Analysis of Influencing Factors in Speech Quality Assessment using Crowdsourcing

vorgelegt von
Ing.
Rafael Zequeira Jiménez

an der Fakultät IV - Elektrotechnik und Informatik
der Technischen Universität Berlin
zur Erlangung des akademischen Grades

Doktor der Ingenieurwissenschaften
-Dr.-Ing.-

genehmigte Dissertation

Promotionsausschuss:
Vorsitzender: Prof. Dr. Axel Küpper
Gutachter: Prof. Dr.-Ing. Sebastian Möller
Gutachter: Prof. Dr. Oliver Hohlfeld
Gutachter: Prof. Peter Pocta
Tag der wissenschaftlichen Aussprache: 12. Juli 2021

Berlin 2022

Dedicated to my parents Mardely and Braulio, my sister Raisa, and my grandparents Aurora, Narciso, and Guillermina.

My parents always supported and encouraged me to be a good student. Without their guidance and unconditional love, I wouldn't be here today.

My brilliant sister accompanied me throughout most of my academic life. She helped me improve myself every day.

To my beautiful grandmother Aurora which I love very much. Thanks for teaching me the multiplication tables; that was the beginning of all this all.

.

.

.

.

.

.

Dedicado a mis padres Mardely y Braulio, a mi hermana Raisa y a mis abuelos Aurora, Narciso y Guillermina.

Mis padres siempre me apoyaron y animaron para que fuese un buen estudiante. Sin su guía y su amor incondicional, no estaría aquí hoy.

Mi inteligente hermana me compañó durante la mayor parte de mi vida académica. Ella me ayudó a mejorar cada día.

A mi linda abuela Aurora, a la que quiero mucho. Gracias por enseñarme las tablas de multiplicar; ese fue el comienzo de todo esto.

Abstract

Crowdsourcing has emerged as a competitive mechanism to conduct user studies on the Internet. Users in crowdsourcing perform small tasks remotely from their computer or mobile device in exchange for monetary compensation. Nowadays, multiple crowdsourcing platforms offer a fast, low cost and scalable approach to collect human input for data acquisition and annotations. However, the question remains whether the collected ratings in an online platform are still valid and reliable. And if such ratings are comparable to those gathered in a constrained laboratory environment. There is a lack of control to supervise the participant and often not enough information about their playback system and background environment. Therefore, different quality control mechanisms have been proposed to ensure reliable results and monitor these factors to the extent possible [1, 2, 3].

The quality of the transmitted speech signal is essential for telecommunication network providers. It is an important indicator used to evaluate their systems, services, and to counterbalance potential issues. Traditionally, subjective speech quality studies are conducted under controlled laboratory conditions with professional audio equipment. This way, good control over the experimental setup can be accomplished, but with some disadvantages: conducting laboratory-based studies is expensive, time-consuming, and the number of participants is often relatively low. Consequently, the experiment outcomes might not be representative of a broad population.

In contrast, crowdsourcing represents an excellent opportunity to move such listening tests to the Internet and target a much wider and diverse pool of potential users at a fraction of the cost and time. Nevertheless, the implementation of existing subjective testing methodologies into an Internet-based environment is not straightforward. Multiple challenges arise that need to be addressed to gather valid and reliable results.

This dissertation evaluates the impact of relevant factors affecting the results of speech quality assessment studies carried out in crowdsourcing. These factors relate to the test structure, the effect of environmental background noise, and the influence of language differences. To the best of the author's knowledge, these influencing factors have not yet been addressed.

The results indicate that it is better to offer test tasks with a number of speech stimuli between 10 and 20 to encourage listener participation while reducing study response times. Additionally, the outcomes suggest that the threshold level of environmental background noise for collecting reliable speech quality scores in crowdsourcing is between 43dB(A) and 50dB(A). Also, listeners were more tolerant of the TV-Show noise compared to the street traffic noise when executing the listening test. Furthermore, the feasibility of using web-audio recordings for environmental noise classification is determined. A Multi-layer Perceptron Classifier with an *adam* solver achieved an accuracy of 0.69 in noise classification. In contrast, a deep model based on a *"Long Short-Term Memory"* architecture accomplished an RMSE of 4.58 on average (scale of 30.6dBA to 81.3dBA) on the test set for noise level estimation.

Finally, an experiment was performed to determine if it is possible to gather reliable speech quality ratings for German stimuli with native English and Spanish speakers in a crowdsourcing environment. The Person correlation to the laboratory results was strong and significant, and the RMSE low despite the listeners' mother tongue. However, a bias was seen in the quality scores collected from the English and Spanish crowd-workers, which was then corrected with a first-order mapping.

Zusammenfassung

Crowdsourcing hat sich als wettbewerbsfähiger Mechanismus zur Durchführung von Nutzerstudien im Internet herauskristallisiert. Diese Benutzer führen kleine Aufgaben aus der Ferne von ihrem Computer oder Mobilgerät aus und erhalten dafür eine finanzielle Entschädigung. Heutzutage bieten mehrere Crowdsourcing-Plattformen einen schnellen, kostengünstigen und skalierbaren Ansatz, um menschliche Eingaben für die Datenerfassung und Annotationen zu sammeln. Es bleibt jedoch die Frage, ob die gesammelten Bewertungen in einer Online-Plattform noch gültig und zuverlässig sind, und ob solche Bewertungen mit denen vergleichbar sind, die in einer Laborumgebung gesammelt wurden. Es fehlt die Kontrolle, um den Teilnehmer zu überwachen, und oft gibt es nicht genügend Informationen über das Wiedergabesystem und die Hintergrundumgebung. Daher wurden verschiedene Qualitätskontrollmechanismen vorgeschlagen, um zuverlässige Ergebnisse zu gewährleisten und diese Faktoren so weit wie möglich zu überwachen [1, 2, 3].

Die Qualität des übertragenen Sprachsignals ist für Anbieter von Telekommunikationsnetzen essentiell. Sie ist ein wichtiger Indikator, um ihre Systeme und Dienste zu bewerten und um möglichen Problemen entgegenzuwirken. Traditionell werden Studien zur subjektiven Sprachqualität unter kontrollierten Laborbedingungen mit professionellem Audio-Equipment durchgeführt. Auf diese Weise kann eine gute Kontrolle über den Versuchsaufbau erreicht werden, allerdings mit einigen Nachteilen: Es ist teuer, zeitaufwendig und die Anzahl der Teilnehmer ist oft relativ gering. Folglich sind die Ergebnisse des Experiments möglicherweise nicht repräsentativ für eine breite Population.

Im Gegensatz dazu stellt Crowdsourcing eine hervorragende Möglichkeit dar, solche Hörtests ins Internet zu verlagern und einen viel größeren und vielfältigeren Pool von potenziellen Nutzern zu einem Bruchteil der Kosten und des Zeitaufwands anzusprechen. Dennoch ist die Implementierung bestehender subjektiver Testmethoden in eine internetbasierte Umgebung nicht einfach. Es ergeben sich mehrere Herausforderungen, die angegangen werden müssen, um valide und zuverlässige Ergebnisse zu erhalten.

Diese Dissertation evaluiert den Einfluss relevanter Faktoren, die die Ergebnisse von Studien zur Bewertung der Sprachqualität, die im Crowdsourcing durchgeführt werden, beeinflussen. Diese Faktoren beziehen sich auf die Teststruktur, den Einfluss von Umgebungsgeräuschen und den Einfluss von Sprachunterschieden. Nach bestem Wissen des Autors sind diese Einflussfaktoren bisher noch nicht behandelt worden.

Die Ergebnisse deuten darauf hin, dass es besser ist, Testaufgaben mit einer Anzahl von Sprachstimuli zwischen 10 und 20 anzubieten, um die Hörerbeteiligung zu fördern und gleichzeitig die Reaktionszeiten der Studie zu reduzieren. Darüber hinaus deuten die Ergebnisse darauf hin, dass der Schwellenwert des Umgebungsgeräusches für die Erfassung zuverlässiger

Sprachqualitätswerte beim Crowdsourcing zwischen 43dB(A) und 50dB(A) liegt. Außerdem waren die Hörer bei der Durchführung des Hörtests toleranter gegenüber dem Lärm der TV-Show als gegenüber dem Straßenverkehrslärm. Darüber hinaus wird die Machbarkeit der Verwendung von Web-Audio-Aufnahmen für die Klassifizierung von Umgebungsgeräuschen ermittelt. Ein Multi-Layer-Perceptron-Klassifikator mit einem “Adam”-Solver erreichte bei der Geräuschklassifikation eine Genauigkeit von 0,69. Im Gegensatz dazu erreichte ein tiefes Modell, das auf einer Long Short-Term MemoryArchitektur basiert, einen RMSE von durchschnittlich 4,58 (Skala von 30,6dBA bis 81,3dBA) auf dem Testset zur Geräuschpegelschätzung.

Schließlich wurde ein Experiment durchgeführt, um festzustellen, ob es möglich ist, zuverlässige Sprachqualitätsbewertungen für deutsche Stimuli mit englischen und spanischen Muttersprachlern in einer Crowdsourcing-Umgebung zu sammeln. Die Personenkorrelation zu den Laborergebnissen war stark und signifikant, und der RMSE trotz der Muttersprache der Hörer niedrig. Allerdings wurde eine Verzerrung in den von den englischen und spanischen Crowd-Workern gesammelten Qualitätsbewertungen festgestellt, die dann mit einem Mapping erster Ordnung korrigiert wurde.

Acknowledgements

During my time working on this dissertation at the Quality and Usability Lab, I had the pleasure to meet, work, and get to know a large number of awesome people.

I would like to thank my supervisor Prof. Dr-Ing. Sebastian Möller for his support, advice, scientific assistance, and for providing me the opportunity to pursue my doctoral degree. I also would like to thank Prof. Dr. Oliver Hohlfeld and Prof. Peter Pocta for co-examining this dissertation and serving on my doctoral committee.

Many thanks to Irene Hube-Achter, Yasmin Hillebrenner, and Tobias Jettkowski for their administrative and technical support during these years at the QU Lab.

Thanks to all former and current colleagues at the Quality and Usability Lab, including Dr.-Ing. Benjamin Bähr, Dr. Benjamin Weiss, Dr. Falk Schiffner, Dr.-Ing. Tilo Westermann, Dr. Patrick Ehrenbrink, Dr. Babak Naderi, Dr. Jan-Niklas Voigt-Antons, Dr. Dennis Guse, Dr. Stefan Josef Uhrig, Gabriel Mittag, Steven Schmidt, Saman Zadtootaghaj. Thank you all for the numerous talks, discussions, and collaborations.

Special thanks to my former colleague and friend Dr. Laura Fernández Gallardo, for her vital support at the beginning of my scientific career. Thank you for helping me write my first article.

Thanks to all my friends who supported me in all sorts of ways.

A special thanks to my partner Rafael for his patience and for cheering me up during many difficult and stressful times. Without your support and love, I probably would not have made it.

It was a pleasure meeting you all and becoming part of my life. All others who are not mentioned be aware of my appreciation.

Table of Contents

Acronyms

2G 2nd Generation 38

3G 3rd Generation 38

AAC Advanced Audio Coding 100

ACR Absolute Category Rating 3, 19

AMR Adaptive Multirate Codec 100

AMR-NB Adaptive Multirate Codec – Narrowband 38

AMR-WB Adaptive Multirate Codec – Wideband 38, 102

AMT Amazon Mechanical Turk 12, 72

ANOVA Analysis of Variance 42

API Application Programming Interface 69

CCR Comparison Category Rating 21

CS Crowdsourcing 48

DCR Degradation Category Rating 21

EVRC Enhanced Variable Rate Codec 99

EVS Enhanced Voice Services 38, 101

FB fullband 38, 101

GSM Global System for Mobile Communications 39

GUI Graphical User Interface 23

HIT Human Intelligence Task 72

ICC Intraclass Correlation Coefficient 43

IQR Interquartile Range 50

ITU International Telecommunication Union 20

ITU-T International Telecommunication Union Telecommunication Standardization Sector 19

LOT Listening-Only Tests 3, 19, 20

LSTM Long short-term memory 66

LTE Long Term Evolution 38

M2M Machine to Machine 100

Mdn Median 35, 37

MNRU Modulated Noise Reference Unit 99

MOS Mean Opinion Score 3, 12, 13, 19

MW microWorkers 2, 81

NB narrowband 36

NLP Natural Language Processing 27

QoE Quality of Experience 2

RMSD Root Mean Square Deviation 45

RMSE Root Mean Square Error 33

RNN Recurrent Neural Network 66

SNR Signal to Noise Ratio 100

SOS Standard Deviation of Opinion Scores 13

SPL Sound Pressure Level 49

SWB super-wideband 36, 99

UMTS Universal Mobile Telecommunications System 39, 102

VoIP Voice over Internet Protocol 39, 100

VoLTE Voice over LTE 38

WB wideband 36

1 Introduction

1.1 Speech Quality

The primary purpose of using a speech telephony service is for vocal human-to-human communication. The technological developments within traditional and modern packet-based (Voice-over-IP) telephony networks can disturb and even impair the transmitted voice signal. The components responsible for these communication impairments are the codecs, delay, bandwidth limitations, packet-loss, linear and non-linear filters, echo, noise, and others [4].

Therefore, it is important for telecommunication network providers to understand how end-users perceive and experience degradations. Estimating the quality of transmitted speech over telecommunications systems enables them to improve their services and counteract possible issues. In this context, the quality of transmitted speech is also referred to as the so-called *Quality of Experience* (QoE).

The term *Quality of Experience* was first introduced in [5]. Then, based on [6], the QoE definition was extended to:

Quality of Experience *"is the degree of delight or annoyance of a person whose experiencing involves an application, service, or system. It results from the persons evaluation of the fulfillment of his or her expectations and needs with respect to the utility and/or enjoyment in the light of the persons context, personality and current state."*

1.1.1 Speech Quality Assessment

A common means to study and understand the QoE of telephony services is by conducting passive subjective experiments with human participants. Such experiments are commonly carried out in constrained laboratory environments with controlled conditions regarding room acoustics and background noise. This way, the impact of confounding factors can be limited while ensuring valid and reliable study outcomes.

This test procedure is the so-called *Listening-Only Test* (LOT), which permits to gather overall quality ratings on a five-point *Absolute Category Rating* (ACR) scale. The collected scores are then averaged by listener, condition, or file to yield a *Mean Opinion Score*

(MOS). However, these subjective laboratory studies are expensive and time-consuming. Quality evaluations come at a relatively high expense, including the cost to build a controlled environment, maintenance, fees for administering test participants, and participants remunerations.

As a result, controlled experiments are usually only carried out by a limited number of institutions, e.g., telecommunication providers, research institutes, or universities that can afford such expenses. Additionally, the controlled laboratory environment leads to artificial test situations in some cases. It was exposed in [7] that users rate services differently in their life context than in a laboratory testing environment. Thus, the validity of the test suffers from the likely higher reliability created by the controlled environment. Consequently, the demand for instrumental models to predict the overall quality of transmitted speech and alternative test methods has increased in recent years [8, 9].

1.1.2 Crowdsourcing

Crowdsourcing represents a valid alternative to counteract some of the limitations of laboratory studies. Researchers are increasingly using it to collect data in more realistic test environments.

The term *"crowdsourcing"* was coined in 2005 by journalists Jeff Howe and Mark Robinson. They later materialized this idea in 2006 with an article explaining how businesses were using the Internet to *"outsource"* to the *"crowd"* work that was once performed by a designated agent, e.g., an employee, freelancer, or a separate company [10]. This article quickly led to the *portmanteau "crowdsourcing"*.

"Crowdsourcing" is a model in which organizations or individuals acquire services and goods that can vary from idea generation, tasks, votes, finance, and medicine. Such goods or services usually are collected from a large, relatively open, and rapidly evolving group of participants commonly referred to as "workers" or "crowd-workers". This model involves using the Internet through crowdsourcing platforms to attract and divide the work between participants to achieve a cumulative result.

Figure 1.1 depicts a simplified crowdsourcing workflow process excluding the payment steps. First, the *"work provider"* creates a task on the platform. In some cases, these tasks may include external content that could be stored in the crowdsourcing platform or external storage service. Frequently, the platforms provide multiple task templates (e.g., survey, image tagging, sentiment analysis, voting) and allow importing HTML and Javascript code for greater task customization and flexibility. At this point, the *"work provider"* can specify different tasks' properties, such as the reward, the target audience, number of responses, and others. Afterward, the tasks are accumulated in a pool that is available to the workers for execution. The validation of the submitted answers is mostly carried out by the *"work provider"*. However, some crowdsourcing platforms implement certain automatic validation mechanisms. Finally, crowd-workers' responses are stored and merged in the platform repository, from where the *"work provider"* can download the results.

Crowdsourcing facilitates the materialization of opinions and evaluations through a multitude of evaluators and contributors working in an open and participatory manner. Over the years, multiple systems based on different collaborative methods have been created to address a wide range of tasks [11]. All of these systems fall under the scope of "crowdsourcing".

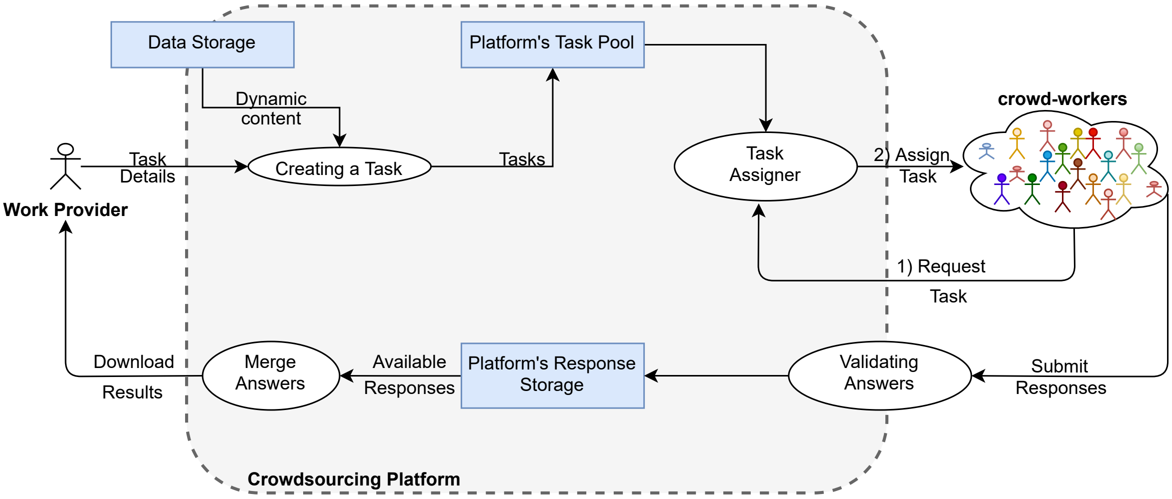


Figure 1.1: This figure represents a crowdsourcing workflow involving the main steps from task creations to results gathering.

Some popular examples are: “Stack Overflow”[1], “Linux”, “Wikipedia”[2], “Amazon Mechanical Turk”[3], and “microWorkers”[4](MW).

The advantages of using crowdsourcing include cost improvements, speed, flexibility, scalability, and diversity. These benefits have led to the wide adoption of crowdsourcing as a powerful instrument for carrying subjective user-centered experiments. Consequently, researchers have been using crowdsourcing to investigate different aspects of images [12, 13], videos [14, 15, 16], mobile gaming [17], speech [18] and audio [19] multimedia applications. The authors of [3] provide a good summary of the use of crowdsourcing for multimedia assessments.

1.1.3 Speech Quality Assessment in Crowdsourcing

The wide availability of the Internet has led to the creation of multiple crowdsourcing services and platforms. These platforms offer low cognitive tasks or jobs to a demographically diverse pool of workers. And the workers execute such tasks with their personal computer from the comfort of their home in exchange for monetary compensation.

The execution of speech quality evaluations in crowdsourcing benefits from reduced experimental turn-around times at a lower cost. Additionally, crowdsourcing permits to reach a broad and diverse audience for collecting quality ratings in a more ecologically valid context than traditional practices of in-Lab annotations. However, conceptual and technical challenges arise due to the remote test settings, and multiple mechanisms have been proposed to ensure valid results [20, 21].

Crowdsourcing users work without supervision, and they may not follow the instructions given. Consequently, they could end up performing a listening test in a noisy environment or with inappropriate equipment. For instance, researchers in [22] found that workers exhibit a low discrimination capacity and don’t perceive certain speech characteristics when performing the listening test with loudspeakers.

[1] https://stackoverflow.com
[2] https://www.wikipedia.org
[3] https://www.mturk.com
[4] https://www.microworkers.com/

To contrast some of the challenges of carrying speech quality assessments in crowdsourcing, researchers in [23] proposed the use of temporal expiring training certificates as a qualification requirement. With this method, the authors achieved an improvement in terms of correlation to the laboratory test results. Moreover, authors in [24] suggested using “trapping questions” as a mechanism to detect inattentive listeners and to discard unreliable ratings.

1.1.4 Differences Between Laboratory-based and Crowdsourcing-based Speech Quality Assessments

As mentioned above, conducting speech quality assessment studies in the laboratory is costly and time-consuming, requiring the availability of testing facilities and human participants. Thanks to the artificial setup of the laboratory tests, it is possible to quantify small differences between speech stimuli that would otherwise be imperceptible under normal usage conditions of a given speech service. The need for participants to access test facilities limits the demographic characteristics of users that can be covered in a single test. Therefore, despite showing high sensitivity and reliability of the results, the laboratory tests could show relatively low ecological validity, in the sense that their results are not representative of the daily use of the service.

The use of the crowdsourcing paradigm could help overcome some of these limitations. A broader, demographically balanced group of users can be reached at a lower cost. However, it is limited to connected and Internet-savvy users. Quality evaluations are usually performed under normal conditions of service usage, with the user’s standard equipment. In this way, ecological validity can be greatly increased, albeit at the cost of rather little control over the test setup, procedure, participants, and environment. The new ITU-T Recommendation P.808 has recently been established to limit the impact of poor experimental control of crowdsourced speech quality assessment studies. More details of Recommendation P.808 can be found in Subsection 3.3.1 and in [25].

1.2 Influencing Factors in Speech Quality Assessment using Crowdsourcing

As previously introduced, the execution of speech quality assessment studies in crowdsourcing is subject to several influencing factors. According to [5], an *influencing factor* is defined as follows:

Influencing Factor: *“Any characteristic of a user, system, service, application, or context whose actual state or setting may have influence on the Quality of Experience for the user.”*

The ITU-T offers a technical report about the subjective evaluation of multimedia using crowdsourcing [26]. This report focuses on the general aspects of crowdsourcing, and it also provides an overview of the influencing factors, which can be grouped into four categories, i.e., user, task, platform, and context related. Figure 1.2 presents an overview of the identified influencing factors of performing speech quality assessment experiments in crowdsourcing.

