihre bisherigen Erfahrungen mit den o.g. Telekommunikationsmittel?	What is your previous experience with the above-mentioned means of telecommunication?
Wie viele Stunden pro Woche nutzen Sie Social Media?	How many hours per week do you use social media?
Wie viele Stunden pro Woche nutzen Sie Online-Zeitungen?	How many hours per week do you use online newspapers?
Wie viele Stunden pro Woche nutzen Sie den Fernseher?	How many hours per week do you use the television?
	How many hours per week do you use

Wie viele Stunden pro Woche nutzen Sie das Radio? Haben Sie schon einmal an einem Hörtest/-Experiment teilgenommen? Wenn Ja : Wann? (Jahr) Haben Sie regelmäßig (z.B. über Ihre Arbeit) mit Audiotechnikverfahren/ -forschung zu tun? Wenn Nein: Hatten Sie früher einmal damit zu tun? Wenn Ja: Wann? (Jahr) Haben Sie Konzentrations- oder Hörschwächen? Wenn Ja: Benutzen Sie Hilfsmittel (z.B. Hörgerät, Medikamente)? Wenn Ja: Welche? (Wenn Sie spezielle Medikamente einnehmen ist es ausreichend mit 'Medikamente' zu antworten)	the radio? Have you ever taken part in a listening test/experiment? If yes : When? (year) Are you regularly involved with audio technology procedures/research (e.g. through your work)? If No: Have you ever been involved with it before? If yes: When? (year) Do you have any concentration or hearing difficulties? If yes: Do you use aids (e.g. hearing aid, medication)? If yes: Which ones? (If you take special medication, it is sufficient to answer 'medication')

Post-questionnaire

Wie identifizieren Sie sich? Fühlen Sie sich romantisch zu Frauen bzw. weiblichen/femininen Personen hingezogen? Fühlen Sie sich romantisch zu Männern bzw. männlichen/maskulinen Personen hingezogen? Was ist Ihr höchster erreichter Bildungsabschluss? Sind Sie hungrig? Vor wie vielen Stunden haben Sie zuletzt gegessen? Wie viele Stunden haben Sie letzte Nacht geschlafen? Hatten oder haben Sie heute Zeitdruck? Wie wichtig ist Ihnen Audio- bzw. Telefonqualität? Elektronische Geräte helfen, an Informationen zu gelangen. Elektronische Geräte ermöglichen einen hohen Lebensstandard. Elektronische Geräte erhöhen die Sicherheit. Elektronische Geräte machen unabhängig. Elektronische Geräte erleichtern mir den Alltag. Hat sich Ihre Stimmung während der	How do you identify yourself? Do you feel romantically attracted to women or female/feminine people? Do you feel romantically attracted to men or male/masculine people? What is your highest level of education? Are you hungry? How many hours ago did you last eat? How many hours did you sleep last night? Did you have or do you have time pressure today? How important is audio or telephone quality to you? Electronic devices help us to access information. Electronic devices enable a high standard of living. Electronic devices increase security. Electronic devices make you independent. Electronic devices make my everyday life easier. Did your mood change during the survey?

Befragung geändert? Anmerkung (z.B. an welcher Stelle oder aus welchem Grund?)	Comment (e.g. at what point or for what reason?)

7.4. Evaluation Categories

Overall Quality (Gesamtqualität)

- Describes how pleasant and appealing the sound was overall perceived. Consider factors such as clarity, precision and tonal richness. This should happen regardless of individual factors.

Timbre (Klangfarbe)

- Describes the characteristics or specific sound of a sample. Consider harmonic properties, pitch distribution and other characteristics. Is the sound rather warm, cool, lively or muted? A good value indicates that the audio files are authentic and unadulterated.

Discontinuity (Diskontinuität)

- Describes the extent to which interruptions, jumps or irregularities occur in given samples. Low discontinuity indicates smooth audio playback, while a high discontinuity value indicates annoying breaks or jumps in the sound progression.