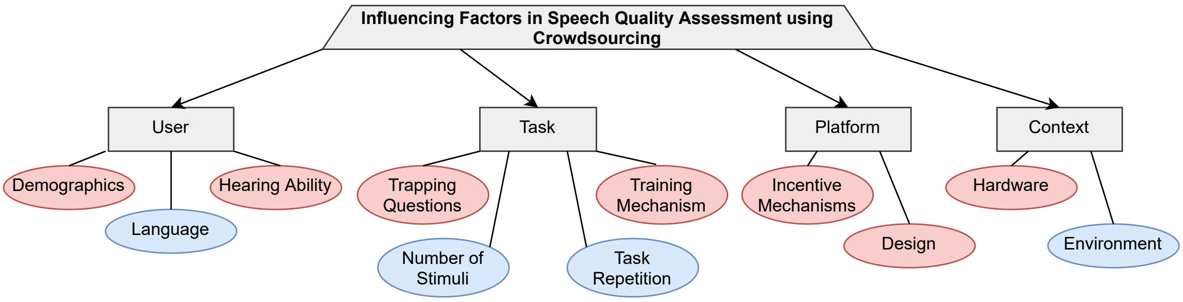


Figure 1.2: The figure presents the main influencing factors of conducting speech quality assessment studies in crowdsourcing. These factors are categorized under four headings, i.e., user-, task-, platform-, and context-related factors. Depicted in blue are the influencing factors addressed in this dissertation. The remaining factors were already addressed in the literature or should be investigated in future work.

The user influencing factors are related to the crowd-worker's characteristics and their ability to execute the speech quality assessment properly. Included in this category are the demographics factors (e.g., age, gender) [27], listener hearing impairments, and language proficiency. For instance, it is important to estimate the hearing abilities of participants, as crowd-workers may not realize that they may have mild or moderate hearing loss [27]. Researchers in [28, 29] proposed an adapted version of a digit-triplet test to detect hearing-impaired listeners. However, to the author's knowledge, the influence of linguistic differences has not yet been investigated. Therefore, it is addressed in this dissertation.

The task influencing factors relates to the speech quality assessment task's properties. Researchers at [24] proposed to include trapping questions within the listening test to promote motivation and detect inattentive workers. Furthermore, [23] encourages using a temporal expiring training certificate as a prerequisite to perform the speech quality assessment test. Moreover, the test duration in crowdsourcing is usually short, i.e., between 5 and 15 minutes [30], which leads to a mixed within-between-participants study design, as each task would contain only a subset of the entire dataset. This dissertation addressed the influence of task-related factors that, to the authors' knowledge, has not been investigated so far, i.e., the optimal number of speech stimuli included in a task and the influence of task repetition.

The platform influencing factors are the design and incentive mechanisms [31]. These factors influence workers' decision to participate in a speech quality assessment task versus other tasks that may be available on the crowdsourcing platform [31]. Consequently, these factors indirectly affect the workers' performance. However, the influence of these platform-related factors are not investigated in this dissertation and should be addressed in future research.

The context influencing factors are related to the hardware employed by the workers to conduct the listening test and the environment in which they ultimately perform the speech quality evaluation. Crowd-workers take part in user studies employing their equipment. This equipment can vary widely, may not be calibrated, and could be of poor quality compared to the hardware provided in laboratory-based studies. In a listening experiment, it is essential to transmit the stimuli without distortions other than those to be judged. Additionally, it is equally important that poor-quality sound cards or headphones do not compromise the stimuli reproduction. Researchers in [24, 23] reported that narrow-band (NB) speech files were rated lower in quality compared to the laboratory in a crowdsourcing speech quality assessment test. Authors in [32] analyzed the influence of employed headphones when performing the speech listening test, i.e., users' regular headphones vs. professional ones. Furthermore, the acoustic environmental characteristics, i.e., noise, reverberation, and the context (i.e., distractive), can directly influence the quality of the responses by masking the underlying test condition or indirectly by affecting the attention of crowd-workers [32, 33, 34]. Hence, this dissertation investigates the influence of environmental background noise on speech quality evaluations performed in web-crowdsourcing.

1.3 Research Questions and Thesis Outline

This dissertation addresses relevant questions related to the influencing factors introduced in the previous section that has not yet been addressed. These questions are related to the test structure, the impact of environmental background noise, and the influence of language differences in speech quality assessment studies executed in crowdsourcing. Specifically, the following research questions are answered:

- What is the optimal number of speech stimuli to include in a speech quality assessment task in crowdsourcing to achieve valid and reliable results?
- What is the influence of conducting a speech quality assessment task multiple times?
- Which environmental noises and distractions are workers exposed to when performing crowd-work?
- What is the impact of the environmental background noise on the speech quality ratings collected in crowdsourcing?
- Which level of environmental background noise is acceptable to collect reliable speech quality scores in crowdsourcing?
- Can environmental background noise recordings collected through the audio web-API be used to identify the type of noise?
- Can non-native German listeners provide reliable speech quality scores to a German speech dataset?

This thesis is structured as follows: In this chapter, the crowdsourcing paradigm is introduced, and also a brief presentation of human speech quality perception. Additionally, I expose some of the main challenges of conducting speech quality studies in crowdsourcing, which set the ground and motivates the investigation carried out in this dissertation.

Chapter 2 reviews relevant work related to the research questions addressed in this thesis. Chapter 3 describes concepts and methods for the practical assessment of speech quality in crowdsourcing. Furthermore, the speech material employed in the experiments executed in this thesis is presented.

Chapter 4 investigates the number of stimuli to include in a single task and analyzes the influence of performing the speech quality evaluations multiple times. Chapter 5 addresses the research questions regarding the influence of environmental background noise on the speech quality scores and evaluates different models to classify background noise from web audio recordings.

Chapter 6 studies whether it is possible to employ non-German listeners to gather reliable speech quality scores from a German dataset. Finally, Chapter 7 concludes and discusses the main findings of this dissertation and presents directions for future work.

Most of the scientific contributions presented in this thesis have been published in the form of conference articles. The author of this dissertation has been the first author and main contributor of all of them:

- [35] Rafael Zequeira Jiménez, Laura Fernández Gallardo, and Sebastian Möller. "Outliers Detection vs. Control Questions to Ensure Reliable Results in Crowdsourcing. A Speech Quality Assessment Case Study". In: *Companion Proceedings of the The Web Conference 2018*. WWW '18. Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee, 2018, pp. 1127–1130. ISBN: 978-1-4503-5640-4. DOI: 10.1145/3184558.3191545. URL: https://doi.org/10.1145/3184558.3191545

- [36] Rafael Zequeira Jiménez, Laura Fernández Gallardo, and Sebastian Möller. "Influence of Number of Stimuli for Subjective Speech Quality Assessment in Crowdsourcing". In: *2018 Tenth International Conference on Quality of Multimedia Experience (QoMEX)*. May 2018, pp. 1–6. DOI: 10.1109/QoMEX.2018.8463298

- [37] Rafael Zequeira Jiménez, Laura Fernández Gallardo, and Sebastian Möller. "Environmental Noise Recording as a Quality Control for Crowdsourcing Speech Quality Assessments". In: *44. Deutsche Jahrestagung für Akustik (DAGA)*. Deutsche Gesellschaft für Akustik DEGA e.V., Mar. 2018, pp. 303–306. ISBN: 978-3-939296-13-3

- [38] Rafael Zequeira Jiménez, Gabriel Mittag, and Sebastian Möller. "Effect of Number of Stimuli on Users Perception of Different Speech Degradations. A Crowdsourcing Case Study". In: *2018 IEEE International Symposium on Multimedia (ISM)*. 2018, pp. 175–179. DOI: 10.1109/ISM.2018.00-16

- [39] Rafael Zequeira Jiménez, Babak Naderi, and Sebastian Möller. "Evaluating Acoustic Features from Environmental Audio Recordings via Web. A Crowdsourcing Survey on Background Noise Characteristics". In: *45. Deutsche Jahrestagung für Akustik (DAGA 2019)*. Deutsche Gesellschaft für Akustik DEGA e.V., Mar. 2019, pp. 1190–1193. ISBN: 978-3-939296-14-0

- [40] Rafael Zequeira Jiménez, Babak Naderi, and Sebastian Möller. "Background Environment Characteristics of Crowd-Workers from German Speaking Countries Experimental Survey on User Environment Characteristics". In: *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*. 2019, pp. 1–3. DOI: 10.1109/QoMEX.2019.8743208

- [41] Rafael Zequeira Jiménez, Babak Naderi, and Sebastian Möller. "Effect of Environmental Noise in Speech Quality Assessment Studies using Crowdsourcing". In:

2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX). May 2020, pp. 1–6. DOI: `10.1109/QoMEX48832.2020.9123144`

- [42] Rafael Zequeira Jiménez, Babak Naderi, and Sebastian Möller. "Effect of Environment in Speech Quality Assessment in Crowdsourcing". In: *Proceedings of Forum Acusticum.* European Acoustics Association. 2020
- [43] Rafael Zequeira Jiménez, Sebastian Möller, and Gabriel Mittag. "Removing the Bias in Speech Quality Scores Collected in Noisy Crowdsourcing Environments". In: *submitted to: 13th International Conference on Quality of Multimedia Experience (QoMEX).* 2021
- [44] Rafael Zequeira Jiménez, Babak Naderi, and Sebastian Möller. "Influence of Language in Speech Quality Studies in Crowdsourcing". In: *submitted to: 13th International Conference on Quality of Multimedia Experience (QoMEX).* 2021

The studies conducted in the publications above were designed and executed by the author. The author is also responsible for the collection and analysis of the results, as well as the writing process. The co-authors contributed valuable discussions to shape the studies' settings, the analysis of the data, and publications proofreading.

Furthermore, the papers [36] and [38] are a fundamental part of the Section 4.1. The publication [40] is part of Section 3.3.4, whereas [41], [42], and [43] are a crucial part of Chapter 5. Finally, [43] is a fundamental part of Chapter 6.

- Rafael Zequeira Jiménez, Anna Llagostera, Babak Naderi, Sebastian Möller and Jens Berger. "Modeling Worker Performance Based on Intra-rater Reliability in Crowdsourcing : A Case Study of Speech Quality Assessment". In: *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX).* 2019, pp. 1–6. DOI: `10.1109/QoMEX.2019.8743148`
- Rafael Zequeira Jiménez, Anna Llagostera, Babak Naderi, Sebastian Möller and Jens Berger. "Intra- and Inter-rater Agreement in a Subjective Speech Quality Assessment Task in Crowdsourcing". In: *Companion Proceedings of The 2019 World Wide Web Conference.* WWW '19. New York, NY, USA: ACM, 2019, pp. 1138–1143. ISBN: 978-1-4503-6675-5. DOI: `10.1145/3308560.3317084`. URL: `http://doi.acm.org/10.1145/3308560.3317084`

The above two publications are related to this thesis with respect to the research question on test structure. Precisely, the impact of task repetitions addressed in Section 4.2. The author was responsible for all necessary steps for the studies conducted in crowdsourcing, i.e., study design and supervision, data collection, analysis of the results, and paper writing. The laboratory study was planned and executed by Anna Llagostera and Jens Berger. They also wrote the laboratory section of both publications. Furthermore, Babak Naderi and Sebastian Möller contributed to the study design and discussion of the study results.

Additionally, the author of this thesis was strongly involved in the activities at the ITU-T SG12 in the P.CROWD work item. These activities led to the ITU-T recommendation P.808 [25]. The following is a list of the ITU-T contributions the author was part of, which are also a key component of Chapter 4 and Chapter 5 of this dissertation:

- Rafael Zequeira Jiménez and Sebastian Möller. *Investigating the Influence of Number of Stimuli in Speech Quality Assessments in Crowdsourcing.* ITU-T Contribution SG12-C.290. CH-Geneva: International Telecommunication Union, Nov. 2018, pp. 1–8
- Babak Naderi, Sebastian Möller, and Rafael Zequeira Jiménez. *Evaluation of the Draft of P.CROWD Recommendation.* ITU-T Contribution SG12-C.290. CH-Geneva: International Telecommunication Union, Nov. 2018, pp. 1–8
- Rafael Zequeira Jiménez, Babak Naderi, and Sebastian Möller. *Influence of environmental background noise on Speech Quality Assessment in a simulated Crowdsourcing scenario.* ITU-T Contribution SG12-C.0425. CH-Geneva: International Telecommunication Union, Nov. 2019, pp. 1–9

2

Related Work

This chapter reviews relevant work about speech quality studies in crowdsourcing, with a focus on literature addressing the impact on the results of factors related to the test structure, the effect of environmental background noise, and the influence of language differences in subjective listening tests.

2.1 Number of Stimuli

Participants in a laboratory study usually evaluate the entire dataset. In contrast, workers in crowdsourcing are presented with just a portion of the samples under investigation. In this way, experimentation time can be kept short while avoiding the participant's boredom [1, 2]. However, such experimental fragmentation leads to a mixed within-between-participants study design. Consequently, workers can easily abandon a multi-part experiment and limit themselves to grading a subset of the available samples [22, 50].

Multiple works in the literature propose the use of crowdsourcing to analyze different aspects of audio [19] or speech [51, 22, 52, 53, 54, 55, 24, 56] applications. It is common to all these articles that the results collected in crowdsourcing were highly correlated to previously gathered laboratory outcomes. Still, the number of stimuli per task was different in most of the cases.

The following papers are presented focusing on the number of stimuli that were used in those studies. It is reported whether the results correlated with those of the laboratory and whether there was worker drop-out due to the high number of samples submitted in crowdsourcing:

- Authors in [22] carried out a study in crowdsourcing where listeners employed a discrete five-point scale to rate the naturalness of synthesized speech. Crowd-workers were presented with tasks containing eight to ten stimuli that were between three and five seconds long. The results were strongly correlated (r=0.95) to paid participants in the laboratory. Additionally, the authors measured repeatability by conducting a second study. They found a decrease in the correlation coefficient (r=0.92 to r=0.78) when comparing the laboratory results with those of workers who used headphones and those who employed loudspeakers.

- Moreover, [52] analyses word recognition in noise. The authors described the outcome of a web-based listening study designed to discover consistent confusions between words presented in noise, alongside an identical task performed using traditional laboratory methods. A web interface was prepared so listeners could evaluate 50 stimuli in less than three minutes in crowdsourcing. The correlation between a subset of the crowd-workers and the laboratory results was strong and significant (corr=0.8) despite the relatively high number of stimuli presented to the listeners in the crowd (i.e., 50).

- Research in [55] employed Amazon Mechanical Turk (AMT) to gather speech ratings regarding /r/ misarticulation in single word utterances. Naive listeners in AMT rated 100 stimuli. The authors found a high agreement (r=0.98) with ratings provided by speech experts in the laboratory.

- Researchers in [24] used mobile-crowdsourcing to investigate the influence of trapping questions in a speech quality assessment task. Crowd-workers were confronted with tasks consisting of six stimuli (9s long on avg.), and a high correlation was achieved between the laboratory Mean Opinion Score (MOS) and the crowdsourcing MOS ($\rho = 0.909$).

- Authors in [19] investigated the viability of AMT for the subjective evaluation of audio with intermediate impairments, a.k.a, MUSHRA [57]. The stimuli were five seconds long, and workers evaluated ten samples per task. The results in terms of overall audio quality were correlated (r=0.78) to previous ratings collected in the laboratory.

Furthermore, work in [56] evaluates the use of AMT for spoken word recognition. Although crowdsourcing and the laboratory results were strongly correlated (r=0.87), the workers were presented with 100 stimuli, and a drop-out was seen in the crowdsourcing experiment. As well, authors in [51] experienced a drop-out rate of 15% when presenting 50 stimuli per task in a crowdsourcing study investigating the intelligibility of synthetic speech.

Section 4.1 of this dissertation focuses on finding an optimal compromise between the number of stimuli and results' accuracy. The goal is to boost workers' performance and mitigate drop-out effects in speech quality studies carried out in crowdsourcing.

2.2 Worker Performance and Task Repetition

As previously noted, crowdsourcing workers are normally presented with a portion of the dataset, which leads to a fractioned experimental design. Hence, a worker would need to perform multiple tasks to evaluate the entire dataset.

The quality of experimental results gathered in crowdsourcing is frequently a function of the workers' performance. This idea has been explored in the literature from different viewpoints. Research in [58] employed Amazon Mechanical Turk to examine crowd-workers' performance in the context of the subjective evaluation of search results. The authors investigated performance in terms of the aggregated majority voting accuracy. Work in [59] also recognized workers' performance as a function of accuracy. For that, they evaluated implicit and explicit training within different common micro-tasks types.

Researchers in [60] introduced the *"SOS hypothesis"* to measure consistency in subjective QoE measurements. This hypothesis models a square relationship between the standard deviation of the opinion scores (SOS) and the Mean Opinion Score. The authors proposed this approach as an alternative to audit the reliability of the test results. Also, they encourage its use to test outcomes comparability across multiple QoE studies in crowdsourcing.

Researchers in [61] also studied result comparability over different studies. The authors carried out four subjective listening tests in three different laboratories to investigate inter- and intra-lab test result repeatability. The speech stimuli were arranged according to Recommendation P.800, and listeners assessed an English dataset containing 22 speech degradation conditions, e.g., wideband versions of AMR, 3GPP TS 26 071, EVS, and selected background noise cases such as cafeteria or road. All listeners were native English speakers from the United States. The Pearson correlation between tests was high and above 0.97 in all cases.

However, the test methodology in [61] was P.835 [62], whereas the experiments in this dissertation follow Recommendation P.808 [25]. ITU-T P.835 is particularly suitable for samples processed by noise-canceling algorithms. Listeners repeat the assessment of each speech sample three times and focus on a different aspect of the sample quality during each playback. Furthermore, the authors found that listeners provided higher overall MOS scores when confronted with a broader set of speech degradation conditions.

On the other hand, authors in [63] studied the influence of test duration on users' fatigue and the reliability of the subjective quality ratings. The fatigue data was gathered objectively by physiological means and subjectively by a questionnaire. The authors performed an analysis based on measures from common QoE laboratory studies involving different task profiles, e.g., audio, video, and web browsing. Intra- and inter-rater reliability was calculated, and the researchers demonstrated to what extent the test duration can be increased without compromising the users' ratings reliability.

Furthermore, authors in [64] used Bayesian networks to model crowd-workers as a function of human capabilities they expose when executing crowd-sourced tasks. Authors employed such a model to estimate workers' performance on new tasks. They aimed at improving task assignments eventually. However, contrary to the work carried out in this dissertation, they focused on different tasks such as fact verification, image comparison, and information extraction.

Research in [65] also investigated workers' performance in crowdsourcing markets, but as a function of users' intrinsic and extrinsic motivation. The authors demonstrated that the crowd-workers' accuracy improves significantly with intrinsic motivators when the extrinsic motivation is low. Moreover, work in [31] employed Amazon Mechanical Turk to evaluate features from micro-tasks in crowdsourcing that would influence the performance of workers.

The study of intra- and inter-rater reliability in speech quality experiments in crowdsourcing is relatively poor. Often, workers evaluate just a portion of the dataset under test. Such a study setup makes it difficult or even impossible to compute the intra- and inter-rater agreement. However, this is usually the case to keep test sessions short [1] while avoiding boring the workers [2]. For instance, participants of the speech quality assessment study carried out in [24] and in [23] evaluated just 2.5% of the speech dataset every time they participated in the crowdsourcing experiment; listeners of the study described in [22] only judged 3.1% of the available samples each time they assessed the naturalness of synthesized speech. Nevertheless, the authors repeated the experiment but only to prove the validity of the framework they proposed.

Additionally, even when workers have the option to evaluate the entire speech dataset, sometimes they choose to take part in the study just one or two times, or other times the available jobs are completed faster by multiple different workers. As a result, almost no participant evaluates the dataset completely. For example, in the experiment carried out in [36] and detailed in Section 4.1, more than 200 workers participated in the study, and only two evaluated the entire speech dataset. Thus, no analysis was made regarding intra- or inter-rater reliability.

To the best of the author's knowledge, no study has reported the use of intra-rater reliability as an input variable to determine workers' performance, especially in the speech quality research domain in crowdsourcing. Section 4.2 presents an experiment where listeners had the chance to evaluate four times all the speech stimuli in the dataset. Hence, enough data is collected for analyzing the agreement between and within-subjects. Additionally, a novel approach is proposed to model the relationship between the consistency of the workers' ratings at different time points and the overall performance.

2.3 Environmental Background Noise

Previous work endeavored to investigate the influence of the environmental background noise in speech quality assessments carried out in crowdsourcing. Authors in [32] performed a study in which participants rated the quality of speech samples first in the laboratory and afterward in a specific crowdsourcing scenario, i.e., cafeteria, metro station, or living room. In their experiment, 14 participants executed a P.800 [66] listening test in the laboratory with simulated "crossroad" noise, and in crowdsourcing also 14 workers performed the listening test under the "living room" test condition. All in all, researchers discovered that the presence of a "cafeteria" background noise at 62.7dB(A) or a "crossroad" noise at 64.7dB(A) would decrease the correlation to ratings collected through a traditional P.800 listening test.

However, the authors focused on two main factors: the effect of the employed playback hardware, i.e., users' regular headphones vs. professional ones, and mobile-crowdsourcing, where crowd-workers have the freedom to work from many different places. In contrast, this dissertation concentrates on web-crowdsourcing. It has been exposed that users of the main crowdsourcing platforms work mostly from home [67]. This fact limits the range of noises workers might be exposed to when they perform crowd-work.

On the other hand, work in [33] examined the impact of the information contained in background noise from environmental sources on the speech quality perception in VoIP applications. The authors investigated the interaction between multiple background noise conditions and realistic network impairments such as coders and packet loss. The tested background noises were: the voice from a TV source, cocktail party, circuit noise, restaurant, and city noises. Researchers discovered that speech quality perception could be affected by the meaning of the noise in telephony contexts and not by the interaction between the percentage of packet loss and the different noises.

Work in [68] also investigated the listener perception of speech affected by different environment background noise conditions. Multiple P.800 [66] speech quality studies were conducted in different laboratories. The authors discovered that users perceived the quality of the speech higher on those samples affected by the “Vehicle” noise condition in comparison to those degraded by the “Street” and “Hoth” noise condition.

Similarly, researchers in [69] studied the speech quality perception for situations in which different noises are present in the speech signal. They attempted to determine the link between different noise types and their impact on the quality judgments. The authors carried out a listening test to gather quality scores for a set of speech samples mixed with different background noises. They then executed a cluster analysis on the quality ratings and identified three classes based on the noise impact. Their experiments suggest that it is possible to create and predict noise classes based on the speech quality scores.

Differently from the user perception of speech affected by noise, this dissertation examines the influence of environmental background noise on the listener’s side, on the speech quality perception in crowdsourcing scenarios.

Furthermore, speech intelligibility under different background noise circumstances is another challenging research stream that has received significant attention in recent years. [70] tackled the problem of speech recognition in ecologically valid natural background noise scenarios. The authors found that the vowel identity was mostly preserved despite the noise under test, and they also confirmed the functional role of consonants during lexical identification.

Work in [71] assessed three different speech coding standards regarding intelligibility in near-end background noise and packet loss conditions. Researchers in [72] studied the user perception of speech containing ambient background noise. Specifically, they looked at how people value speech-based take-over requests as a function of background noise, speakers’ gender, speech rate, and emotional tone. To this end, the authors carried out an experiment in crowdsourcing where workers rated speech samples according to urgency, commanding, and pleasantness. Researchers found that the female voice was easier to understand under background noise conditions.

Still, to the best of our knowledge, the effect of environmental background noise on speech quality evaluations performed in web-crowdsourcing has not been studied so far.

2.4 Influence of Language Differences

Multiple studies in the literature have analyzed different aspects of conducting an audio listening test with users of different nationalities. Work in [68] presents the results of multiple P.800 speech quality evaluation tests carried out in multiple laboratories with listeners from different countries. The participants assessed the quality of speech stimuli in their native language. The authors found that Japanese listeners provided lower quality scores per condition than the native French, English, German and Norwegian speakers. This outcome suggests that the speech quality scores collected from listeners of a particular nationality may be biased when compared to the same ratings provided by a different demographic group of users.

The research in [73] presents the results of a speech quality assessment test in which users evaluated stimuli that were in a language other than their mother tongue. Listeners were either native Czech, Slovak, or Italian speakers. They assessed the quality of an English speech database comprising different codecs and two noise conditions. The authors found that listeners with insufficient English knowledge (i.e., beginner and intermediate level) rated the speech quality systematically lower than the participants with an advanced level. Researchers believed this outcome was due to the listeners' inability to understand what was being said in the stimulus, even in the less distorted samples. Therefore, they provided lower quality scores.

Similarly, French listeners in a speech quality study carried out in [74], rated systematically lower the quality of speech items that they were able to understand but lacked semantical meaning. In contrast, the German listeners participating in the study that could not understand French scored similarly the quality of both types of French speech stimuli, i.e., those with no semantic meaning and those carrying typical telephone content.

Contrary to the experimentations conducted in this thesis, the listening test in [73] and in [74] was carried out in the laboratory following Recommendation P.800, whereas I executed the listening test in crowdsourcing according to Recommendation P.808 [25]. Additionally, no insight was given in [73] about how the quality ratings from native English speakers compared to the ratings from the non-native listeners.

Furthermore, authors in [75] carried out a speech quality study to investigate the listener's perception of a speech dataset containing sentences of American English. There were two groups of participants, i.e., native English listeners from the United States and native Igbo speakers of an African tone language. The authors discovered that Igbo subjects overestimated the impact of the noise on the speech samples' quality. They were more disturbed by the additive noise than other degradations compared to the native English participants. In contrast, low-level listening or attenuated impairment conditions did not significantly affect the perceived speech quality.

On the other hand, work in [76] examines whether a foreign language influences the ratings and listening times in a subjective evaluation of audio with intermediate impairments, a.k.a, *"MUSHRA"* [57]. The authors performed a study with native German and Mandarin Chinese speaking listeners and items of these two languages. They discovered that for high audio quality items where the sample artifacts are difficult to perceive, non-native listeners executed more comparisons and needed 20% more time to conduct the listening test. Moreover, no overall difference was found between the ratings of the native and non-native listeners.

Unlike our work, the quality of items in a *"MUSHRA"* test is higher, and intelligibility is not an issue. Additionally, for high-quality items, listeners typically focus on small audio sections and listen to it repeatedly. Often these sections are as short as half a second. Thus, any semantic meaning is lost, and understanding does not play a role. In contrast, I investigate how listeners of different mother tongues perceive degraded speech German stimuli.

2.5 Conclusion

This chapter reviews the literature for relevant research addressing the influence of factors related to the test structure, the effect of environmental background noise, and the influence of language differences in speech quality assessment studies.

The test structure relates to the number of stimuli to include in a speech quality assessment task in crowdsourcing. This decision is frequently made based on a rule of thumb about how many samples can be squeezed into a test without compromising the result quality and the test duration [1]. To the best of the author's knowledge, no studies have been conducted in which the influence on the results of using a different number of stimuli in a speech quality assessment context has been analyzed. The findings of this dissertation in this respect, will help the research community when designing subjective user studies to find a good compromise between the number of speech samples, task length, and result reliability.

Additionally, this chapter examines different articles addressing the effect of background noise on speech quality assessments. However, most of the cited work analyses the listeners' perception in scenarios where the noise is coupled to the speech signal, or situations where the noise affects the speech intelligibility. As yet, to the best of the author's knowledge, the effect of environmental background noise on speech quality assessment studies performed in web-crowdsourcing has not been studied so far.

Finally, this chapter reviewed the literature for research where participants of a study evaluated speech or audio material in a language different from their mother tongue. Still, no studies have been found that analyze the influence of presenting a German speech dataset to English or Spanish listeners in crowdsourcing.

3
Method

This dissertation investigates the effect of different factors influencing the results of speech quality assessment studies conducted in crowdsourcing. To this end, I use the listening-only quality of transmitted speech rated in an Absolute Category Rating paradigm, as this is the most popular rating method used in practice. This chapter introduces relevant information regarding the test procedures in crowdsourcing and the laboratory necessary to understand the studies conducted throughout this thesis. Additionally, the speech databases employed in most of the experiments are presented.

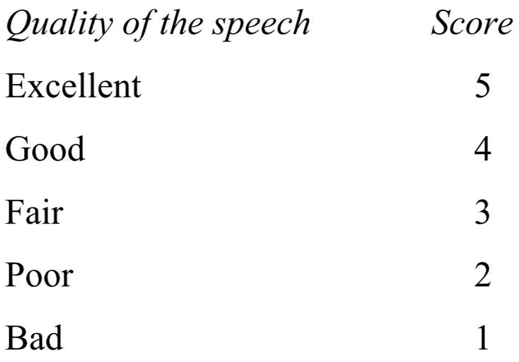
3.1 Laboratory Test

A common means to study and understand the Quality of Experience of telephony services is by carrying passive subjective experiments with human participants in a laboratory context. These experiments are the so-called *"Listening-Only Test"* (LOT), where overall quality ratings are collected on a five-point Absolute Category Rating (ACR) scale.

The most frequent listening-only test employed in the laboratory is the overall quality evaluation using an ACR paradigm. The test stimuli are presented individually, and listeners express their opinion about the overall quality on a unipolar five-point rating scale with numeric values 5, 4, 3, 2, and 1, and the corresponding labels: "excellent", "good", "fair", "poor", and "bad" (or their respective language-specific counterparts). Figure 3.1 presents this five-point ACR scale as defined in [66].

The average of the ratings from several test participants per stimulus results in an absolute quality metric called *"Mean Opinion Score"* (MOS). The scores are commonly reported on a per-stimulus and a per-test-condition level, and a test condition corresponds to the treatment of a source stimulus, e.g., due to transmission impairments and coding effects.

The International Telecommunication Union Telecommunication Standardization Sector (ITU-T) has published recommendations and guidelines for carefully conducting subjective speech quality assessment studies in the laboratory, i.e., ITU-T P.800 [66]. The recommendation states that "balanced" speech material, "normal" speakers, as wells as "normally-hearing" test participants should be chosen, and the test stimuli should be presented in a neutral



Quality of the speech	*Score*
Excellent	5
Good	4
Fair	3
Poor	2
Bad	1

Figure 3.1: Five-point Absolute Category Rating (ACR) scale as defined in [66] that is used for Listening-Only Tests.

reproducible situation. The test room guidelines define that the experiment should be carried out in an acoustically treated listening environment with limited sources of external noise and reverberation.

Furthermore, the test participants should be selected according to their perceptual abilities and experience with listening tests. Then, they get invited to the laboratory, carefully instructed about the listening test, monitored during the experiment, and debriefed after the test. This procedure ensures constant and controllable conditions across participants and increases reliability. However, little effort is taken to mimic real-life usage situations, e.g., simulating background noise, asking content-related questions. Therefore, limiting the ecological validity of the test results to some extent. Such a gap is meant to be shortened by crowdsourcing-based speech quality evaluations.

3.2 Speech Database

The stimuli employed in the studies described in this dissertation were taken from database number 501 and 502 from the ITU-T Recommendation P.863 [9], competition. And also from the ITU-T Recommendation P.501 Annex D [77]. These databases were kindly provided by SwissQual AG, Solothurn, for research purposes.

3.2.1 SwissQual 501

The database SwissQual 501 includes different types of speech degradations that were created following the ITU-T Recommendation P.863. Four German speakers were recorded per condition uttering four different German sentences. Overall, 200 stimuli were arranged, accounting for 50 speech impairments conditions. These degradation conditions represent different audio bandwidths (narrowband 300-3400 Hz, wideband 50-7000 Hz, super wideband 50-14000 Hz), temporal clipping, signal-correlated as well as uncorrelated noise, speech coding at various bitrates, packet loss with multiple temporal loss profiles, ambient background noise of diverse types, different frequency distortions, and also, combinations of these degradations. Table A.1 in Appendix A presents a summary of these speech impairments conditions.

The database encloses subjective quality assessments to the 200 stimuli made by 24 different native German listeners. These quality scores were gathered in a laboratory following the

ITU-T Recommendation P.800 [66]. The resulting Mean Opinion Scores (MOS) for each stimulus and condition are taken as a reference for the analyses made in the multiple studies carried out in this thesis (from now on referred to as "Lab-MOS").

3.2.2 SwissQual 502

The speech database SwissQual 502 is very similar to SwissQual 501, with slightly different but comparable speech impairment conditions. This database was also created for developing the ITU-T Rec. P.863 [9]. It contains 50 speech degradation conditions, e.g., send-side ambient background noise, white background noise, different audio bandwidths (narrowband 300-3400 Hz, wideband 50-7000 Hz, super wideband 50-14000 Hz), speech coding at various bitrates, and combinations of these degradations.

This database was employed in the study detailed in Section 5.1. Out of the 50 conditions, only 16 were used due to the time constraints of the study. Table B.1 in Appendix B presents a summary of these 15 speech impairments.

SwissQual 502 also includes subjective quality evaluations made by 24 different native German listeners, following ITU-T Rec. P.800. I use the resulting MOS scores per file and condition as a reference for the different analyses made in this thesis's studies.

3.2.3 SwissQual P.501 Annex D

The database SwissQual P.501 Annex D is a mixed fullband set of samples containing audio bandwidths from below narrowband and up to fullband. The speech stimuli are encoded with state-of-the-art codecs under ideal and live good/average/bad coverage situations. Furthermore, 71% of the speech samples are live conditions from real-field recordings, whereas 29% are based on offline processed speech and anchor conditions. The full list of these conditions is presented in Table C.1 of Appendix C. More details about the speech material can be found in [78].

3.3 Crowdsourcing Test

3.3.1 Standardized Evaluation Method for Speech Quality in Crowdsourcing

The ITU-T has recently published Recommendation P.808 [25] for subjective speech quality assessments in crowdsourcing. This standard considers the fundamental differences between laboratory and crowdsourcing, as well as information about the experimental design, test material, and the procedure for conducting ACR listening tests in crowdsourcing. The Recommendation does not provide information on other listening opinion tests, such as Degradation Category Rating (DCR) and Comparison Category Rating (CCR), as they are the subject of future research.

P.808 encourages dividing the test procedure into three phases, i.e., *"qualification"*, *"training"*, and *"assessment"*. The *"qualification"* phase serves to test the eligibility of the test participants. The *"training"* phase should contain a temporal restriction and anchoring conditions to prepare the listeners for the evaluation task. Finally, the *"assessment"* phase includes a small set of stimuli, and participants give their opinion about the overall quality on the corresponding ACR scale.

Recommendation P.808 es the outcome of years of research regarding crowdsourcing for speech quality assessment studies. This research has been conducted by the "Quality and Usability Lab" of the "Berlin Institute of Technology" (from which the author is a member) and other laboratory partners. Thus, the methodology defined in the following Subsection 3.3.3 for executing speech quality evaluations in crowdsourcing is in line with the specifications of P.808.

3.3.2 Crowdsourcing Platforms

The crowdsourcing platforms maintain a dedicated crowd of workers and provide the necessary infrastructure like a pool of tasks, payment mechanisms, and in some cases, additional services like quality control or worker selection mechanisms [30].

The crowdsourcing platforms employed to conduct some of the studies detailed in this dissertation are Amazon Mechanical Turk (AMT)[1] and clickworker[2]. AMT is the best known, used, and researched platform within the academic literature and in a more public context [67]. Due to the platforms' payout policy[3], the vast majority of its crowd-workers are from the United States and India[4].

On the other hand, clickworker is a crowdsourcing platform based in Germany that claims to have a crowd of more than 2.2 million workers from all over the world[5]. Most of their users come from North America (46%), Europe (30%), and Asia (15%). And overall, 47% are native English speakers and 12% native German speakers.

3.3.3 Test Setup and Procedure

The experiments carried out in this thesis were executed on multiple crowdsourcing platforms and in the laboratory. To ensure a consistent test layout across the different studies, an HTML JavaScript-based framework[6] was implemented to administer the listening test, in conjunction with a Node.js server for the data collection.

The speech quality assessment studies conducted in crowdsourcing contain three main phases, i.e., *Qualification*, *Training*, and *Assessment*. These phases were implemented in the framework mentioned above and are detailed in this subsection.

3.3.3.1 Qualification

The ***Qualification*** phase serves to screen the workers' population to find suitable participants for the listening test based on their mother tongue. Additionally, basic demographics information is collected at this point, i.e., gender, age group, country of residence, and mother tongue. Specifically, the Qualification consist of a short introduction to the study, a consent request for the data being collected, a demographic questionnaire, and a quiz to test the German language knowledge. This test comprises three audios that are 23 seconds long on average. Workers then listen to a German passage and are asked to select the correct

[1] https://www.mturk.com

[2] https://www.clickworker.com

[3] https://www.mturk.com/mturk/help?helpPage=worker

[4] http://demographics.mturk-tracker.com

[5] https://www.clickworker.com/clickworker-crowd

[6] https://gitlab.com/zequeira/SQAT-Cr.git

statement out of three options. Workers failing this quiz are not invited to the study. Finally, workers answer whether or not they have been involved in any type of listening test in the last year. Those responding positively to this question are prevented from participating in the study [66].

3.3.3.2 Training

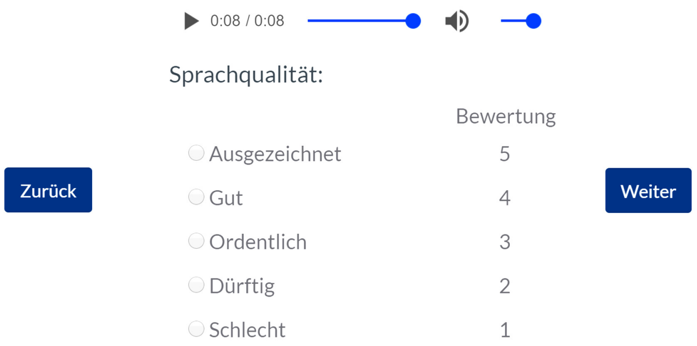
The ***Training*** phase permits to check whether the workers' system is optimal for conducting the listening test. Listeners specify how they use their headphones, i.e., in-ear or over-ear. They are not allowed to use loudspeakers. Otherwise, they would present a smaller discrimination capacity [22]. Additionally, workers are presented with a short math exercise with digits panning left to right in stereo. Then, they insert the total sum in an input field. This way, the two-eared usage of headphones is controlled. Furthermore, they play a short audio clip and set their device volume to a comfortable level. After this, they are instructed not to change the volume; otherwise, the results would be invalidated. Finally, workers listen to five speech stimuli carefully selected from the dataset to cover the entire MOS range. Thus, they could get to know what to expect on the rating task while becoming familiar with the interface. They are not informed about the samples' quality to avoid bias on the ratings. Additionally, this anchoring step serves to overcome the scale usage problem reported in previous studies [79]. All the questions are mandatory, and the workers failing the math exercise are prevented from participating in the study for 12 hours. On the other hand, when they pass the Training, they are granted an hour-long time frame. During this time, workers can execute the assessment as long as it was available. After one hour, they need to perform the Training again to continue participating in the study.

3.3.3.3 Assessment

The ***Assessment*** phase is presented immediately after the workers complete the Qualification successfully. They are informed that each speech sample contains one or two sentences and that they just need to rate the overall speech quality on the provided five-point scale, see Figure 3.2. Workers could not give their opinion on the scale unless they listened first to the speech sample. They could not go forward until the audio was played entirely and one option selected on the scale. They could listen to each speech sample as many times as they wished.

Inspired by work in [24], one trapping question (TQ) is inserted randomly within the first five stimuli from every ten speech samples. This TQ consists of a speech stimulus that starts as the rest of the stimuli but is interrupted after four seconds and includes a new voice that requests the worker to select one specific option in the five-point scale (the TQ's Graphical User Interface (GUI) is the same as the rest of stimuli, Figure 3.2). The TQ includes as well a motivational message and highlights the value of the listeners' work. When workers fail the TQ, the ratings from the set of those ten stimuli are considered unreliable. Like when failing the training phase, they are prevented from participating in the study for one hour.

All the text, feedback messages, and labels in each study were translated to match the target audience's language.



Sprachqualität:

	Bewertung
○ Ausgezeichnet	5
○ Gut	4
○ Ordentlich	3
○ Dürftig	2
○ Schlecht	1

Figure 3.2: Graphical interface employed by the listeners for the speech quality assessment in crowdsourcing (in line with [66]). The text translate from German: "Speech Quality" and "Rating". The scale (in descending order): "Excellent", "Good", "Fair", "Poor", and "Bad".

3.3.4 Environment

As previously stated, users in crowdsourcing work remotely and unsupervised. This lack of control introduces some "noise" into the study setup, which needs to be controlled for gathering valid results. Therefore, understanding the characteristics of the environments in which workers perform crowdsourcing tasks is crucial to mitigating these study impairments. Specifically, the common sources of noise and distractions they experience when doing crowd-work. The user environment characteristics are essential in audio and speech quality assessment studies as listeners might not be able to evaluate appropriately certain degradations in speech samples if they conduct the task from a place where they could be distracted. Inattentive crowd-workers are prone to work sloppy, which in turn leds to poor experimental results.

This subsection investigates the environment characteristics of crowd-workers from German-speaking countries. To this end, a study was conducted in which workers gave details about the surroundings in which they usually perform crowdsourcing tasks. Audio and visual data was collected per user, which contributed to aggregate more information on the users' input.

3.3.4.1 Study Setup

The German-based clickworker crowdsourcing platform was used. Clickworker has an active pool of workers from Germany, Austria, and Belgium. Thus, a good fit for the experimental needs. The study consisted of three phases, i.e., "Audio Recording Setup", "Environment Video Recording", and "Environment Questionnaire". A three-pages HTML JavaScript-based framework with a Node.js back-end to collect the data was implemented.

I targeted users from Germany, Austria, and Belgium. Overall, 325 crowd-workers participated in the study. Unfortunately, a number of them dropped out of the experiment at different stages. Consequently, it was not possible to collect the same amount of data at each study phase.

3.3.4.2 Audio Recording Setup

This phase consisted of a tutorial that guided workers in deactivating some noise reduction options that are frequently enabled by default on Windows and macOS computers. This step was important for collecting accurate audio recordings of the workers' environmental scene.

They were also asked to submit a screenshot of their system's configuration to show they had the proper audio recording setup. Otherwise, the recordings would be too corrupt to extract any useful information from them.

A total of 248 workers submitted a screenshot and 183 of them from a Windows computer. I noticed that 54% of the participants configured their computer correctly, whereas 23.4% failed on this task. Some of these workers did not pay attention to the instructions and selected the wrong option on the setup. Others uploaded an image file different from the requested screenshot. The remaining 22.6% could not set the requested audio configuration, either because their system was old or because the options they needed to select were not available on their computers. Uploading the screenshot was a requirement to proceed to the next phase.

3.3.4.3 Environment Video Recording

Once crowd-workers uploaded the screenshot, they continued with the second phase in which they recorded and submitted a short video of their current environment. For this purpose, I prepared an example video so that they could see how to do the recording. They were advised to perform the recording with their smartphone. Additionally, they were instructed to avoid recording any element that allows them to be identified nor other people, and neither documents containing confidential data.

These videos were analyzed to verify that the recorded scene corresponded with the one reported by the workers in the questionnaire (Subsection 3.3.4.6). I collected a total of 230 videos from different crowd-workers.

I recognized a house's room in 82.17% of the recordings. Most frequently, a living room, bedroom, workroom, or kitchen. I also identified an office space in 3.48% of the videos. In 6.09% of the cases workers failed to do a proper recording, and no scene was identified. Also, 6.52% of the participants submitted a video different from what we requested.

Moreover, a few workers were not concern about privacy, and despite our instructions, they recorded themselves or other people in the room. Additionally, I found that workers tended to turn down the TV, radio, or music volume prior to recording, as the video audio was low or quiet compared to the web audio recording.

3.3.4.4 Environment Questionnaire

Finally, after workers submitted the video, they proceeded to the last phase to answer questions regarding their place's environmental characteristics, where they usually perform crowdsourcing tasks. They also gave information about the environment in which they were at that moment.

Additionally, a JavaScript code embedded within the first question permitted to record the workers' environment background noise for 15 seconds. This data was used to check whether the information they provided in the questionnaire correlated with the sounds captured in the recordings.

3.3.4.5 Audio Recording Analysis

A total of 131 environmental background noise files from different workers were collected. I hypothesize that some users did not grant access to their microphone or that hardware or browser issues happened on the worker side, which prevented the recording from happening; therefore, the low number of files. These audios were labeled manually according to the sound they contained and whether they carried any information. Table 3.1 presents the labels assigned to these background noise recordings.

I found that 43.08% of the recordings were corrupted, and it was not possible to identify any sound. According to the screenshots of these workers' audio setup, 19.6% failed to configure their system correctly, while 32.1% selected the instructed options. For the rest of the recordings (48.2%), a screenshot was not provided.

It is important to note that some of the recordings were made with a properly configured system, yet the audio files did not carry any information. We believe that the microphone of these users was defective. Moreover, we labeled some files as "quiet" because the background noise was low or non-existent. More information is presented in Table 3.1. In addition, we detected in some files noises from the kitchen, bird sounds, mobile ring tones, ventilation systems, water flowing, and radio. However, these noises were detected in only one recording in most cases and, therefore, not included in Table 3.1.

Table 3.1: Labels assigned to the audio recordings in relation to the type of noise that could be heard on them.

Background Noise	Number of Files	Percentage
not defined (NA)	56	43.08%
quiet	31	23.85%
TV	18	13.85%
electric device	11	8.46%
music	8	6.15%
street noise	4	3.08%
people talking	3	2.31%

3.3.4.6 Questionnaire Results

A total of 213 workers submitted answers to the questionnaire. 93.4% reported *Home* as the regular workplace where they usually perform crowdsourcing tasks, whereas 3.8% execute the tasks from *Work*. This information seems accurate since 86.9% of them expressed being at home while running our study, and 7.5% selected *Work* as their current workplace.