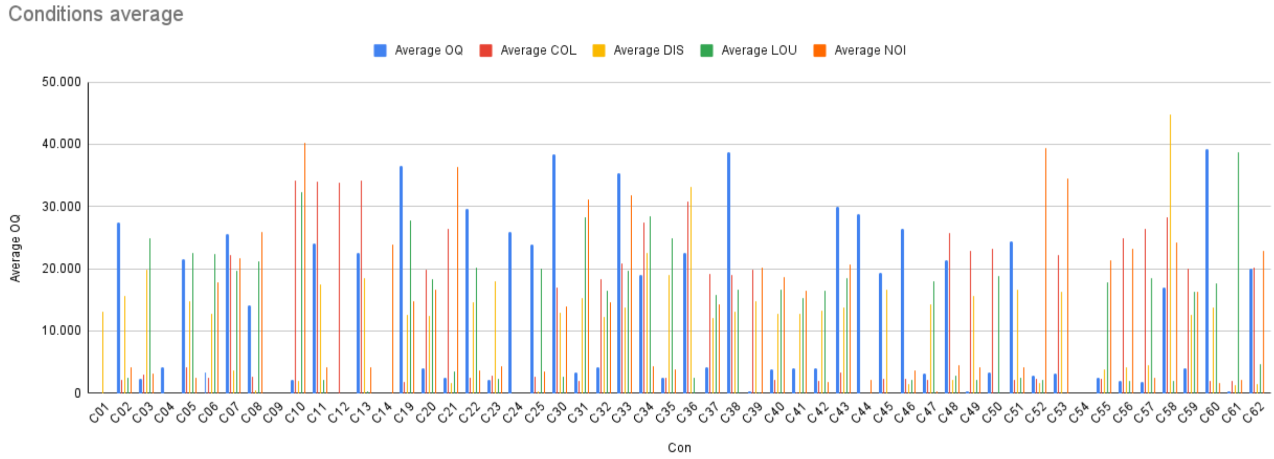
Loudness (Lautstärke)

- Describes the subjective perception of the strength or intensity of the sound in a sample. An appropriate volume enables clear audibility. If the volume is too low, details may be obscured . If the volume is too high, distortion may occur.

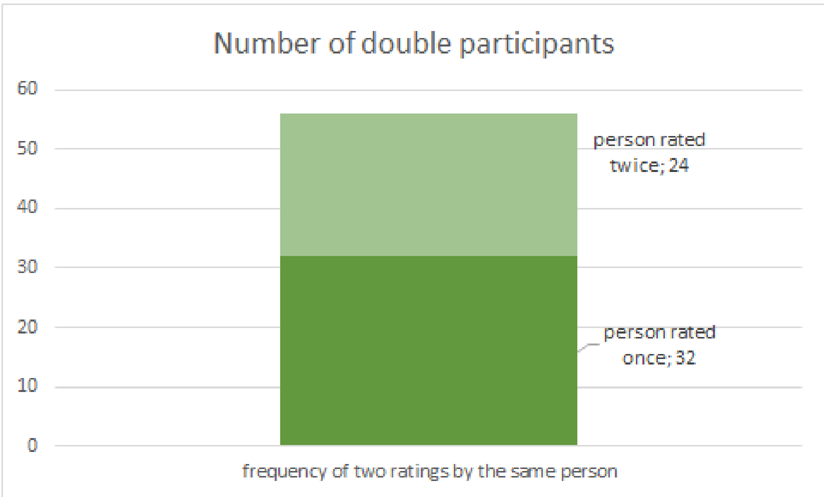
Noise (Rauschen)

- Describes the presence and intensity of unwanted background noise or interference in a sample. A low noise intensity is rated positively as it contributes to a clear and clean sound. In contrast, strong noise can impair the perception of details and disrupt the listening experience.