Moreover, 79.6% of the workers reported being alone while conducting our experiment, while 14.7% announced that there was one more person in the same room. They also reported an estimate of the number of hours per week they spend performing crowdsourcing tasks. I discovered that approximately 60% of the participants spend one to three hours per week conducting crowdsourcing tasks. Table 3.2 presents the questionnaire results.

Additionally, workers responded to (translated from German): *"What other tasks or activities do you normally conduct when you run Clickworker Jobs?"*. Answers were collected in a multiple-choice question format with an open text field. 24.8% of the crowd-workers stated that they listen to some music, 18.3% take part in other activities on their computer, 16.3% execute other crowdsourcing tasks, 15.1% watch TV, and 10.8% get into social media. Only 1.2% reported not taking part in any other activity. See Table 3.3 for more details.

Table 3.2: Answers to the questionnaire from 213 workers. Information regarding the workers' regular workplace, workers' current workplace, number of persons on the same room while executing the study, and the amount of hours per week workers spend doing crowdsourcing tasks. Numbers are expressed in percentages.

Regular Work Place	
Home	93.4%
Work	3.8%
onTheGo	1.9%
Cafeteria	0.5%
Other	0.5%
Current Work Place	
Home	86.9%
Work	7.5%
Other	3.3%
onTheGo	2.3%
Persons in the Same Room	
Alone	79.6%
One	14.7%
Two	4.3%
Three	1.4%
Hours/Week doing Crowdsourcing	
1	23.5%
2	23.9%
3	10.3%
5	9.9%

The last two questionnaire items were two open-ended questions. First, crowd-workers described their current environment and reported the sounds or noises they could hear at that moment. Afterward, they reported the noises that usually distract them when performing tasks on clickworker or other crowdsourcing platforms. I collected these answers in an open text field, which I then analyzed with the "IBMgermanb® SPSSgermanb® Text Analytics for Surveys" (v4.0.1) tool. This software uses Natural Language Processing (NLP) to code the input text into terms (either simple words or phrases), from which categories are also developed.

This analysis revealed that 21.6% of the crowd-workers could hear cars or street noises, 18.8% the TV, 14.6% of them were in the presence of other people talking, 13.6% reported being in a quiet place, and 13.6% could hear the noises produced by their computer. More details can be found in Table 3.4(a).

This information is in line with the noises extracted manually from the environmental noise recordings (see Table 3.1). In both tables (Table 3.1 and Table 3.4(a)), the same type of noise was identified, e.g., TV, car or street noise, music, people talking, quiet, etc. This suggests the validity of the background noise recordings and the workers' answers regarding their current environmental noise.

To the second open-ended question, workers reported that when conducting crowdsourcing tasks, they get distracted frequently by other people (33.3%), phone calls (28.6%), and their

Table 3.3: Answers to the question: *"What other tasks or activities do you normally conduct when you run Clickworker Jobs?"*, from the 213 workers that submitted the questionnaire.

Home Other Activities	%
listening to music	24.8%
other activities in computer	18.3%
other crowdsourcing tasks	16.3%
watching TV	15.1%
social media	10.8%
listening to the radio	6.0%
texting, chatting	3.8%
talking with people	3.4%
none	1.2%
first task in clickworker	0.2%

smartphone (6.6%). 17.8% stated that they execute crowdsourcing tasks while being in a quiet place. More details in Table 3.4(b).

Table 3.4: Categories found in the open ended answers from 213 workers that reported about the noises or sounds heard while conducting our study (Table 3.4a), and the sources of distractions they face when doing crowdsourcing tasks (Table 3.4b).

(a) Noises or sounds heard.

Category	%
car	21.6
TV	18.8
people	14.6
quiet	13.6
computer	13.6
music	7.0
bird	5.2
radio	3.8
baby	3.3
keyboard	2.8
mouse click	2.3
mowing	1.4
dog	0.9

(b) Sources of distractions.

Category	%
people	33.3
phone calls	28.6
quiet	17.8
smartphone	6.6
baby	5.2
TV	5.2
music	2.8
social media	2.8
pets	1.4
WhatsApp	1.4
chat	0.9
autos	0.5

3.3.4.7 Discussion

This section investigated the environmental characteristics of German crowd-workers. I found that 86.9% of our study participants perform crowdsourcing tasks mostly from home. Additionally, I observed that the environmental background noise that might influence the results of a speech quality assessment experiment in crowdsourcing could be grouped into two main types, i.e., constant mechanic noise like car engines and street traffic noises (as reported in Table 3.4(a)), and periodic, melodic, or voiced types of noise like from TV, music, radio, or people talking. I also discovered that 24.8% and 15.1% of the workers take part in other activities when conducting crowdsourcing tasks, i.e., listening to music and watching TV, respectively.

Based on these findings, when addressing studies to German crowd-workers, it is advised to include instructions for turning off multimedia devices, i.e., TV, radio, music player, and to request workers to close windows and doors to avoid outside noises. It is also recommended to collect environment audio recordings to identify those workers not following the instructions. Moreover, these findings might be partly generalizable to western users from Amazon Mechanical Turk or microWorkers, being that societies and customs are similar. However, I cannot ensure this as a comparable experiment to this one would be needed in AMT and MW.

3.4 Simulated Crowdsourcing Test in Laboratory

Two requirements need to be fulfilled to simulate a crowdsourcing environment in the laboratory to carry out subjective user studies. First, the environmental characteristics related to background noise must be met. And second, the study participants must use their own devices to execute the study.

To simulate environmental background noise situations in the laboratory, I employed a four-speaker setup as defined in [80]. The loudspeakers were positioned two meters apart from the center of the room where the test participant was seated. The height of the four loudspeakers was set to 1.5 meters so that the center of the acoustic field (i.e., the listeners' ears while seated) is of the same height as the loudspeaker positions. [80] also gives guidelines to use a subwoofer for the noise simulation. However, I did not employ one in the setup, as the "Studio Monitor Genelec 8030A" speakers were used, which were powerful enough with 5" woofers, 110dB SPL, and frequency range from 54Hz to 20KHz. Figure 3.3 shows a diagram of the speakers' positioning.

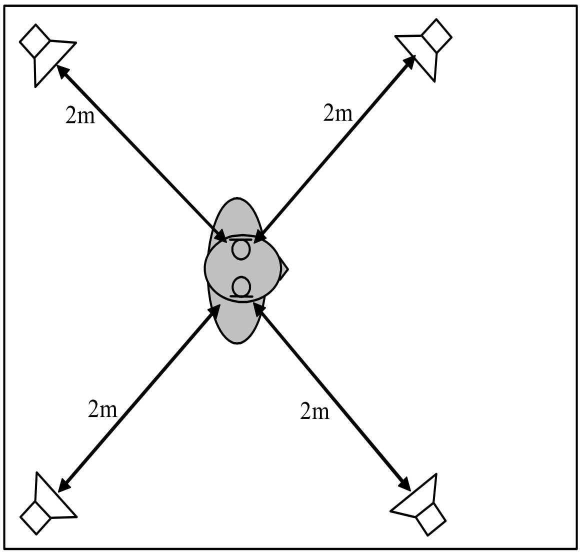


Figure 3.3: Loudspeakers arrangement in the laboratory room for environmental background noise simulation.

Moreover, I used the audio interface “FIREFACE UCX” to reproduce the noise signal, and the background noise levels were measured with the fullband artificial head HMS II.3 HEAD acoustics.

Additionally, to fulfill the second requirement, the study participants were requested to bring their computer and headset they use in everyday life to execute the listening test. To accomplish the speech quality assessment study, they connected to the Internet to access the same HTML and JavaScript-based framework[7] that I have used successfully in all speech quality studies conducted in crowdsourcing [36, 45, 46, 41]. Specifically, this framework is a web-based implementation of the three phases of a crowdsourcing speech quality assessment study detailed in Subsection 3.3.3.

3.5 Result Metrics

Reliability and validity are two important criteria that must be fulfilled by any measurement to ensure that results are robust and potentially generalizable to more natural contexts. These two criteria must also be met to guarantee that theoretical conclusions derived from these results are well-founded.

The *reliability* is defined as the accuracy of measurements, which describes how much initial test results would deviate from the results of follow-up tests (retests). For experimental studies, the reliability metric is crucial as it warrants the replication of the initial experiment with a sample drawn from the original population.

The *validity* refers to which extend a measurement actually captures its intended measurands. It relates to the truthfulness of the responses and to whether a test measures what is intended to measure. In order to measure the validity of the test answers, a “ground truth” or alike is needed. For instance, the validity of the ratings collected in a speech quality assessment task in crowdsourcing can be determined by comparing them to the results from a laboratory test, assuming that the latter is close to the “ground truth”.

Furthermore, the ratings collected in a crowdsourcing speech quality assessment study must undergo a data screening process. The submitted responses to a specific task should be discarded when listeners answer wrong the gold standard or trapping questions included. Additionally, an analysis should be made to detect and discard the ratings deemed extreme outliers, i.e., those located at a distance from the median equal or higher than $3.0 \cdot IQR$ (interquartile range) [81]. Other methods to detect outliers include the use of boxplots, which flag the extreme outliers ratings when they fall beyond the outer fence [82].

After the data screening process, the Mean Opinion Score (MOS) per stimulus and condition should be calculated. These MOS values can be complemented with the calculation of Confidence Intervals (CI). Then, the Spearman’s rank-order correlation or Pearson’s product-moment correlation and Root Mean Square Error should be computed to determine the relationship between MOS scores collected in the laboratory and crowdsourcing. These correlations provide an estimate of the validity of the ratings gathered in crowdsourcing. It gives insights into the strength and direction of the association between two continuous or ordinal variables.

[7] https://gitlab.com/zequeira/NoStimuli-SQA.git

3.6 Conclusion

This chapter details the study setup and procedures common to the different experiments carried out as part of this dissertation. Additionally, the speech material employed in the studies is presented as well as insights about the environmental conditions of the German crowd-workers. Finally, a definition of reliability and validity is given in the context of subjective speech quality testing, and guidelines are given for the data screening process of speech quality scores collected in crowdsourcing.

4

Test Structure

This chapter investigates aspects concerning the structure of a speech quality assessment task in crowdsourcing and its influence on worker performance. Precisely, I determine the optimal number of speech stimuli to include in a single task. Additionally, the impact of executing the rating task multiple times on different rater reliability metrics is analyzed.

Two sets of separate studies were conducted. In the first study, participants were assigned to one of three user-groups, each of which was confronted with tasks consisting of a different number of speech stimuli. The study setup and results are detailed in Section 4.1.

The second study was also conducted in crowdsourcing. A group of listeners was recruited to assess the overall quality of speech samples four times. Then, I analyze the impact of task repetition in the intra- and inter-listener agreement and the relationship between intra-rater agreement with the workers' accuracy. Further details can be found in Section 4.2.

4.1 Influence of Number of Stimuli

One of the fundamental differences between carrying speech quality studies in the laboratory versus a crowdsourcing environment is that listeners in the laboratory evaluate the entire dataset. In contrast, in crowdsourcing, participants assess just a portion of the stimuli per task to keep the listening session short and avoid the workers' boredom. Aspects like listeners' workload and fatigue are essential as they relate to an important question, i.e., how to optimize the study design without compromising the results' quality by tiring the test participants? This section examines the impact of the number of presented speech stimuli per task on the reliability of listeners' ratings in the context of subjective speech quality assessments.

4.1.1 Study Setup

A crowdsourcing experiment was conducted in which three groups of workers were recruited, i.e., G1, G2, and G3. Listeners in each group assessed the quality of different numbers of speech stimuli. I then computed the Mean Opinion Scores (MOS) per file and condition, and the results were contrasted with previously collected ratings in the laboratory (Lab-MOS).

4.1.1.1 Speech Database

The speech stimuli were taken from the database SwissQual 501 from the ITU-T Rec. P.863 [9], competition. This database contains 200 speech samples carrying 50 different degradation conditions. Additionally, it includes subjective quality assessments made by 24 different native German listeners, in accordance with the ITU-T Rec. P.800 [66]. The resulting Lab-MOS scores are taken as a reference for the analysis presented in this section. Further information regarding SwissQual 501 can be found in Section 3.2.1.

4.1.1.2 Method

The study was conducted in the clickworker crowdsourcing platform. Clickworker is based in Germany, and it has an active pool of workers from Germany, Austria, and Belgium. Hence, a good fit for the experimental needs.

The study consisted of three phases, i.e., Qualification, Training, and Assessment. The details of these phases are presented in Section 3.3.3.1, 3.3.3.2, and 3.3.3.3, respectively.

Differently from the specifications of Section 3.3.3.3, the Assessment task in this study contained a different number of stimuli according to each group. Listeners in G1 evaluated ten speech samples, whereas listeners in G2 and G3 assessed the quality of 20 and 40 speech stimuli per task, respectively. I selected these quantities, considering the number of samples that could be included in a task without making it too long to avoid boring workers [1]. The 200 stimuli in the dataset were between eight and ten seconds long. To assess them all, crowd-workers in G1, G2, and G3, could perform the Assessment task up to 20, 10, and 5 times, respectively. Due to the inserted trapping questions, the final number of speech samples in G1, G2, G3 were 11, 22, and 44, respectively (see Section 3.3.3.3).

Moreover, listeners were assigned a 12-hour window when they completed the Assessment task successfully (i.e., without failing the included trapping questions). During this time, they could not participate in the study. Otherwise, e.g., a worker from G1 (that evaluates ten stimuli), if it would execute the Assessment task two times in a row, would become like a crowd-worker from G2 that assesses 20 speech stimuli. Since our goal is to determine the optimal number of stimuli to include in a single task, it is crucial to avoid such a scenario and differentiate the number of samples presented in each group.

Finally, the Assessment included one last question in which crowd-workers selected on a slider from 1 to 11 how exhausted they felt after completing the listening task, with 1 (translated from German): “not exhausted at all” and 11 being “extremely exhausted”. Selecting the midpoint of the scale, i.e., 6, was not possible.

4.1.2 Results

4.1.2.1 Qualification

A total of 466 workers executed the Qualification phase successfully. Table 4.1 below presents the collected demographics. It can be seen that 94.8% of the participants were from Germany, and 95.9% stated German as their first language. I analyzed the responses of the workers who were not native Germans (4.1%, 19 workers in total). All their answers were correct, so it was assumed that they were fluent in German and therefore included in the study.

Furthermore, 25 workers stated that they were involved in a listening test in the last year and were excluded from our experiment [22]. The remaining 441 crowd-workers were equally demographically distributed in three non-overlapping groups, so a different number of stimuli could be presented to each group.

Table 4.1: Demographic information of the 466 workers that executed the Qualification phase successfully. Values are expressed in percentages. "NP" stands for *"Not Provided"*, i.e., workers did not provide that information.

Language		Country		Gender		Age	
German	95.9	Germany	94.8	Male	56.2	18-25	22.7
English	0.9	Austria	4.5	Female	42.7	26-35	32.8
Russian	0.9	Belgium	0.2	NP	0.9	36-45	21.0
Other	2.3	NP	0.4	Other	0.2	>45	23.4

4.1.2.2 Training and Assessment

The Training phase proved valuable and stopped 15, 8, and 9 workers in G1, G2, and G3, respectively, from taking part in the assessment task. They answered wrong the math question. Overall, 92 crowd-workers in G1 (35.9% female and 93.5% native Germans), 53 in G2 (39.6% female, 96.2% native Germans) and 64 in G3 (42.2% female, 98.4% native Germans) provided a total of 5230, 4840 and 5080 ratings, respectively.

A total of 50 quality scores in G1 and 60 in G3 were labeled as unreliable by our quality control mechanism and were removed. All listeners in G2 answered the trapping questions correctly, and all their ratings were considered for further analysis. I collected at least 24 assessments per stimuli from different listeners in groups G1 and G2. Workers in G3 did not finish all tasks during the study time, so at least 23 assessments per speech sample were gathered.

I calculated the Spearman's rank-order correlation to determine the relationship between the laboratory and the ratings collected in crowdsourcing in G1, G2, and G3 (only the trustworthy assessments). Preliminary analysis showed the relationship to be monotonic, as assessed by visual inspection of a scatter-plot. Additionally, the Root Mean Square Error (RMSE) between the Lab-MOS and the G1-, G2-, and G3-MOS was computed. The correlation with the Lab-MOS was strong and significant for all groups regardless of the number of stimuli included. However, the highest correlation and the lowest RMSE was accomplished in G3 with 40 stimuli ($\rho = 0.89 (p < .001)$), which suggests the highest validity for this group. I believe that this result might be because workers in G3 listened to more speech samples, which allowed them to assess the differences between stimuli better. Table 4.2 presents a summary of these results and the valid ratings collected in each crowdsourcing group.

4.1.2.3 Influence of Number of Stimuli

I conducted an analysis per group to determine if there was a positive or negative trend in the correlation coefficient when considering the listeners' evaluations in intervals of ten stimuli. To this end, the Spearman's correlation was computed between the Lab-MOS and the first ten ratings of a single listener. Then, for the next ten ratings, and like this, four times in G3, two

Table 4.2: Study sizes and Spearman's correlation (ρ) between the Lab-MOS and the CS-MOS (only ratings deemed reliable). The correlation coefficient resulted to be strong and significant in every group.

Group	# of valid ratings	# of listeners	ρ	$RMSE$
G1 (10 stimuli)	5180	92	0.87*	0.45
G2 (20 stimuli)	4840	53	0.86*	0.47
G3 (40 stimuli)	5020	64	0.89*	0.40

*$p < 0.001$

times in G2, and only one in G1. This analysis was conducted with all crowd-workers in all groups. I call this division subgroups (SG).

Results in Figure 4.1 show a slight increase in the correlation coefficient in the second half of the test in G2 (SG1 to SG2). The same tendency can be seen in G3 (i.e., an increase from SG1 to SG2). However, the correlation starts decreasing slightly from SG2 to SG4. This trend suggests that increasing over 40 the number of stimuli per task might lead eventually to unreliable results.

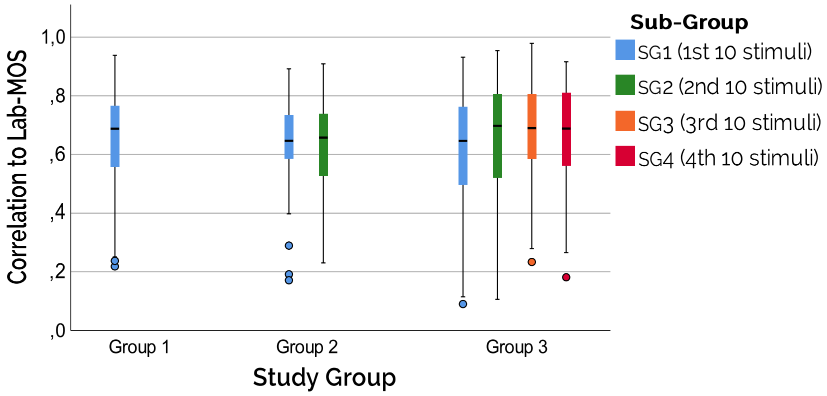


Figure 4.1: Trend in the Spearman's rank-order correlation between the Lab-MOS and the crowdsourcing groups when analyzing in intervals of 10 stimuli.

Figure 4.2 presents the fatigue scores per listening session that were gathered in each group in crowdsourcing. Values are expressed in percentages. It can be seen that in G2 and G3, a higher amount of assessments (14.1% and 13.2%, respectively) were made by listeners reporting being exhausted at the end of the Assessment task (i.e., selecting a value on the fatigue scale above six). In contrast, only 4.7% of workers in G1 were exhausted when finishing each listening session.

The fact that crowd-workers in G3 were fatigued by the end of the listening test could be one reason for a higher proportion of them participating in the study only once compared to those in G1 and G2, see Figure 4.3. Although the number of listeners executing the Assessment decreased linearly in all groups, there were a number of workers completing the listening session 12 times in G1 and eight times in G2. This behavior was not observed within listeners of G3. This outcome indicates that workers preferred to execute short tasks as it was more comfortable for them.

As stated before, the speech material assessed in this study carried 50 degradation conditions. I run a Kruskal-Wallis H test [83] to investigate the differences between the rating scores

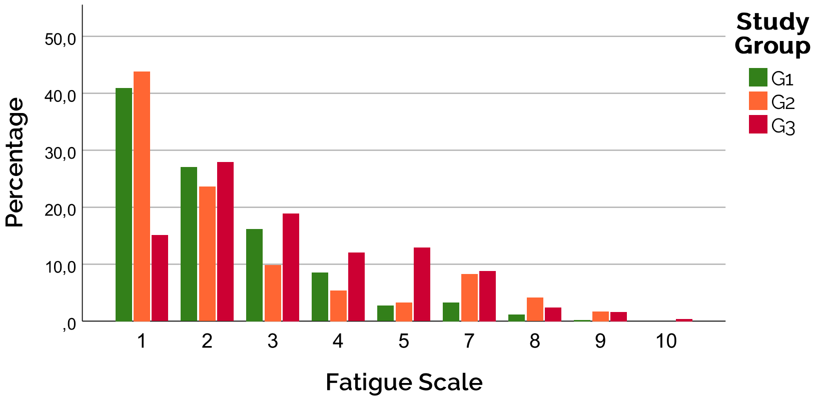


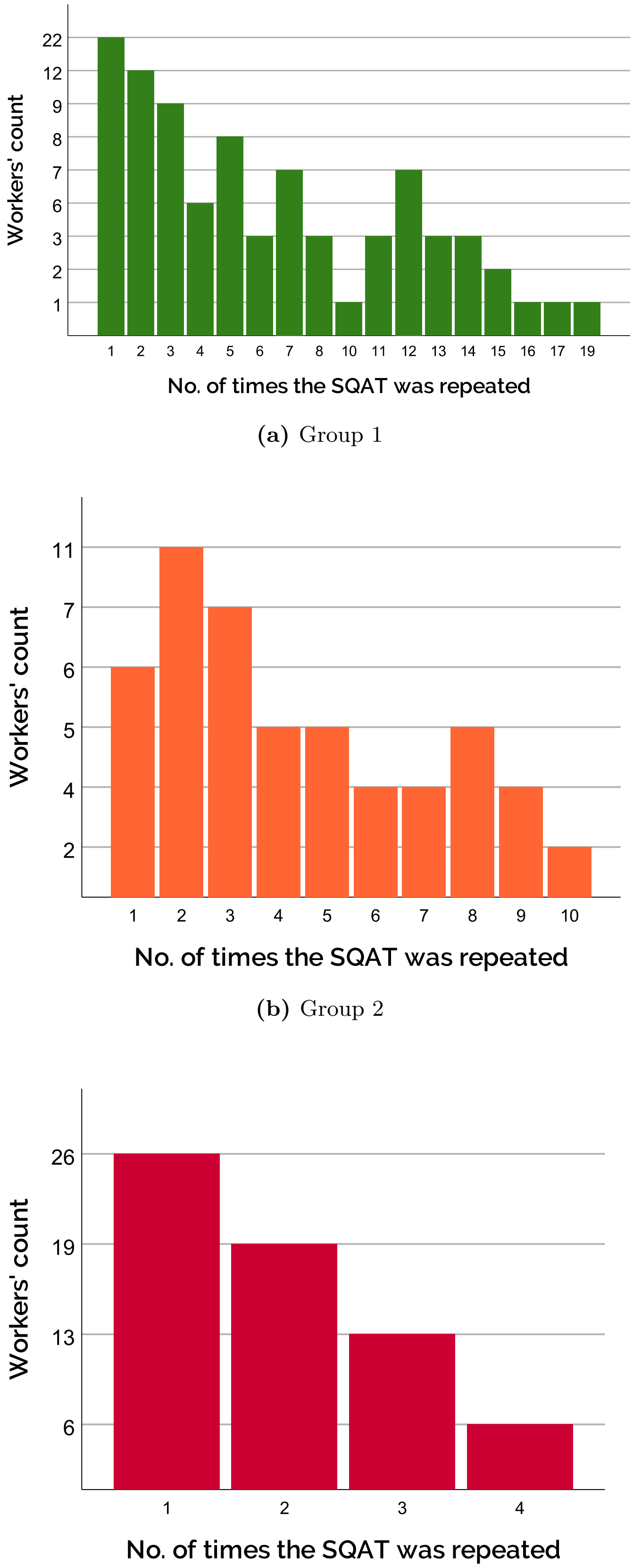
Figure 4.2: Fatigue score per assessments expressed in percentages. In G2 and G3 a greater number of evaluations were made by workers who announced being exhausted at the end of the listening session. The scale was: 1 not exhausted and 11 extremely exhausted. Workers could not select the middle of the scale, i.e., 6.

provided by participants in the laboratory and the ratings from workers of G1, G2, and G3 that evaluated a different number of stimuli per task (i.e., 10, 20, and 40, respectively). Distributions of the quality scores were similar for all groups, as judged by visual inspection of a boxplot. This test revealed that the Median (Mdn) of the rating scores were statistically significantly different in 26 conditions.

Consequently, to determine the cases in which the differences laid between the laboratory and at least two crowdsourcing groups, I conducted a pairwise comparison analysis according to Dunn's procedure [84]. Bonferroni [85] correction for multiple comparisons was made with a statistical significance level accepted at $p < .0083$. This posthoc test exhibited significant differences in 7 and 12 conditions when comparing the laboratory against two and three crowdsourcing groups, respectively. Figure 4.4 presents these 19 conditions with 95% confidence intervals. Information about the degradation conditions can be found in Appendix A. Furthermore, Table 4.3 shows the results (χ^2 values and pairwise comparisons) for the cases in which the laboratory differed from all the crowdsourcing groups.

Additionally, I analyzed the results per group of the previous test. I found that the ratings given to 17 of the conditions were statistically significantly different between the laboratory and workers in G1. In turn, for G2 and G3, there were significant differences to the laboratory for 22 and 18 conditions, respectively. These results suggest that the quality scores were more accurate in G1 and again less reliable in G2. I hypothesize that workers in G1 evaluating a lower number of stimuli were only partially exhausted, so they could better judge the overall speech quality based on the different degradations.

A closer look into these results revealed that particularly wideband (WB) (e.g., conditions 13, 20, 26, 27, 43, 45, and 46) and super-wideband (SWB) (e.g., conditions 5, 11, 12, 22, and 34) speech stimuli were overrated in crowdsourcing in comparison to the laboratory, see Figure 4.4. In contrast, narrowband (NB) speech files seem to provoke a lower quality score in crowdsourcing (e.g., conditions 29, 31, 36, and 42). The labels and details of the degradation conditions are presented in Appendix A. These results are in line with [24], where NB conditions were rated significantly lower in crowdsourcing than in the laboratory.



(a) Group 1

(b) Group 2

(c) Group 3

Figure 4.3: The figure shows the number of workers that executed the speech quality assessment task a certain number of times. For instance, 22 workers in G1 conducted the Assessment phase one time. “SQAT” stands for “Speech Quality Assessment Task”.

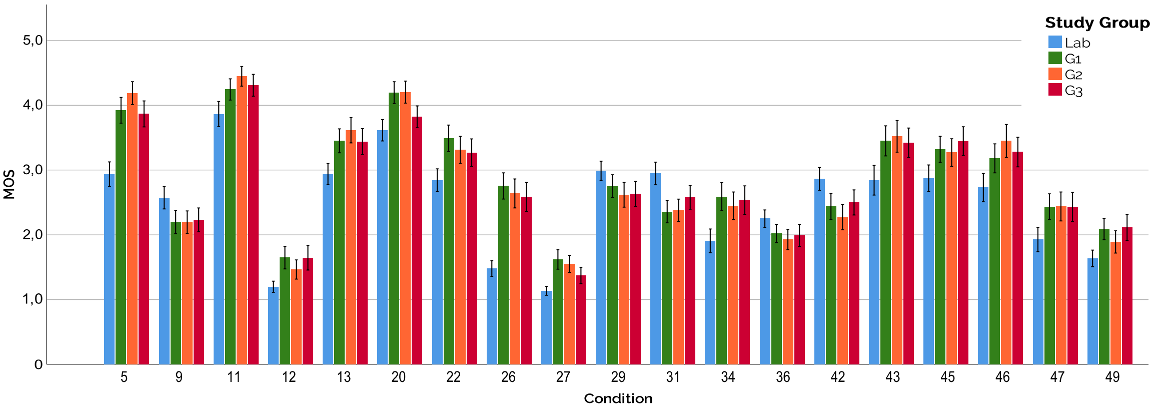


Figure 4.4: Comparison between the Lab-MOS and the CS-MOS with 95% confidence intervals. Represented only the conditions for which there is a statistically significant difference between laboratory and at least two crowdsourcing groups. Information about the degradation conditions can be found in Appendix A and more details in [9].

Table 4.3: The table presents the 12 conditions for which the median (Mdn) of the rating scores were statistically significantly different between the laboratory and all the groups in crowdsourcing.

Cond. No.	$\chi^2(3)$	Lab	G1		G2		G3	
		Mdn	Mdn	p-value	Mdn	p-value	Mdn	p-value
5	75.6*	3	4	< .001	4	< .001	4	< .001
9	12.1⋆	3	2	= .022	2	= .019	2	= .046
11	26.6*	4	4	= .001	5	< .001	5	= .001
13	26.6*	3	3	= .001	4	< .001	3.5	= .001
22	22.7*	3	4	< .001	3	= .005	3	= .013
26	94.4*	1	3	< .001	3	< .001	2	< .001
31	26.4*	3	2	< .001	2	< .001	3	= .019
34	25.0*	2	2	< .001	2	= .002	2	< .001
42	20.8*	3	2	= .007	2	< .001	2	= .042
43	20.2*	3	4	= .002	4	< .001	4	= .003
46	20.4*	3	3	= .037	4	< .001	3	= .007
47	16.6●	2	2	= .002	2	= .007	2	= .008

*$p < .001$; ⋆$p = .007$; ●$p = .001$

4.1.3 Discussion

This section investigates the influence of the number of stimuli on the validity of the speech quality ratings collected via crowdsourcing. A study was carried out with 209 crowd-workers divided into three non-overlapping groups. Each group was presented with tasks consisting of a different number of stimuli, i.e., 10, 20, or 40. The results in the three groups were highly correlated to previously collected laboratory ratings. The best performance in terms of correlation to the laboratory ratings was achieved in G3 when employing 40 stimuli. However, a significant number of workers in G3 reported being exhausted when finishing the assessment task. Thus, most listeners in this group participated in the study only one time.

Additionally, an analysis was made to determine whether listeners in crowdsourcing perceived significantly different speech impairments compared to listeners in the laboratory.

This test revealed that G1 with ten stimuli had a lower number of degradation conditions for which the quality ratings were significantly different from those provided in the laboratory. Therefore, it is desirable to offer tasks with a reduced number of speech stimuli at the expense of sacrificing ratings' accuracy to some extent.

Furthermore, I found that workers in crowdsourcing tended to rate higher the quality of wideband (WB) and super-wideband (SWB) speech stimuli than participants in the laboratory experiment. In contrast, narrowband (NB) speech samples seem to provoke lower quality ratings in crowdsourcing than in the laboratory. Further investigation would be needed to determine the reasons for the difference in these quality evaluations for certain WB, SWB, and NB speech files.

4.2 Impact of Task Repetition

Crowdsourcing is a convenient instrument for carrying subjective speech quality studies. Nonetheless, data gathered in crowdsourcing can be corrupt due to users' neglect. Therefore, participants who are consistent in their answers or exhibit a high intra-rater reliability score are preferred for speech quality assessments in crowdsourcing.

This section investigates the impact of executing a speech quality assessment task multiple times on the intra- and inter-listener agreement. Additionally, I determine the relationship between the intra-rater reliability and listener performance. Two studies were conducted, one in the laboratory and the other one in crowdsourcing. Listeners in both experiments rated four times the quality of the speech stimuli. Finally, I propose a model as a function of intra-rater reliability, root-mean-square, and listeners' age to predict worker performance. Such a model is intended to measure how valid the crowdsourcing results are when there are no laboratory results to compare with.

4.2.1 Study Setup

In the following, I present the speech material employed in our studies and the experiment carried out in the laboratory, which aimed at evaluating the listener perception of modern mobile telephony services. Finally, the crowdsourcing study is detailed, intended to replicate the results gathered in the laboratory test.

4.2.1.1 Speech Material

The constructed speech database is a mixed fullband set of samples containing different audio bandwidths, i.e., from below narrowband and up to fullband. The bandwidths distribution between the speech stimuli is the following: 21% narrowband (NB), 50% wideband (WB), 23% super-wideband (SWB), and 6% fullband (FB).

The listening test in the laboratory targeted the assessment of speech stimuli encoded with state-of-the-art codecs, e.g., EVS [86], AMR-WB, AMR-NB [87, 88], and Opus [1] under ideal and live good/average/bad coverage situations. The test focused on live conditions from real-field recordings, which were 71% of the speech samples. The remaining 29% were based on offline processed speech and anchor conditions.

[1] https://opus-codec.org/

The real-field recordings were collected in Switzerland during September 2018, under good, average, and bad coverage network conditions. These samples were gathered using modern equipment from Rohde & Schwarz SwissQual, and it included state-of-the-art measurements such as:

- VoLTE calls with EVS at 24.4 kbit/s SWB
- WhatsApp calls in LTE with Opus at 20 kbit/s WB
- 3G mobile to mobile calls with AMR-WB at 12.65kbit/s and 23.85 kbit/s
- 3G mobile to mobile calls with AMR-NB at 12.2 kbit/s
- 3G/2G mobile to mobile calls with transcoding from AMR-WB at 12.65 kbit/s to AMR-NB at 12.2 kbit/s

The offline (simulated) coded conditions in the test aimed at emulating everyday situations that can be seen in the field. Consequently, we selected:

- EVS 24.4 kbit/s SWB
- EVS 13.2 kbit/s SWB
- Opus 20 kbit/s WB
- AMR-WB 23.85 kbit/s
- AMR-WB 12.65 kbit/s
- AMR (NB) 12.2 kbit/s.

The three fullband conditions in the test were one fullband reference and two anchors with packet loss. Four additional low quality simulated conditions were obtained by adding packet loss to some codec conditions or by re-encoding the reference sample multiple times with the same settings. The sample used was the composed female/male German sample from the ITU-T Rec. P.501 Annex D [77]. Overall, 53 speech stimuli were arranged, carrying 53 degradation conditions. The full list of these conditions is listed in Table C.1 of Appendix C. More details about the speech material can be found in [78].

As stated above, the test included a high number of live recordings in real-field mobile networks. For these recordings, it does not apply the use of several speech samples forming a “condition”. Thus, we used only one speech stimulus that was presented four times in total (non-consecutive) to the listeners. This procedure resulted in a comparable significant statistical confidence, as achieved with traditionally designed P.800 listening tests.

4.2.1.2 Laboratory Study

The laboratory study intended to evaluate the user perceptions of popular mobile telephony services such as VoLTE, circuit-switched mobile (i.e., GSM and UMTS), and VoIP OTT applications like WhatsApp.

The experiment was conducted at the SwissQual test laboratory in October 2018. The listening panel consisted of 24 native German listeners (11 female and 13 male). They were invited individually to run the test, and only one person was listening at a time. Subjects assessed the quality of speech samples using a discrete five-point ACR scale with the possibilities: “ausgezeichnet” (excellent), “gut” (good), “ordentlich” (fair), “dürftig” (poor), and “schlecht” (bad).

All participants evaluated all 53 conditions, which in this study were 53 speech stimuli. To accomplish small enough confidence intervals, listeners assessed four times the quality of the 53 speech samples. Then, we collected 96 ratings per stimulus. The stimuli presentation order was randomized for each participant. The test started with five training speech sequences to let the listeners get used to the interface and the test setup. The speech samples were presented in a diotic (binaural) form to the subjects through a diffuse field equalized headphones (Grado SR 60). The presentation level was 73 dB(A) SPL at each ear (equivalent to -26dB OVL). More details about the laboratory study can be found in [78].

I collected subjective quality assessments from 24 different native German listeners. Then, Mean Opinion Scores (MOS) were computed for each stimulus. I refer to these scores as “Lab-MOS” and use it as a reference for the analysis presented in this subsection.

Additionally, Kendall’s coefficient of concordance (W) [89] was calculated to determine the agreement among the listeners’ ratings. This test revealed a statistically significant agreement across the laboratory participants when they evaluated all the speech samples, $W = 0.86, p < 0.001$. The data’s reliability is verified when a high agreement exists in the ratings that different participants provide to the same speech stimulus. This high Kendall’s coefficient exposes a low variability across the individual ratings, demonstrating high confidence in the collected MOS.

It is valid to mention that we cannot prove the influence of the training effect directly (i.e., listeners assessing the samples four times). However, when confronting the predictions of the ITU-T Rec. P.863 [9] model with the collected subjective scores shows no anomaly regarding common, classically designed P.800 [66] studies with varying or non-repeating speech samples. Neither the spread of the prediction rank-order nor the bias nor codec or bandwidth dependencies were larger than for classically designed experiments [78]. For this indirect proof, it should be noted that P.863 is trained on classically designed databases but predicts successfully in our case of an experiment with repetition of the same sample across all conditions.

4.2.1.3 Crowdsourcing Study

The crowdsourcing experiment was executed in the clickworker crowdsourcing platform. The study consisted of three phases, i.e., Qualification, Training, and Assessment. The details of these phases are presented in Section 3.3.3.1, 3.3.3.2, and 3.3.3.3, respectively.

Unlike the details defined in Section 3.3.3.3, the Assessment task in this study comprised 58 speech stimuli, i.e., 53 speech samples plus five trapping questions inserted every ten stimuli. The Assessment also included a slider at the end so crowd-workers could state from 1 to 11 how exhausted they were after completing the listening task. One meant “not exhausted at all” and 11 “extremely exhausted”. Workers could take part in the Assessment task up to four times.

4.2.2 Results

A total of 8321 ratings were collected and provided by 52 crowd-workers. 51.9% were female, 96.2% were from Germany (2 workers were from Austria), and all were native German speakers.

The trapping question was useful and permitted us to identify 119 unreliable ratings given by three participants. Two of them failed all of the trapping questions the single time they conducted the Assessment task. Then, 106 ratings were invalidated. The other worker failed the last trapping question in the listening test; then, those 13 ratings were also invalidated. These 119 ratings were discarded as I assumed those crowd-workers were performing the listening test carelessly. The resulting 8202 ratings are then considered for further analysis.

Out of all participants, 29 crowd-workers executed the Assessment task four times, like in the laboratory, and provided 6148 quality scores. These subjective ratings were averaged per degradation condition to compute the MOS scores, referred to as "CS-MOS". I calculated the Pearson's product-moment correlation between the laboratory ratings and the ratings given by these 29 workers. Preliminary analyses showed the relationship to be linear with both variables approximately normally distributed, as assessed by visual inspection of Normal Q-Q Plots, and there were no outliers. Additionally, the Root Mean Square Error (RMSE) between the Lab-MOS and the CS-MOS was computed. The correlation coefficient was strong and statistically significant, and the RMSE low, i.e., $r = 0.978$ ($p < .001$), $RMSE = 0.441$.

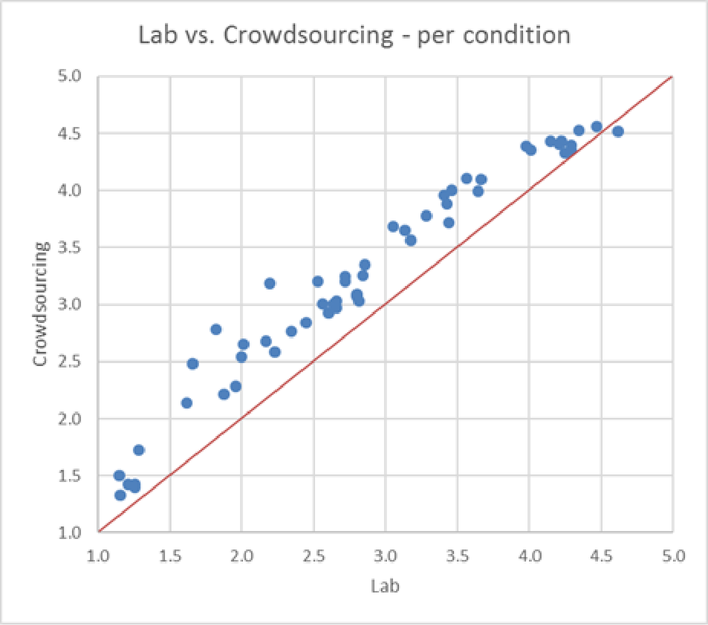
Moreover, a scatterplot between the Lab-MOS and the CS-MOS exposed a slight "banana shaped" effect between the two sets of results. Such an effect is usually due to the differences in the test conditions between the laboratory and crowdsourcing environment, e.g., equipment, listening panel, test presentation. This effect was corrected by applying the first and third order mapping [90]. While the correlation coefficient did not significantly increase, this analysis helped correct the bias and improved the RMSE to 0.17. Table 4.4 presents these results, and Figure 4.5 shows the scatterplots between the Lab-MOS and the CS-MOS before and after applying the third order mapping.

Table 4.4: The table shows the Pearson correlation (r) and RMSE between Lab-MOS and CS-MOS after applying the first and third order mapping [90].

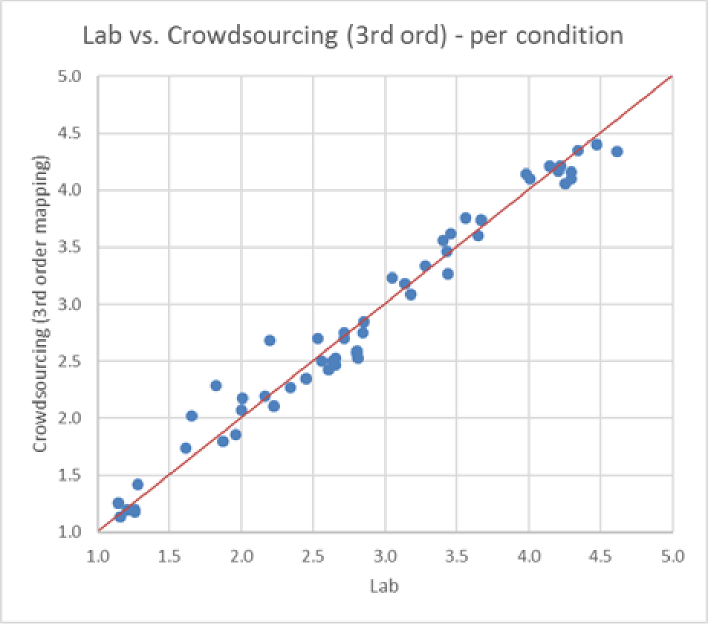
	r	$RMSE$
no mapping	0.978*	0.4409
1^{st} order mapping	0.979*	0.2025
3^{rd} order mapping	0.986*	0.1701

*$p < 0.001$

Furthermore, I computed the Kendall's concordance coefficient (W) [89] to determine the agreement among the workers' ratings. I found that the 29 crowd-workers statistically significantly agreed in their assessments, $W = 0.807, p < 0.001$. Our results reveal that, despite the task's subjectivity, listeners in crowdsourcing seemed to apply the same criteria when assessing the overall quality of these speech stimuli encoded with state-of-the-art codecs. This is probably due to our mature methodology. Such a high coefficient indicates that most of the variance across the ratings can be explained by the differences between the speech samples and not by individual differences in the workers' evaluations. All in all, our results suggest that the crowdsourcing ratings might be as reliable as the ones gathered in the laboratory



(a) no mapping



(b) 3^{rd} order mapping

Figure 4.5: The figures show the Lab-MOS versus CS-MOS per condition before and after applying the third order mapping [90].

and that most probably, the differences would be the same if repeating the experiment in a different laboratory. In fact, the agreement between listeners in the laboratory, i.e., $W = 0.86$ (see Subsection 4.2.1.2), is comparable to workers' agreement in crowdsourcing.

4.2.2.1 Inter-rater Reliability

In the following, I investigate the listeners' agreement in each of the four stages in which they conducted the crowdsourcing Assessment task. Our goal is to determine if the agreement fluctuated from the first to the fourth time workers executed the Assessment and if it varied with respect to the Pearson correlation and the RMSE. This analysis was only performed for listeners in crowdsourcing. Participants in the laboratory assessed four times the quality of the 53 speech stimuli at once, which resulted in a randomized set of 212 stimuli. Thus, no distinction can be made in terms of repetitions.

The Kendall's concordance coefficient (W) [89] can be used as such a measure of inter-rater agreement for continuous and ordinal variables when there are two or more raters [91, 92]. I considered the single rating given by the 29 workers to the 53 speech stimuli at each repetition. Consequently, four Kendall's W coefficients were calculated on this ordinal data.

Additionally, I calculated the correlation and RMSE between the Lab-MOS and the ratings provided by the 29 crowd-workers averaged per file, at repetition one, two, three, and four. Table 4.5 outlines these results. While the RMSE and the correlation prevailed almost constant, the agreement between workers increased slightly (not monotonous) from repetition one to four (i.e., 0.67 to 0.68). Such an increase can be explained by listeners becoming more confident with their assessments, the more they participate in the study. Nevertheless, the lowest agreement ($W = 0.65$) was seen during the third repetition. To investigate further, I conducted a two-way mixed ANOVA [93, 94]. The analysis showed that indeed the main effect of repetition presented a statistically significant difference in the mean ratings at the third repetition for three of the speech degradation conditions, i.e., $C07$, $C15$, and $C37$. The results are presented in Table 4.6.

Table 4.5: The table presents the Kendall's (W) coefficient of agreement, the Pearson's product-moment correlation (r), and RMSE between the Lab-MOS and the CS-MOS for each of the times the workers conducted the Assessment task.