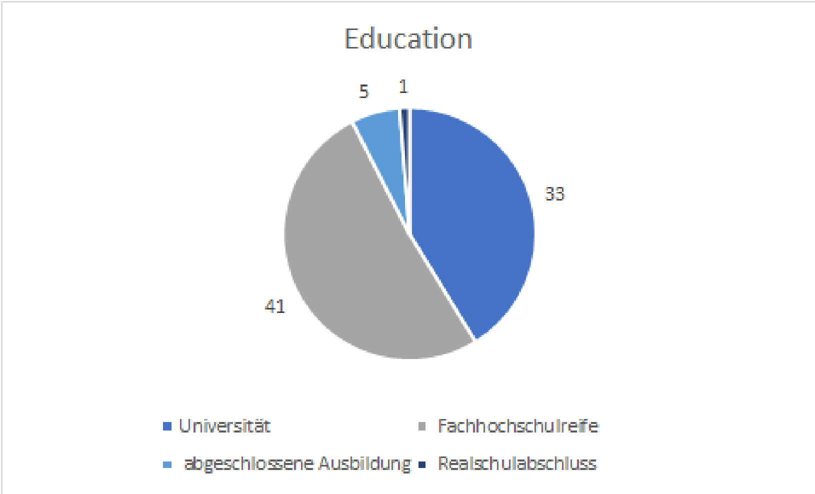
7.5. Data Analysis Results



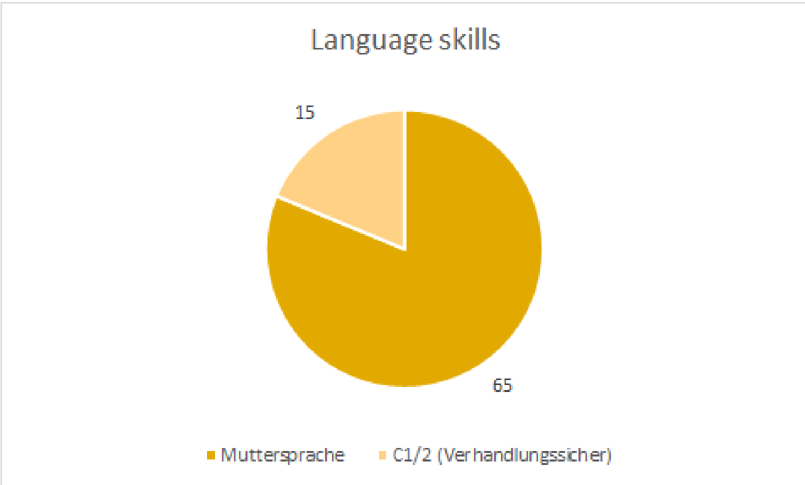
Conditions-Rating (Con) with the overall quality (Average OQ), the coloration rating (Average COL), the discontinuity rating (Average DIS), the loudness rating (Average LOU) and the noisiness rating (Average NOI).



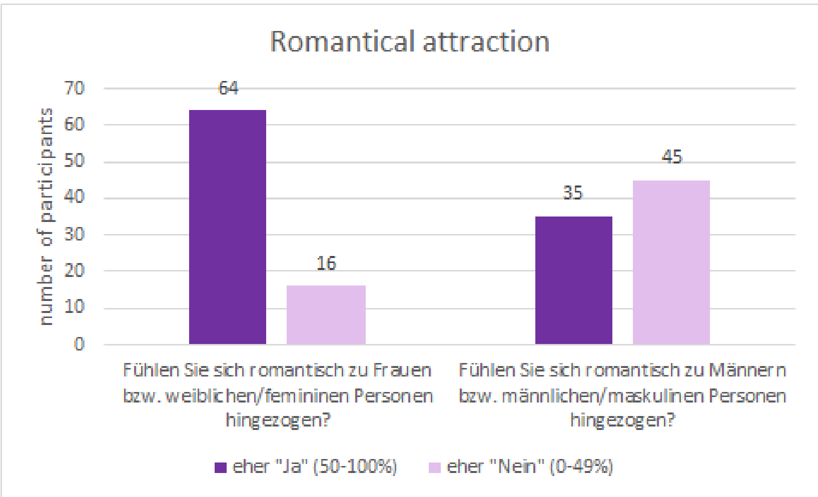
Number of participants, who participated twice out of all ratings



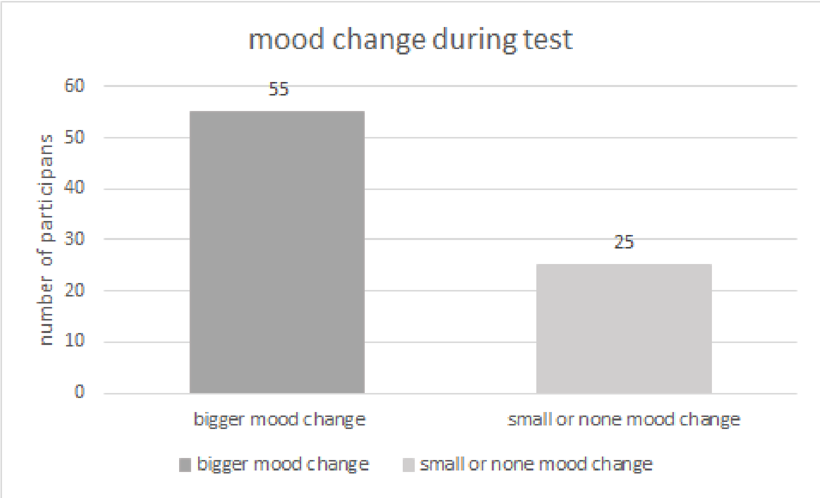
Education (upper left to right: University, A-Levels, Apprenticeship, Secondary school diploma)



Language skills (left to right: Mother tongue, C1/C2)



Romantical Attraction (left to female, right to men, purple: rather yes, pink: rather no)



Mood change during test