Repetition	W	r	$RMSE$
1	0.6712*	0.9758*	0.4452
2	0.6673*	0.9745*	0.4413
3	0.6582*	0.9770*	0.4460
4	0.6836*	0.9777*	0.4409

*$p < 0.001$

Table 4.6: Results of the two-way mixed ANOVA showing the three conditions for which a statistically significant difference was seen in the mean ratings at the third repetition. More details about these speech impairments can be found in Table C.1 of Appendix C and in [78].

Condition	Condition Description	$F(3, 53)$	$p - value$	η^2
$C07$	EVS 13.2 kbps SWB	5.23	= 0.002	0.093
$C15$	2 x AMR-WB 6.6kbps	2.891	= 0.037	0.054
$C37$	M2M UMTS call AMR-WB 23.85kbps avg. network condition 3	7.287	< 0.001	0.125

The slight decrease in the between listener agreement helped identify those three speech impairments that were the most difficult for the listeners to assess. I believe that, in order to gather reliable speech quality annotations for these degradation conditions, the assessment task should be addressed to a large pool of listeners. This way, the quality ratings' confidence intervals could be lowered down.

All in all, the high Kendall's coefficient indicates that the speech quality assessment study in crowdsourcing was well designed and with low ambiguity. It can be assumed that all workers understood the instructions and that most of the variance across the ratings can be explained by the differences between the speech degradation conditions and not by different test interpretations. This outcome confirms the reliability of the collected ratings.

4.2.2.2 Intra-rater Reliability

Additionally, I examined which of the four times the workers performed the Assessment were most confident in their evaluations. With this aim in mind, I calculated the intra-rater reliability (IRR). The IRR provides a measure of the consistency in the ratings that a single worker gives to the same sample at different time points. Then, the IRR was estimated for each worker considering the first two, three, and four repetitions by calculating the Intraclass Correlation Coefficient (ICC) over the ratings given at each Assessment task [95].

As noted earlier, the laboratory listeners evaluated the quality of the 53 speech stimuli four times at a time, resulting in a random set of 212 speech samples. Consequently, it was not possible to make a distinction on repetitions. Therefore, this analysis is only conducted for listeners in crowdsourcing.

The ICC is a statistical method frequently used for assessing IRR for ratio, interval, and ordinal variables. It is particularly suitable when the "cases" under investigation are evaluated two or more times [96] like in our study. I used a "two-way random" model to compute an $ICC(2,1)$ coefficient as I was interested in the degree of agreement in the absolute values across the ratings from a single worker [97]. For this purpose, the *"icc"* function of the R package "IRR" was employed using "agreement" and "single" as parameters [96]. The boxplots in Figure 4.6 reveal these results. The graph shows that the point in time workers executed the Assessment did not influence the IRR. This outcome suggests that all crowd-workers executed each Assessment task with high conscientiousness, which confirms the reliability of the collected ratings.

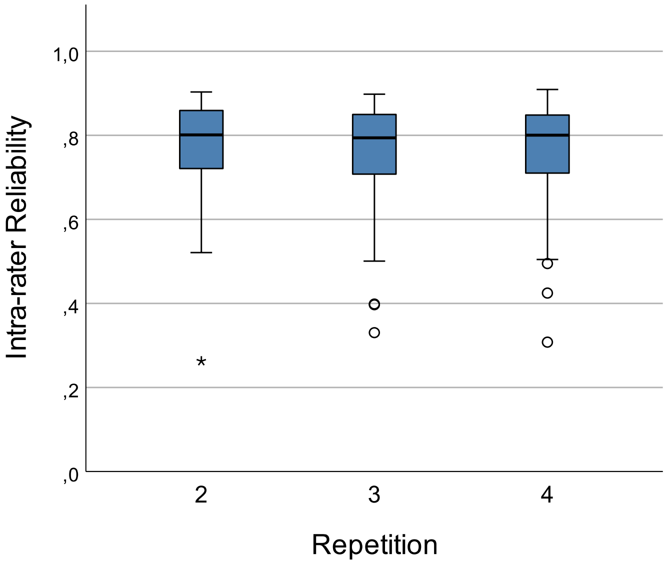
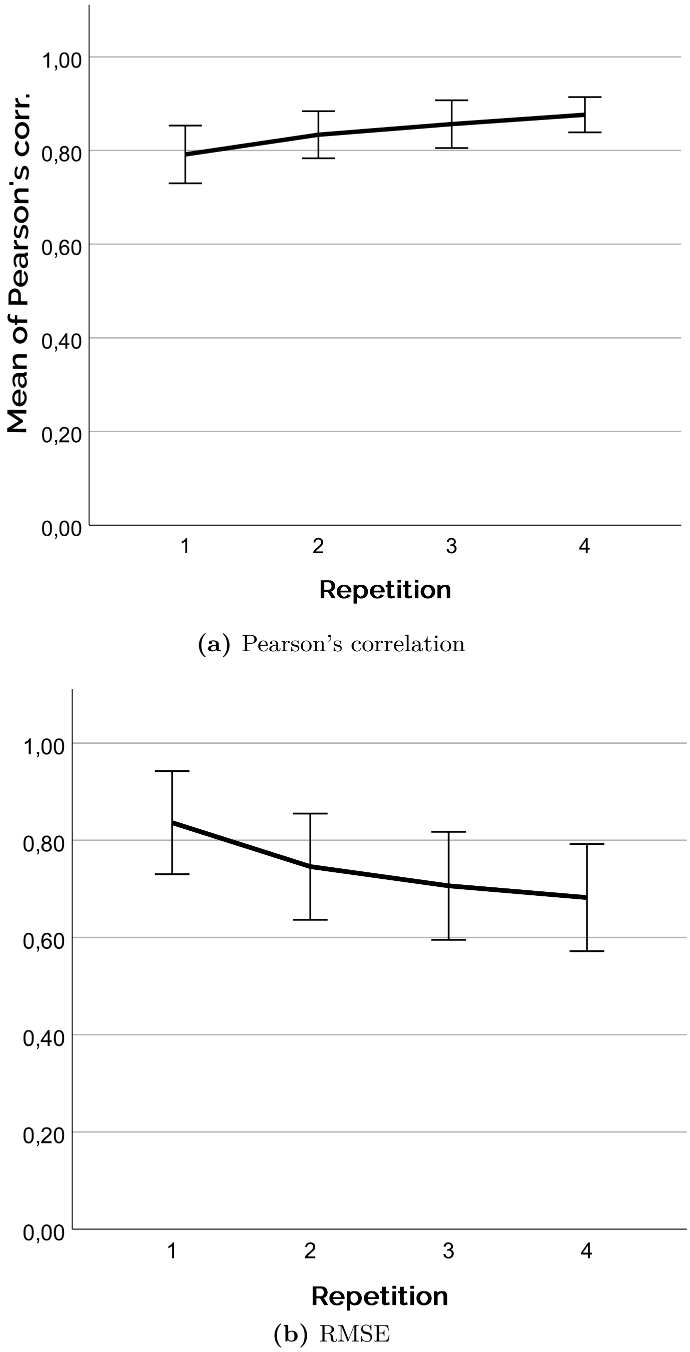


Figure 4.6: Reliability of the workers at the second, third and fourth time they conducted the Assessment task.

Furthermore, I investigated how the Pearson correlation (r) and the RMSE changed through the whole crowdsourcing study, i.e., from repetition one to four. To this end, I calculated the correlation and RMSE between the Lab-MOS and the scores given by a single crowd-worker at each Assessment task. This analysis was made per worker and repetition. Figure 4.7 shows these results with 95% confidence intervals. It can be seen that both the correlation and the RMSE improved with the number of repetitions, i.e., r increased from repetition one to four, whereas the RMSE decreased. This outcome evidences that conducting the Assessment task multiple times contributed to collecting more accurate speech quality scores. Hence, better results were achieved after the fourth repetition (when comparing to the laboratory results).



(a) Pearson's correlation

(b) RMSE

Figure 4.7: Pearson's correlation and root-mean-squared-error (RMSE) with 95% confidence intervals between the Lab-MOS and the ratings provided by each worker.

4.2.2.3 Predicting Workers' Performance

A standard multiple regression analysis to predict the workers' performance was executed. Features were derived from intra-rater reliability (IRR), root-mean-squared-deviation (RMSD), listeners' age and gender, workers fatigue, and time they spent in each Assessment task. I considered the ratings of 29 and 7 crowd-workers who ran the Assessment four and three times, respectively. The rest of the workers participated in the study only once or twice, so their ratings were not considered.

The workers' performance (WPer) I expressed as a percentage and defined it as the Pearson's correlation between the Lab-MOS and the ratings provided by a single worker. As stated before, the IRR determines to what extent a listener's ratings gathered at different time-points are consistent [98]. The IRR was calculated in the previous Subsection 4.2.2.2, and it is used now for this analysis as well.

The RMSD was determined by computing the root-mean-squared-deviation between a workers' MOS and the rest of the workers' ratings. The Fatigue (FTG) data was gathered on a scale varying from 1 (not exhausted) to 11 (extremely exhausted). The collected FTG scores and the time (T) were averaged over the three or four times each listener conducted the Assessment task. Moreover, workers' age was collected as an age range, i.e., 18-25, 26-35, 36-45, and above 45; thus, I treated it as a categorical variable. All of the input features were normalized based on z-scores to avoid the bias of the regression coefficients and to evaluate better their impact on the prediction.

The first multiple regression analysis revealed that time, fatigue, and gender, were not contributing significantly to predict workers' performance, i.e., $p = 0.121$, $p = 0.879$ and $p = 0.108$, respectively. Then, a stepwise multiple regression was conducted with the remaining input variables and 5-fold cross-validation. This test unveiled linearity as assessed by partial regression plots and a plot of studentized residuals against the predicted values. Residuals were independent [99], as determined by a Durbin-Watson statistic of 1.808. There was homoscedasticity, as assessed by visual inspection of a plot of studentized residuals versus unstandardized predicted values. There was no multicollinearity evidence, as judged by tolerance values greater than 0.1 [100], and the correlation between the independent variables was below 0.7. There were no studentized deleted residuals greater than ±3 standard deviations, no leverage values greater than 0.498, and values for Cook's distance above 1. The assumption of normality was met, as assessed by a Q-Q Plot. The multiple regression model statistically significantly predicted WPer, $F(3, 32) = 44.228, p < .001$, adj. $R^2 = 0.787$. All three variables added statistically significantly to the prediction, $p < 0.05$ for IRR and RMSD, and $p = 0.016$ for age. Regression coefficients and standard errors can be found in Table 4.7. Equation 4.1 presents the yielded multiple regression model.

$$\mathbf{WPer} = 87.925 + 6.36 \cdot \mathbf{IRR} - 3.72 \cdot \mathbf{RMSD} - 1.847 \cdot \mathbf{Age} \tag{4.1}$$

The model accomplished a Pearson correlation of $r = 0.898$ ($p < .001$) and RMSE of 3.9 in predicting the workers' performance. The regression coefficients show that intra-rater reliability has the largest influence on the prediction, while age has the lowest impact. This

Table 4.7: Summary of the multiple regression analysis.

Variable	B	SE_B	β	p
Intercept	87.925	0.689		< 0.05
IRR	6.360	0.742	0.709	< 0.05
RMSD	-3.720	0.727	-0.415	< 0.05
Age	-1.847	0.723	-0.206	$= 0.016$

B (unstandardized) and β (standardized) regression coefficients; SE_B= Standard error of the coefficient;

outcome suggests that it is possible to predict and evaluate the validity of the crowdsourcing results based on how consistent are the ratings of a single user at different points in time.

4.2.3 Discussion

In the study described previously, listeners could evaluate four times the speech samples in the dataset. I analyzed the inter- and intra-listener agreement, as I believed that these two metrics would increase gradually from repetition one to four. It was also believed that the crowdsourcing ratings' accuracy, in terms of correlation and root-mean-squared-deviation to the laboratory results, would increase as well. However, this was not the case with the inter- and intra-rater agreement. While the former fluctuated from the first to the last repetition (see Table 4.5), the latter remained almost constant, as shown in Figure 4.6. This outcome shows that there was no linear relationship between these two metrics, and the fact that listeners, as individuals, were quite consistent with their answers did not contribute to a significant increase in the listeners' agreement as a group.

Interestingly, an opposite effect was seen with the Pearson's product-moment correlation and the RMSE. Both remained almost constant from repetition one to four (see Table 4.5) when contrasting all the laboratory and the crowdsourcing ratings. In contrast, these two metrics improved when considering the ratings of individual workers at each of the Assessment tasks. See Figure 4.7(a) and 4.7(b), respectively.

Furthermore, I developed a model to predict workers' performance based on intra-rater reliability (IRR), Root Mean Square Deviation (RMSD), and listeners' age. The IRR coefficient had the largest impact on the prediction, while age had the lowest influence. Additionally, I demonstrated the importance of the consistency in the workers' ratings to the final performance, and eventually to the overall results' validity.

It is valid to point out that the study setup of the experiments carried out in this subsection is relatively uncommon. Usually, due to financial and time constraints, participants of a speech quality assessment study evaluate only once the samples in the dataset. Frequently, these speech databases contain multiple samples coded with the same degradation condition. Thus, an intra-rater reliability score could be calculated, accounting for the ratings given to a particular degradation condition. With this premise in mind, I hypothesize that the proposed model could be used to evaluate the listeners' performance, considering the scores per condition of a single participant.

4.3 Conclusion

This chapter investigated the optimum number of speech stimuli to include in a speech quality assessment task in crowdsourcing. Additionally, I analyzed important reliability metrics within a speech quality experiment. To this end, two sets of separate studies were conducted.

In the first study, workers were divided into three groups, each of which was confronted with tasks consisting of a different number of speech samples, i.e., 10, 20, or 40. The Mean Opinion Scores in all groups were significantly correlated to ratings collected in a previous laboratory study despite the number of speech stimuli per task. The highest correlation coefficient was achieved in the group assessing 40 stimuli. However, I found that workers in this group were exhausted at the end of the assessment task, and therefore they participated in the study only one time. I encourage offering tasks with a reduced number of stimuli to promote worker participation and reduce study response times.

The second study was first conducted in the laboratory and then replicated in crowdsourcing. Listeners had the chance to evaluate four times the quality of speech samples processed with state-of-the-art codecs under different transmissions types. I then analyzed the between and within listeners' agreement and prosed a model to predict the workers' performance. Features were derived from the gathered data, and the intra-rater reliability was the feature with the most substantial influence on the prediction. Such a model would help determine workers' accuracy in situations where there is no baseline data to compare the crowdsourcing results.

5

Impact of Background Noise

Speech quality assessment studies in crowdsourcing benefit from reduced turnaround times at lower costs. They also induce real-life environmental conditions, as crowd-workers frequently work from home employing their computers and headphones. Still, there is a lack of control over the participants and not enough information about their playback system and background environment. The validity of data gathered in a disturbed environment is questionable, especially in speech quality assessment studies.

This chapter investigates the influence of the environmental background noise in speech quality studies carried out in crowdsourcing. A three-phase speech quality assessment study was conducted in a simulated crowdsourcing environment in the laboratory. In each phase, listeners assessed the quality of speech files under the influence of environmental background noise at different levels.

Furthermore, the feasibility of using web-audio recordings for environmental noise classification is examined. Two background noise datasets were arranged, and standard features were extracted and used for noise classification and noise level estimation.

5.1 Effect of Environmental Background Noise

The methodology for conducting speech quality studies in crowdsourcing has matured. Different mechanisms have been proposed to ensure valid results, e.g., “trapping questions”, “temporal training”, and others [66]. However, the question remains about the influence of the environmental background noise on speech quality ratings. Workers do not always follow the given instructions, and they might be exposed to different environmental conditions while conducting specific tasks. For instance, I surveyed crowd-workers’ environmental conditions and found that they could hear street noises and the TV while answering the given survey. Therefore, it is crucial to determine the influence of common crowdsourcing environmental background noise conditions on the speech quality assessments.

5.1.1 Study Setup

A simulated crowdsourcing study in the laboratory was conducted, as detailed in Section 3.4. Three groups of participants were recruited to evaluate the quality of speech stimuli under the influence of different environmental background noise conditions. Group one (G1) and three (G3) carried out the listening test in the presence of street traffic background noise, and Group two (G2) in the presence of “TV-Show” background noise. I decided to use these noises as it was found in Chapter 3, Section 3.3.4 that these two are the most common environmental noises that German crowd-workers may be exposed to when performing crowdsourcing tasks.

The speech quality assessment tests were conducted following the ITU-T Recommendation P.808 [25], which was already published by the time this study was conducted. P.808 is the result of years of research of our Lab and other laboratory partners into the use of crowdsourcing for speech quality assessment studies. Therefore, our methodology, defined in Subsection 3.3.3 for executing speech quality evaluations in crowdsourcing, is in line with the ITU-T Recommendation P.808.

The study was divided into four sessions. First, participants executed a standard P.808 [25] listening test without background noise. Afterward, the remaining three sessions were also according to Recommendation P.808, but under the influence of environmental background noise at different levels. The order of these last three sessions was randomized, and the speech stimuli were the same in each of the four sessions. Table 5.1 below compiles this information and displays the levels at which the background noise was played during each test session.

Table 5.1: Study setup and levels at which the background noise was played throughout each test session.

Session	Group 1 (G1) (street traffic noise)	Group 2 (G2) (TV-Show noise)	Group 3 (G3) (street traffic noise)	Ordering
CSLvl0	-	-	-	first
CSLvl1	50dB(A)	60dB(A)	36dB(A)	random
CSLvl2	55dB(A)	70dB(A)	43dB(A)	
CSLvl3	65.5dB(A)	80dB(A)	50dB(A)	

Sessions’ name nomenclature: “CS” stands for crowdsourcing, “Lvl” for level and 0 to 3 to indicate the noise level.

The background noise levels were measured with the dummy head HMS II.3 HEAD acoustics. For the presentation of the noise, I employed a four-speaker setup as defined in [80] for simulating the noisy environments, and a “FIREFACE UCX” served as an audio interface. The speakers used were the “Studio Monitor Genelec 8030A” that contain 5” woofers, 54Hz to 20KHz frequency range, and 110dB SPL.

5.1.1.1 Background Noise Signals

The street traffic noise signal was taken from the background noise database available at [80]. Concretely, the 20 seconds long “Outside_Traffic_Crossroads_binaural.wav” audio file was used. On the other hand, for the “TV-Show” background noise group (G2), it was required an audio signal containing speech, instrumental music, and a combination of both. Hence, I crafted a three-minute file using audio clips from the German television program: "Die Harald Schmidt Show".

These noise signals were attenuated or amplified to accomplish the desired noise level. A frequency analysis of the street noise signal revealed that most of its energy was concentrated at the low frequencies between 10Hz and 1000Hz. In turn, the "TV-Show" noise had the most energy between 10Hz and 4000Hz. I used these two frequency ranges as parameters for measuring the different levels of noise with the fullband artificial head HMS II.3. This binaural measurement was conducted in dB SPL (Sound Pressure Level) and A-weighted throughout each noise signal with time weighting response time of one second.

5.1.1.2 Speech Database

I used the speech stimuli from the database SwissQual 502 from the ITU-T Recommendation P.863 [9] competition. This database contains four speech samples per condition and a total of 50 different speech impairment conditions. The stimuli were 8 to 12 seconds long, and I used only 60 speech stimuli to keep the overall study duration under one hour while avoiding listener fatigue. More details about SwissQual 502 can be found in Section 3.2.2.

Moreover, this database also includes speech quality ratings provided by 24 different native German speakers. The MOS per file and condition is used as a reference for the analysis presented in this chapter (referred to as "Lab-MOS"). Table 5.2 provides information about the 15 speech degradation conditions employed in our study. More details can be found in [9].

Table 5.2: Conditions labels referring to the speech impairments under test [9].

Condition Number	Degradation Description
1	SWB
2	SWB+Noise 12dB
3	SWB+Noise 20dB
6	SWB Level -10dB
7	SWB Level -20dB
32	EVRC-A
33	EVRC-A + Noise 18dB SNR + Codec NS
43	VoIP WB-Call + acoust. send
44	VoIP WB-Call + -16dB + acoust. send
45	VoIP WB-Call + -8dB + acoust. send
46	VoIP WB-Call + +5dB + acoust. send
47	VoIP WB-Call + acoust. noise (rcv)
48	VoIP WB-Call + acoust. noise (rcv) + -8dB
49	VoIP WB-Call + acoust. noise (rcv) + -16dB
50	VoIP WB-Call + ampl. clipping + acoust. send

5.1.2 Results

Overall, 24 listeners in G1, 25 in G2, and 20 in G3 participated in our study and produced 5760, 6000, and 4800 ratings, respectively. All participants in the three groups came from Germany and were native German speakers. Table 5.3 shows the demographics of the participants. They all were compensated for their participation.

Table 5.3: The table presents the demographic information of the 24, 25, and 20 participants that executed the study in G1, G2, and G3, respectively. All listeners came from Germany and were native German speakers. Values are expressed in percentages.

		G1 (24 listeners)	G2 (25 listeners)	G3 (20 listeners)
Age:	18-25	12.5	44.0	25.0
	26-35	54.2	44.0	50.0
	36-45	20.8	8.0	10.0
	>45	12.5	4.0	15.0
Gender:	Female	54.2	60.0	45.0
	Male	45.8	40.0	55.0

I analyzed the speech quality scores gathered in each session to identify and remove the extreme outliers ratings, those located at a distance from the median equal or higher than $3.0 \cdot IQR$ (Interquartile Range) [81]. Consequently, I discarded 53, 46, and 85 ratings in G1, G2, and G3, respectively. See Table 5.4 for a summary. The remaining ratings are then considered for further analysis.

Table 5.4: Ratings discarded deemed extreme outliers that were found in each session in groups 1, 2, and 3.

Session	Group 1	Group 2	Group 3
CSLvl0	22	13	37
CSLvl1	0	11	16
CSLvl2	17	22	16
CSLvl3	14	0	16

In the following, I first determine the validity of the mean opinion scores (MOS) collected in the first session (CSLvl0-MOS) by comparing them to the Lab-MOS. Note that both the laboratory study and CSLvl0 were conducted without background noise. Then, to investigate the background noise's influence on the speech quality ratings, I contrasted the CSLvl0-MOS against the MOS values gathered in the remaining three sessions. This analysis was made for each of the three groups.

5.1.2.1 Laboratory vs. CSLvl0

I calculated the Pearson's product-moment correlation and the Root Mean Square Error (RMSE) to determine the relationship between the ratings collected in the laboratory and CSLvl0. The correlation coefficient proved to be strong and significant in each group as well as low the RMSE. Table 5.5 outlines these results. This outcome indicates the validity of the collected quality scores at the first session in each of the groups.

Additionally, I analyzed the scores per degradation condition to investigate if there were statistically significant differences between the speech quality ratings provided by the participants in the laboratory and CSLvl0. To this end, I run a Mann-Whitney U test [101].

Table 5.5: The table presents the Pearson's correlation and Root Mean Square Error (RMSE) between the Lab-MOS and the ratings gathered in each group's first session (CSLvl0) without background noise.

Group	Session	# of listeners	r	$RMSE$
G1	CSLvl0	24	0.88*	0.379
G2	CSLvl0	25	0.87*	0.406
G3	CSLvl0	20	0.83*	0.502

$*p < 0.001$

Out of the 15 conditions that were under test, the median (Mdn) of the rating scores were statistically significantly different for only four conditions in G1 and G3 and only five conditions in G2. Table 5.6 presents these results.

Table 5.6: The table presents the results of the Mann-Whitney U test per condition between the laboratory and CSLvl0 for all groups.

Cond.	G1			G2			G3		
	U	z	p-value	U	z	p-value	U	z	p-value
2	166.5	-2.537	= .011	129.0	-3.447	= .001	144.0	-2.292	= .022
32	483.5	4.049	< .001	473.0	3.476	= .001	448.5	4.936	< .001
33	–	–	–	–	–	–	363.0	2.936	= .003
43	418.0	2.711	= .007	415.0	2.323	= .02	390.0	3.586	< .001
45	395.0	2.228	= .026	421.5	2.460	= .014	–	–	–
47	–	–	–	195.0	-2.115	= .034	–	–	–

Additionally, I ran an independent-sample t-test to determine if the speech stimuli were evaluated differently by listeners in CSLvl0 than the listeners in the laboratory. A visual inspection of a boxplot showed no outliers in the data. The quality scores from the laboratory and CSLvl0 were approximately normally distributed according to a generated Q-Q Plot. There was homogeneity of variances, as determined by Levene's test for equality of variances ($p = .197$, $p = .554$, and $p = .626$ in G1, G2, and G3, respectively). This t-test revealed that there was not a significant statistical difference between the quality scores provided by the listeners in the laboratory and those in CSLvl0 from all groups, i.e., $t(718) = 1.067, p = .286$ in G1, $t(733) = -.208, p = .836$ in G2, and $t(658) = 1.838, p = .067$ in G3. This result set the baseline for the analysis of the other three sessions (i.e., CSLvl1, CSLvl2, and CSLvl3) that were performed in the presence of environmental background noise.

One of the reasons for the differences found with the Mann-Whitney U test might be that listeners in our simulated crowdsourcing study executed the listening test with their headphones and not a professional one. It seems that they did not perceive the background noise in the signal of stimuli from condition two and therefore provided higher quality scores than the laboratory participants. Moreover, they could not distinguish the quality of the speech samples from conditions 32, 43, and 45 accurately and provided lower quality scores than the laboratory listeners. Another reason might be that participants in our experiment assessed only 15 conditions, whereas 50 conditions were tested in the laboratory. Nevertheless, listeners executed a training session before the assessment with an anchoring step to counteract this effect. Still, listeners in the laboratory had a higher discrimination capacity, and thus, these differences were observed.

5.1.2.2 Influence of Background Noise

To investigate the influence of the environmental background noise on the speech quality ratings, listeners performed three more times the listening test in the presence of noise (i.e., CSLvl1, CSLvl2, and CSLvl3). As previously pointed out, participants in G1 and G3 were exposed to street traffic background noise and those in G2 to a "TV-Show" noise signal. The noise was reproduced at a different level in each of the sessions, i.e. 50dB(A), 55dB(A) and 65.5dB(A) in G1, 60dB(A), 70dB(A) and 80dB(A) in G2 and 36dB(A), 43dB(A) and 50dB(A) in G3, respectively. See Table 5.1.

I ran a Wilcoxon signed-rank test to investigate the differences between the listeners' speech quality ratings in CSLvl0 and CSLvl1. I performed this analysis per conditions and for all groups. The difference between scores was approximately symmetrically distributed. This test revealed that the medians of the quality scores were statistically significantly different in 10 and 8 of the speech degradation conditions under test in G1 and G2, respectively. These results are presented in Table 5.7. In G3, significant differences were only found for condition number 33, i.e., $z = 2.391, p = .017$.

Table 5.7: Results of the Wilcoxon signed-rank test showing the speech degradation conditions rated statistically significantly different between CSLvl0 and CSLvl1 that were found in each group. In G3 only condition 33 was rated significantly different, i.e., $z = 2.391, p = .017$.

Cond. No.	G1		G2	
	z	p-value	z	p-value
2	2.567	= .01	2.933	= .003
6	-3.684	< .001	-3.342	= .001
7	-3.824	< .001	-3.507	< .001
32	2.930	= .003	–	–
33	3.943	< .001	3.595	< .001
44	-3.374	= .001	-2.542	= .011
45	-2.649	= .008	–	–
46	2.196	= .028	2.642	= .008
48	-2.122	= .034	2.123	= .034
49	-3.188	= .001	–	–
50	–	–	2.756	= .006

Ten conditions rated statistically significantly different between CSLvl0 and CSLvl1 in G1, and eight in G2 is a relatively high value, considering that there were 15 speech degradation conditions under test. These results infer that a speech quality assessment test conducted in a room with a level of the environmental background noise of 50dB(A) or more would lead to significantly different results compared to the silent condition.

Moreover, I analyzed whether there were statistically significant differences between the quality ratings collected in CSLvl0 and the ones gathered in the remaining two sessions in G3. A Wilcoxon signed-rank test showed that there were only five speech degradation conditions rated statistically significantly different when comparing CSLvl0 against CSLvl2 (43dB(A)), and nine conditions when contrasting CSLvl0 versus CSLvl3 (50dB(A)). More details can be found in Table 5.8.

Table 5.8: Results of the Wilcoxon signed-rank test executed in Group 3. The table shows the conditions rated statistically significantly different between CSLvl0 and CSLvl2, and CSLvl0 and CSLvl3.

Cond.	CSLvl0	CSLvl2			CSLvl3		
	Median	Median	z	p-value	Median	z	p-value
3	3.625	3.875	2.080	= .038	4.000	1.987	= .047
6	4.000	3.625	-2.670	= .008	3.125	-3.642	< .001
7	2.750	2.250	-2.514	= .012	1.625	-3.280	= .001
33	1.500	1.875	2.054	= .040	2.000	3.014	= .003
44	2.625	–	–	–	1.875	-2.490	= .013
45	3.500	2.875	-3.158	= .002	2.625	-3.420	= .001
48	3.000	–	–	–	2.750	-2.619	= .009
49	2.500	–	–	–	1.875	-2.635	= .008
50	1.750	–	–	–	2.125	2.758	= .006

These results are in line with those from G1 when comparing CSLvl0 versus CSLvl1 (50dB(A)). Participants in G1 evaluated statistically significantly different ten degradation conditions while a street traffic noise was being played at 50dB(A) (see Table 5.7). Indeed, an independent-samples t-test revealed that there was not a significant statistical difference between the quality scores provided by the listeners from G1 at CSLvl1 (50dB(A)) and listeners from G3 at CSLvl3 (50dB(A)) that executed the test under the influence of street traffic background noise ($t(658) = 1.861, p = .063$). This outcome confirms that a speech quality assessment test executed under the influence of background noise at a level of 50dB(A) would lead to unreliable results. Moreover, our findings suggest that a speech quality test conducted in the presence of environment background noise of 43dB(A) or lower would only in rare cases differ from a test carried out in silence, comparing the resulting MOS scores.

Furthermore, I executed a one-way repeated measures ANOVA to determine whether the gathered MOS values in G3 were statistically significantly different over the test sessions performed under different background noise conditions. There were no outliers, and the data were approximately normally distributed. The assumption of sphericity was violated, as assessed by Mauchly's test of sphericity, $\chi^2(5) = 133.257, p < .001$. Consequently, a Greenhouse-Geisser correction was applied ($\varepsilon = 0.754$). This ANOVA test revealed that the quality scores were statistically significantly different at the different test sessions, $F(2.261, 676.034) = 5.504, p = .003$. Then, a post hoc analysis with a Bonferroni adjustment showed that the quality scores provided by the listeners at CSLvl3 (under the influence of a 50dB(A) background noise) were statistically significantly different to the ones given at CSLvl0 ($p = .033$), CSLvl1 ($p = .009$) and CSLvl2 ($p = .02$). However, no significant difference was seen among the quality scores gathered between the sessions CSLvl0, CSLvl1, and CSLvl2. All in all, our results imply that the threshold of the level of environmental background noise for collecting reliable speech quality scores lies between 43dB(A) and 50dB(A).

In the case of G1 with street traffic background noise, when comparing CSLvl0 to the other two sessions, i.e., CSLvl2 (55dB(A)) and CSLvl3 (65.5dB(A)), the number of conditions that were rated statistically significantly different were 11 in CSLvl2 and 10 in CSLvl3. On the other hand, in Group 2, under the influence of "TV-Show" background noise, the number of significant differences ranges from 8 to 9 (CSLvl0 vs. CSLvl1 (60dB(A)), CSLvl3 (80dB(A)), respectively). Table 5.9 compiles these results. This outcome indicates that listeners are less distracted by the "TV-Show" noise than by the street traffic noise when executing the listening test. Thus, higher loudness values of the environment background noise in G2 (60dB(A), 70dB(A), and 80dB(A) for CSLvl1, CSLvl2, and CSLvl3, respectively) led to slightly less significant differences in the speech quality ratings when comparing to listeners in G1. Hence, the "TV-Show" environmental background noise was less intrusive for the participants than the street noise.

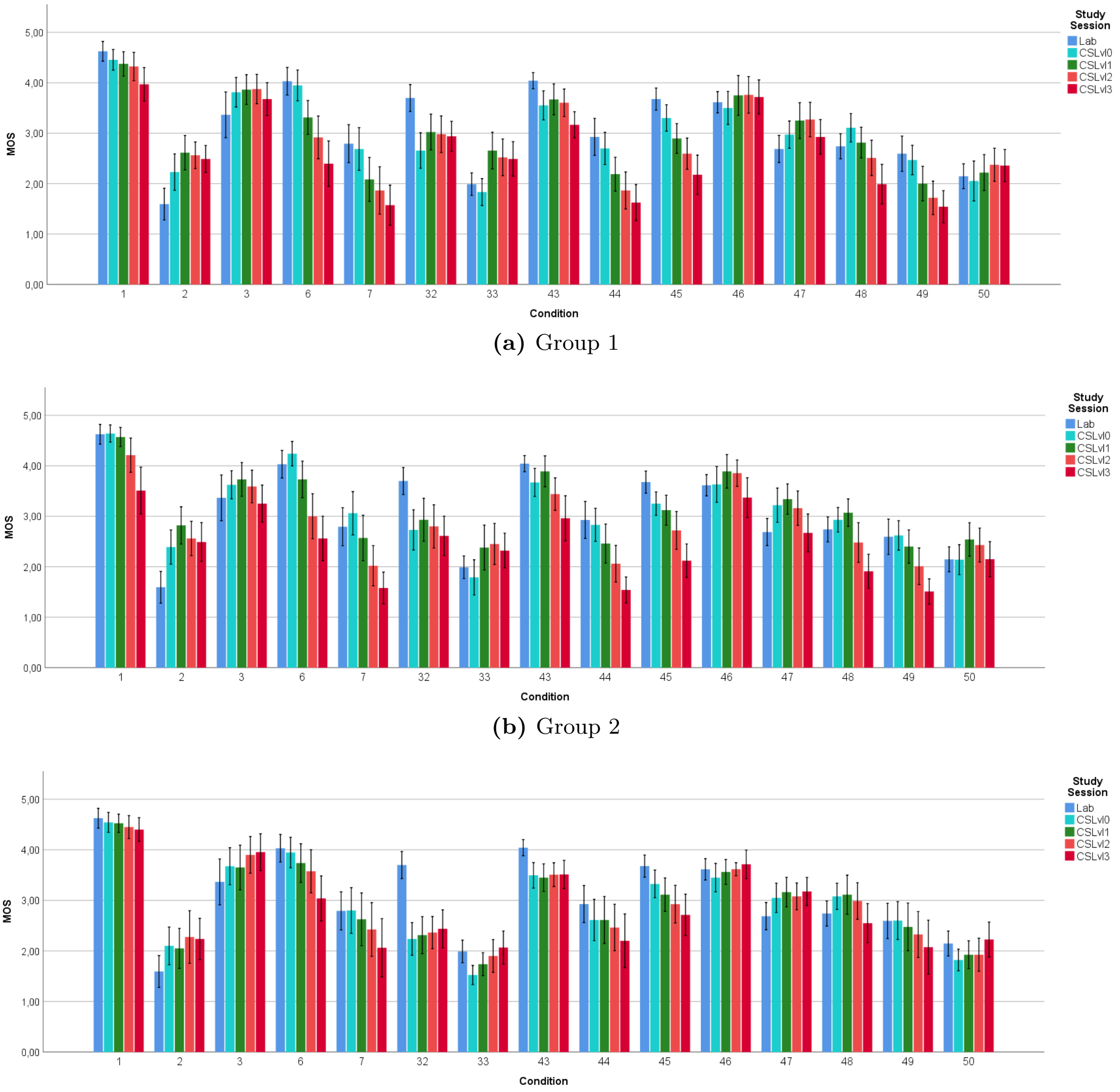
Table 5.9: Number of conditions rated statistically significantly different between CSLvl0 and the rest of the test sessions in Group 1 (G1) and Group 2 (G2).

G1 (street traffic noise)			G2 (TV-Show noise)		
CSLvl0 vs.	CSLvl1 (50dBA)	10	CSLvl0 vs.	CSLvl1 (60dBA)	8
	CSLvl2 (55dBA)	11		CSLvl2 (70dBA)	9
	CSLvl3 (65.5dBA)	10		CSLvl3 (80dBA)	9

5.1.3 Discussion

Figure 5.1 presents the MOS values per condition with 95% confidence intervals given by listeners in the laboratory and our study sessions in Group 1, 2, and 3. Considering these results, it can be inferred that the presence of background noise does not provoke a constant linear decrease or increase of the quality scores among an entire speech quality assessment test. Instead, it influences differently depending on the speech degradation condition under test. For instance, conditions 6, 7, 44, 45, 48, and 49 are characterized as being quiet. A clear trend was seen where listeners scored lower the quality of the speech stimuli than participants in the laboratory. They were unable to hear these speech samples properly due to the environmental background noise. This trend was observed in all groups.

Differently, conditions 2 and 3 were characterized by having background noise. In these cases, it seems that the environmental background noise masked the noise present in the speech signal, which prevented the listeners from perceiving the low quality of those stimuli. Instead, the participants in our study rated higher the quality of those speech samples than the listeners in the laboratory, see Figure 5.1.



(a) Group 1

(b) Group 2

(c) Group 3

Figure 5.1: Comparison between the Lab-MOS and the CSLvl0,1,2,3-MOS values per condition with 95% confidence intervals. Information about the degradation conditions can be found in Table 5.2 and more details in [9].

5.2 Analysis of Noisy Speech Quality Scores collected in Crowdsourcing Environments

As previously stated, a trend was found in the ratings given to similar speech degradation conditions as the background noise increased. More precisely, listeners in our study provided lower quality scores than the laboratory participants to the quiet speech impairment conditions. In contrast, conditions with background noise received higher ratings in our experiment than in the laboratory, see Subsection 5.1.3. Therefore, there is a need for instrumental models to correct the bias that can be found in the speech quality ratings collected in noisy crowdsourcing environments. Such models would be beneficial for using quality scores that would otherwise be discarded.

This section investigates a method for correcting the bias found in quality ratings given to speech stimuli with background noise and attenuated speech samples in the presence of environmental noise. Specifically, I determine the applicability of different regressor models to estimate the difference between the MOS collected in quiet and noisy environmental conditions.

The models were trained with data collected from the participants of the study groups 1 (G1) and 3 (G3) that conducted the listening test while being exposed to street traffic environmental background noise. The features were derived from statistics metrics calculated from the quality ratings given to each speech degradation condition and the level of environmental noise. Additionally, I tested two approaches to apply prior training to overcome the problem of imbalanced data. Finally, I present the results of an extensive model tuning for optimizing the regressor predictions.

5.2.1 Speech Quality Scores

Out of the 15 conditions under test, I only used the ratings provided to the six attenuated degradation conditions and the assessments made to the four speech impairment conditions with noise. Table 5.10 shows the selected conditions. The other conditions were not included since no relationship was found between the listeners' ratings, the characteristics of the degradation, and the environmental background noise.

I computed the Pearson's product-moment correlation and the Root Mean Square Error (RMSE) to determine the relationship between the ratings collected in the laboratory and CSLvl0, 1, 2, and 3. These results are presented in Table 5.11. A positive and significant correlation with the Lab-MOS and low RMSE was seen in both groups. As expected, the correlation coefficient decreased with increasing background noise.

As previously pointed out, six of the impairment conditions were distinguished by the speech signal's attenuation in the range of -8dB to -20dB (see Table 5.10). A clear trend was seen in both groups (G1 and G3), where listeners scored low the quality of these speech samples with the increase of the background noise. Figure 5.2 presents the MOS scores provided by listeners in both groups to the six speech degradation conditions. It can be seen that the quality scores degraded as the environmental noise at the listening side increased.

The other four conditions were characterized by degrading the speech signal with a background noise at the sending side varying from 12dB to 20dB (see Table 5.10). A slight trend can be observed in two of these conditions (i.e., 2 and 33 (see Figure 5.3)), where users

Table 5.10: Condition numbers and labels referring to the speech impairments under test [9]. The first six conditions were characterized by being attenuated, and the last four by having background noise.

Condition Number	Degradation Description
Attenuated conditions	
6	SWB Level -10dB
7	SWB Level -20dB
44	VoIP WB-Call + -16dB + acoust. send
45	VoIP WB-Call + -8dB + acoust. send
48	VoIP WB-Call + acoust. noise (rcv) + -8dB
49	VoIP WB-Call + acoust. noise (rcv) + -16dB
Conditions with noise	
2	SWB+Noise 12dB
3	SWB+Noise 20dB
33	EVRC-A + Noise 18dB SNR + Codec NS
47	VoIP WB-Call + acoust. noise (rcv)

Table 5.11: Pearson's (r) correlation and root mean squared error ($RMSE$) between the Lab-MOS and the MOS scores collected in all of the test sessions in Group 1 (G1) and Group 3 (G3).

Group	Session	r	p	$RMSE$
G1	CSLvl0	0.877	$< .001$	0.405
	CSLvl1	0.721	$= .002$	0.593
	CSLvl2	0.632	$= .011$	0.726
	CSLvl3	0.526	$= .044$	0.922
G3	CSLvl0	0.784	$= .001$	0.534
	CSLvl1	0.793	$< .001$	0.519
	CSLvl2	0.754	$= .001$	0.562
	CSLvl3	0.71	$= .003$	0.638

provided a higher speech quality score in the test sessions with environmental background noise than without background noise. Figure 5.3 shows the MOS scores with 95% confidence intervals provided by listeners in both groups to the four speech degradation conditions.

5.2.2 Model

In the following, I test the feasibility of some ensemble-based and non-linear models to predict the difference between the MOS collected in the test session with background noise and the MOS gathered in the silent test condition.

5.2.2.1 Feature Selection

The model's input features were calculated per listener, speech degradation condition, and level of noise (i.e., test session). The target variable was computed by subtracting the average of the ratings per speech degradation condition provided by a single user in each of the test session to the MOS per degradation condition given by the rest of the participants in CSLvl0.

I computed the standard deviation (StdDev) out of the four ratings provided by each participant to each speech degradation condition. Additionally, I scaled the ratings using the

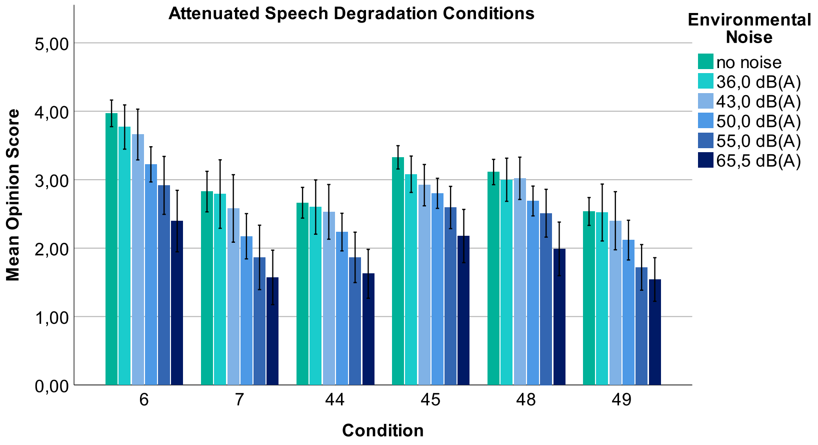


Figure 5.2: MOS provided by listeners in G1 and G3 to the six speech degradation conditions in the presence of environmental background noise. “no noise” and “50,0 dB(A) comprise the samples from both G1 and G3. Information about the degradation conditions can be found in Table 5.10 and in [9].

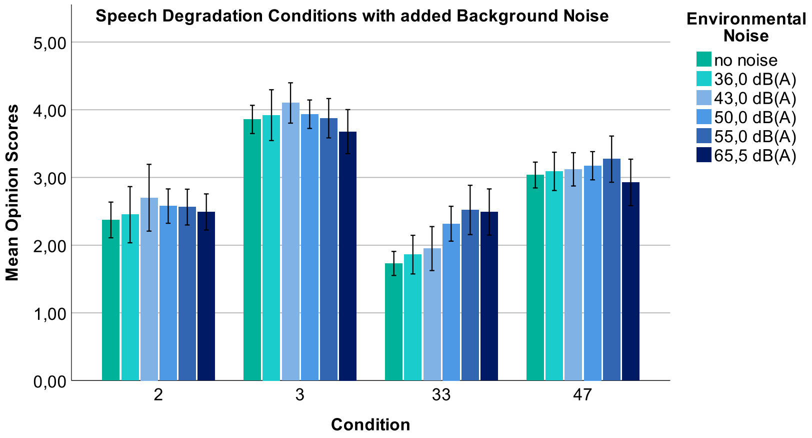


Figure 5.3: MOS provided by listeners in G1 and G3 to the four speech degradation conditions in the presence of environmental background noise. “no noise” and “50,0 dB(A) comprise the samples from both G1 and G3. Details about the degradation conditions can be found in Table 5.10 and in [9].

scale generic function of the R package “base” and calculated the standard deviation of these scaled scores (StdDev2). The scale factor (SCALE), the variance (VAR), and the median absolute deviation (MAD) were also considered as input features. Moreover, I calculated an intra-rater reliability (IRR) score by computing the intraclass correlation coefficient (ICC). Since I was interested in the degree of agreement in the absolute values across the ratings from a single worker, I used a “one-way random” model to compute the ICC coefficient. For this purpose, the *icc* function of the R package “IRR” was employed using “agreement” and “single” as parameters [96]. To be able to calculate the IRR from the ratings provided by a single user to a specific speech degradation condition, the input to the ICC function were these ratings and the same ratings but randomized. All the input features were normalized based on z-scores to prevent bias of the regression coefficients and better evaluate their impact on the prediction.

5.2.2.2 Model Evaluation

Different regression models were trained to find the one yielding the best results on estimating the degradation of the speech quality scores due to the environmental background noise. The training was done on the data corresponding to the attenuated speech stimuli, i.e., six speech conditions, from now on referred to as “quiet speech” data, and also on the data corresponding to the speech samples with background noise, i.e., four speech conditions, from now on referred to as “noisy speech” data.

After computing the features and the target variable, I noted that the data was imbalanced. Figure 5.4 shows a histogram of the target variable. In the case of the quiet or attenuated speech data, it can be seen that most of the samples correspond to values between -0.5 and 2 (approx.). In contrast, for the “noisy speech” data, a large number of samples were between -0.5 and 0.5 (approx.). A common approach with such an imbalanced dataset is calculating weights and using them as parameters when training and testing. Otherwise, a model aiming to minimize the misclassification error would be inherently biased towards the majority class and produce inaccurate predictions.

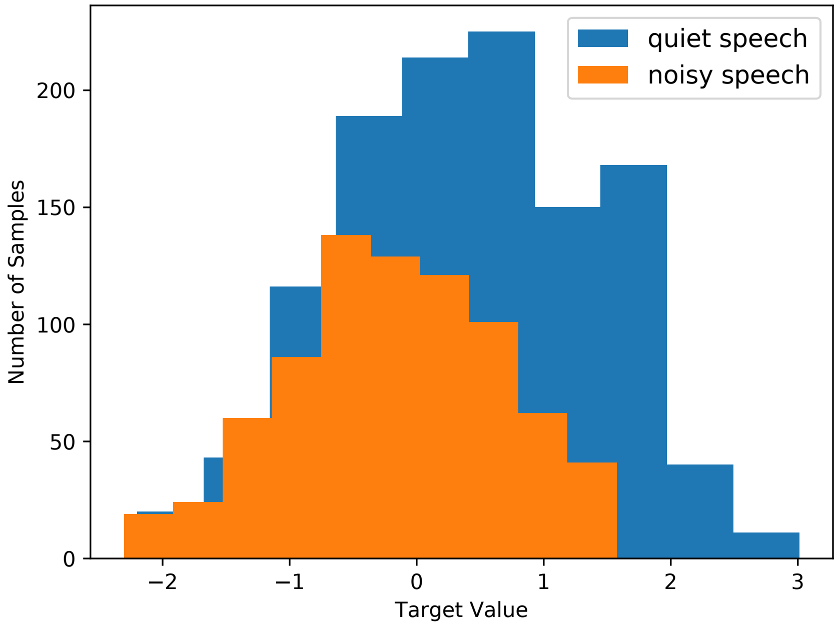


Figure 5.4: The figure shows the imbalanced distribution of the target variable values corresponding to the attenuated speech stimuli and the speech conditions with background noise.

In a classification task, the weights are realized by dividing the total number of samples by the number of samples in a specific class. In contrast, in a regression problem like in our case, this approach can not be made since the target variable is a continuous number. Therefore, I applied k-means clustering to group the target variable into bins and then calculated and assigned the weights to these clusters as if they were classes. In this way, I addressed the imbalance problem of the data so I could obtain accurate predictions from the regressors under test. This approach would penalize the prediction made to the numeric space containing a large number of samples.

The regressors models to be trained with our dataset that support weighting were some *decision-tree based ensemble models* like: Bagging regressor (BR) [102] with an Epsilon-Support Vector Regression model as a base estimator, Random Forest (RFR) [103], Extra Trees

(ETR) [104], Ada Boost (ABR) [105], Gradient Boosting (GBR) [106], and Stacking regressor (SR) with a linear regression model and a RFR as base estimators. Also, I tested some *non-linear* regressors: Support Vector Regressor (SVR) with a polynomial function kernel and a K-nearest Neighbors (KNNR).

The models were evaluated in terms of the coefficient of determination (R^2) and the root-mean-squared-error (RMSE). I employed a 10-fold cross-validation technique, which leads to a robust estimation of the models under test. Our experiments are based on the implementations in the *"scikit-learn"* toolkit [107], and I used default model parameters in all cases for this initial evaluation. Table 5.12 presents these results. It can be seen that when training the model on the data from the speech degradation conditions that provoked an attenuation in the speech stimuli ("quiet speech"), the Gradient Boosting Regressor (GBR) yielded the best results as it had one of the best R^2 and $RMSE$ scores. In contrast, the models performed poorly (i.e., low R^2 and high RMSE) on the data from the speech degradation conditions that corrupted the speech stimuli with background noise ("noisy speech").

Table 5.12: Regressors evaluation. The coefficient of determination (R^2) and RMSE is the average of the 10-fold cross validation. Best results are shown in **bold**.

Regressor	quiet speech		noise speech	
	R^2	$RMSE$	R^2	$RMSE$
Ada Boost	0.816	0.496	0.50	0.65
Bagging	0.809	0.491	0.43	0.69
Extra Trees	0.774	0.521	0.32	0.73
Gradient Boosting	**0.826**	**0.479**	0.50	0.65
K-nearest Neighbors	0.776	0.483	0.40	0.64
Random Forest	0.793	0.495	0.40	0.68
Support Vector	0.754	0.576	0.26	0.74
Stacking	0.791	0.576	0.43	0.72

One of the reasons for the poor performance on the "noisy speech" data might be the fact that there were only four conditions in this case, and in only two it could be seen a slight trend in the speech quality ratings with the increase of the background noise (see Figure 5.3). Therefore, more data would be needed to evaluate the performance of the aforementioned models for correcting the bias on the quality ratings due to the environmental background noise. Thus, in the following, I focus the analysis on the "quiet speech" data.

Next, I investigate if training the model after applying a data augmentation technique would lead to better results. Data augmentation is used to even datasets and overcome the data imbalance problem. Specifically, I used the *Synthetic Minority Over-sampling Technique* (SMOTE) [108]. SMOTE uses an algorithm based on k nearest-neighbors to generate new samples of the minority classes. That is, given a sample x_i, a new sample x_{new} will be generated by selecting one of its k nearest-neighbors x_{zi} and applying the following formula:

$$x_{new} = x_i + \lambda \times (x_{zi} - x_i) \tag{5.1}$$

where λ is a random number in the range $[0, 1]$.

I applied the *SMOTE* algorithm to our “quiet speech” dataset using the implementation from the *imbalanced-learn* toolbox [109]. Only the GBR model was evaluated since it was the one showing the best results on the data, see Table 5.12. I conducted 10-fold cross-validation, and the average of the R^2 coefficient and RMSE was 0.88 and 0.44, respectively. These results suggest that it is better to use a data augmentation technique like SMOTE than a weighting approach, to estimate the deviation of the MOS scores due to the influence of background noise.

5.2.2.3 Model Tuning

The gradient boosting algorithm [106] adds new decision trees based models using an ensemble technique called *boosting*. In doing so, it permits to correct the errors made by existing models and this process is repeated until there are no further improvements. As previously pointed out, the GBR model was trained using the default parameters. I then conducted an extensive hyper-parameter tuning using grid search over the relevant values of the regressor parameters with 10-fold cross-validation. The goal was to find the best parameter constellation. Again, the metrics used for the hyper-parameter optimization were R^2 and RMSE. The default and optimized values of the parameters are shown in Table 5.13. The tuned GBR model achieved a R^2 score of 0.90 and RMSE of 0.416.

Table 5.13: Important hyper-parameters of the GBR model and their default and optimized values. Information regarding the hyper-parameters can be found in [110].

Hyper-parameter	Default	Optimized
number of boosting stages	100	80
min_samples_leaf	1	5
max_depth	3	6
alpha	0.9	0.5

5.2.3 Discussion

The proposed Gradient Boosting Regressor can be applied when information is available about the environmental background noise characteristics of listeners in crowdsourcing. Further investigation would be needed to evaluate different approaches to fulfill this requirement. Either by including a step in the listening test asking crowd-workers to assess the environmental noise with a sound meter app or recording through the audio web-API a few seconds of the workers’ environmental noise and using this data to make automatic predictions about the characteristics of the background noise.

I encourage using the proposed bias correction method on ratings from speech degradation conditions of similar characteristics. Otherwise, the uncertainty of the algorithm would increase as the RMSE would be high.

5.3 Environment Background Noise Classification

This section investigates the use of different features for environmental noise classification in terms of type and level of noise. Our goal is to determine if it is possible to infer information about the workers' environment from small environmental audio recordings gathered through the audio-web API when workers perform the listening test. In a previous section, I showed the noises that might distract workers when executing tasks in crowdsourcing (see Subsection 3.3.4), and in Section 5.1, I determined the influence of the environmental background noise in the speech quality scores. Hence, a mechanism is needed to determine the characteristics of the workers' environment. Such a mechanism would be useful for detecting when crowd-workers do not follow the study's instructions and execute the listening test in noisy environments.

5.3.1 Environment Background Noise Collection

I created a web application for collecting environmental audio recordings. These web-audio recordings conform the two dataset that will be introduced in the following sections for noise classification and noise level estimation. This web-App had four sections. The first section contained instructions to disable the noise reduction options enabled by default in Windows and macOS computers. These options were: *"Disable all sound effects"* in the *"Microphone Properties"* in Windows, and the checkbox: *"Use ambient noise reduction"* in the sound settings in Mac computers. The second section comprised the instructions for performing the recordings emulating a realistic usage scenario (e.g., to record noises from a TV on, users had to turn on the TV at a normal volume and run the web recording App from where they usually sit down to work).

The third section included the list of nine different environmental noises to emulate and record:

- Watching TV or TV-Show
- Listening to Radio
- Listening to Music
- Coffee Machine
- Dishwasher
- Water Heater
- Street/Traffic Noises
- People Talking
- Quiet

These noises were selected based on our findings of the most common environmental noises affecting German crowd-workers when they enroll in crowdsourcing tasks (see Subsection 3.3.4). Finally, the web-App had two text input fields and the button to trigger the recording. The input field served to assign a unique identifier to the recording and to provide information about the noise's loudness.

This web-App was either used in the laboratory with the background noise simulated according to the definitions of Section 3.4 and loudness measurements with the artificial head HMS II.3 HEAD acoustics. Or used at home in real-life scenarios with noise level measurements from the "db Meter" [1] mobile application available for Android and IOS devices.

5.3.1.1 Dataset for Noise Classification

The Dataset for Noise Classification (DNC)[2] [111] consisted of 4377 labeled environmental recordings approximately equally balanced between three main categories, i.e., *"mechanic"*, *"melodic"*, and *"quiet"*. I decided to group the recordings into these three categories, as it was found in Subsection 3.3.4.7, that these are the type of background noises that might impact the results of speech quality evaluations carried out in crowdsourcing. The environmental recordings had an average duration of 14.6 seconds, representing 17.8 hours of background noise data. Table 5.14 shows the number of files per category and the noise classes contained in each category.

Table 5.14: Number of recordings per category and types of noises or classes contained in each category.

Noise Classes	Noise Category	Files per Category
coffee machine dishwasher water heater street, traffic	mechanic	1545
TV, TV-Show music radio people	melodic	1427
quiet	quiet	1405

5.3.1.2 Dataset for Noise Level Estimation

The Dataset for Noise Level Estimation (DNLE)[3] [112] contained 1668 recordings labeled according to type and level of noise. The loudness measurements were 50.7dB(A) on average, varying from 30.6dB(A) to 81.3dB(A). The environmental recordings were 15.0 seconds long on average, which summed up to 6.95 hours of background noise data in total. As in the DNC dataset, the recordings were arranged into three main categories, i.e., *"mechanic"*, *"melodic"*, and *"quiet"*. Table 5.15 below shows the number of files per category, the noise classes in each category, and information regarding each category's loudness levels.

[1] http://dbmeterpro.com/

[2] https://github.com/zequeira/DNC.git

[3] https://github.com/zequeira/DNLE.git

Table 5.15: Number of environmental recordings per category and loudness levels in each category.

Noise Classes	Noise Category	Files per Category	Loudness Average (dBA)	min / max (dBA)
street, traffic dishwasher	mechanic	576	58.6	46.5 / 70.0
TV, TV-Show music radio	melodic	580	59.6	32.0 / 81.3
quiet	quiet	512	31.4	30.6 / 33.0

5.3.2 Experiment BN1

The experiment BN1 examines the suitability of *mel-frequency cepstral coefficients* (MFCC) for classifying the background noise from web audio recordings. Spectral features like MFCC have been widely adopted in speech and music [113] applications due to its coefficients' stability against signal deformations [114, 115, 116]. The MFCC is calculated by applying a *discrete cosine transform* (DCT) to the log-mel-spectrogram as if it were a signal. The amplitudes of the resulting spectrum represent the MFCC coefficients.

I investigate the optimal number of MFCC coefficients for identifying the environment noise category. To this end, I trained multiple state-of-the-art machine learning classifiers with different feature sets. Each set comprised a different number of MFCC coefficients varying from 5 to 31 (i.e., an odd number of coefficients from 5 to 31), and as a result, all classifiers were tested 14 times.

I used the librosa [117] python package with default settings[4] to compute MFCC features from the "DNC" dataset. As previously stated, DNC has 17.8 hours of environmental recordings grouped into three main categories, i.e., *"mechanic"*, *"melodic"*, and *"quiet"* (see Subsection 5.3.1.1). All features were normalized based on z-scores to avoid bias and to evaluate the classifiers' performance better.

The classifiers under test were: k-nearest neighbors (KNN) [118], Random Forest Classifier (RFC) [103], Extra Trees Classifier (ETC) [104], optimized Gradient Boosting Classifier (XGBC) [106], Ada Boost Classifier (ABC) [119], Bagging Classifier (BC) [102], and Multi-layer Perceptron Classifier (MLPC) [120]. Table 5.16 presents a list of these classifiers with their key parameter settings.

The classifiers were evaluated in terms of accuracy. I employed a 5-fold cross-validation technique, which leads to a robust estimation of the models under test. Our experiment is based on the implementations in the *"scikit-learn"* toolkit [107]. When not specified, the parameters had default values.

5.3.2.1 Results

Figure 5.5 presents the experimental results. It can be seen that the accuracy increased with the increase in the number of MFCC coefficients. However, at the cost of more resources and time needed during the training process. In fact, it was not possible to train the "Bagging" classifiers

[4] https://librosa.org/doc

Table 5.16: The table shows the classifiers under test with their main parameter configuration.

Classifier	Parameters Setup	Classifier Abbreviation
k-nearest neighbors (KNN)	No. neighbors: 5	KNN-5
Random Forest Classifier (RFC)	No. estimators: 50	RFC-500
Extra Trees Classifier (ETC)	No. estimators: 50	ETC-50
Gradient Boosting Classifier (XGBC)	tree method: approx	XGBC
Ada Boost Classifier (ABC)	base estimator: RFC(n_estimators=10)	ABC-RFC10
Ada Boost Classifier (ABC)	base estimator: RFC(n_estimators=20)	ABC-RFC20
Ada Boost Classifier (ABC)	base estimator: ETC(n_estimators=20)	ABC-ETC20
Bagging Classifier (BC)	base estimator: KNN(n_neighbors=5)	BC-KNN5
Bagging Classifier (BC)	base estimator: decision tree classifier	BC-DTC
Bagging Classifier (BC)	base estimator: extra tree classifier	BC-ETC
Multi-layer Perceptron Classifier (MLPC)	hidden_layer_sizes: (40, 30, 30, 30, 3); solver: adam	MLPC-adam
Multi-layer Perceptron Classifier (MLPC)	hidden_layer_sizes: (40, 30, 30, 30, 3); solver: lbfgs	MLPC-lbfgs
Multi-layer Perceptron Classifier (MLPC)	hidden_layer_sizes: (40, 35, 35, 35, 3); solver: adam	MLPC-adam+

with more than 23 MFCC coefficients, or the “ExtraTrees”, “Random Forest”, “K-nearest neighbors”, and optimized “Gradient Boosting” classifiers with more than 25 MFCC coefficients. Training these classifiers with a high number of MFCC coefficients was computationally too expensive, and therefore no results were obtained in these cases.

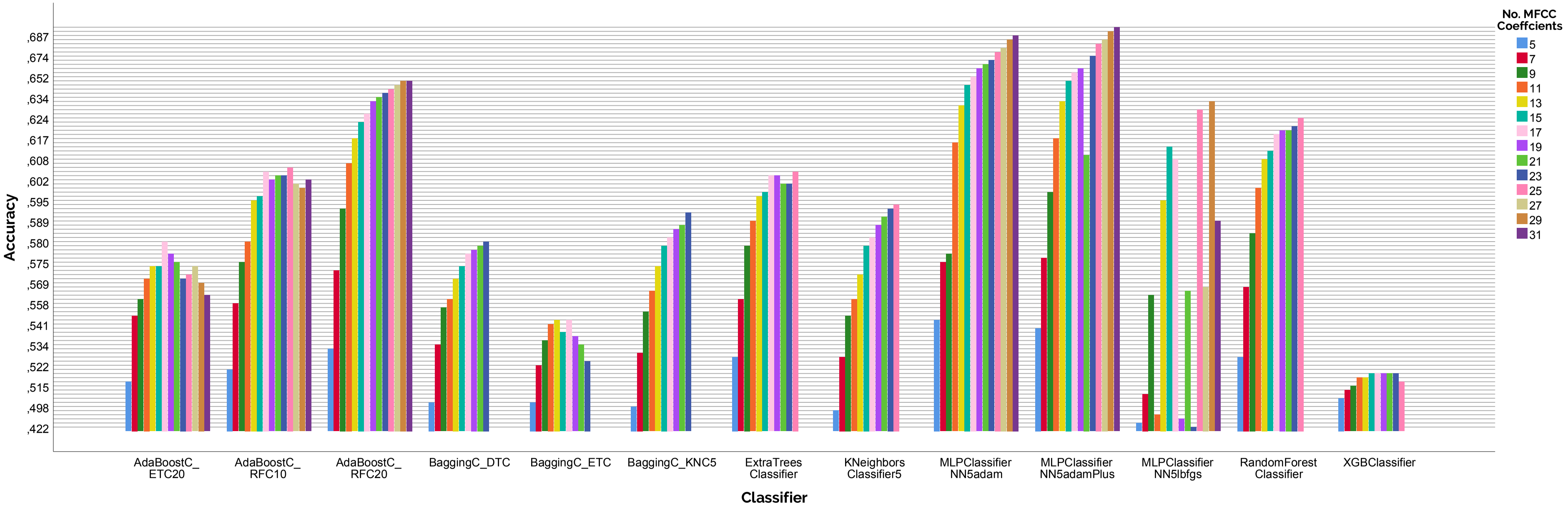


Figure 5.5: The figure shows the classifier's accuracy by increasing the number of MFCC coefficients from 5 to 31. The best accuracies were 0.64 and 0.69 and were achieved with the Ada Boost Classifier with a Random Forest Classifier as a base estimator and a Multi-layer Perceptron Classifier with *adam* solver, respectively.

The two classifiers with the best performance were the Ada Boost Classifier with a Random Forest Classifier with 20 estimators as a base estimator (ABC-RFC20), and the two architectures of the Multi-layer Perceptron Classifier with the *adam* solver for the weight optimization (MLPC-adam). The highest accuracy was 0.64 and 0.69 for the ABC-RFC20 and MLPC-adam classifiers, respectively.

A closer look at the performance evolution of these classifiers with the increase of the MFCC coefficients revealed that the accuracy did not improve significantly after 19 or 21 coefficients. This outcome suggests that a good trade-off might be to use a number of MFCC coefficients around 20 to accomplish a fair balance between accuracy and computational resources for noise classification from web audio recordings.

5.3.3 Experiment BN2

This subsection investigates the suitability of different spectral and chroma features for environmental noise level estimation from web audio recordings. I used the librosa[5] package with default parameters to derive the following 65 features from the DNLE dataset:

- 20 MFCC
- 20 delta-delta
- 1 spectral centroid
- 12 chromagram
- 12 chroma energy normalized (CENS).

Delta-delta features are calculated by computing the local estimate of the second derivative of MFCC features. They are also called acceleration coefficients and introduce an even longer temporal context. Spectral features have been successfully employed in sound classification [114, 115, 121] and audio event recognition [122] tasks. The spectral centroid is calculated by normalizing each frame of a magnitude spectrogram and treating it as a distribution over frequency bins, from which the mean or centroid per frame is extracted [123].

The chroma features include the computation of a chromagram from a waveform or power spectrogram [124] and a variant of “Chroma Energy Normalized” (CENS). Chroma features are a powerful representation of the audio signal in which the entire spectrum is projected onto 12 bins representing the 12 different semitones or chroma of the musical octave. For certain audio signals, knowing the chroma distribution even without the absolute frequency (i.e., the original octave) can provide useful information about the audio, and may even reveal perceived similarities that are not apparent in the original spectra. On the other hand, CENS features take statistics over large windows to smooth out local deviations in tempo, articulation and musical ornaments. Chroma features are robust to dynamics, timbre, and articulation. Therefore, they are commonly used in audio matching and retrieval applications [125].

All features were normalized based on z-scores. A preliminary analysis revealed that the data was imbalanced. The histogram below (see Figure 5.6) shows a high difference between

[5] https://librosa.org/doc

the number of samples available around 40dBA and above 70dBA, compared to the samples in the other ranges. To overcome this imbalanced data problem, I used the same approach employed in Subsection 5.2.2.2. K-means clustering was applied to the target variable and I computed weights based on the resulting groups. Then, I used this weight vector on each training step to correct the result of the "loss function".

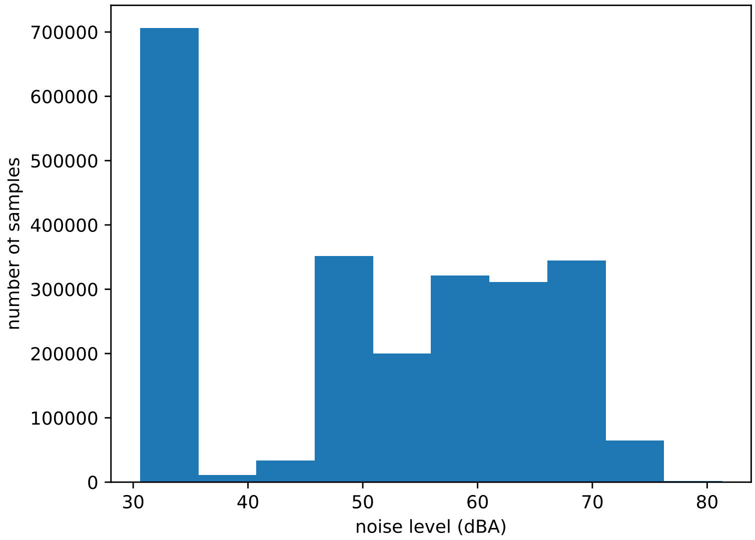


Figure 5.6: The figure presents the imbalanced distribution of the target variable in the dataset for noise level estimation. There is a low number of samples around 40dBA and above 70dBA compared to the rest.

I trained a deep neural network model based on a *"Long Short-Term Memory"* (LSTM) [126] architecture for estimating the level of noise. LSTMs are a specific Recurrent Neural Network (RNN) architecture designed to model temporal sequences and their long-range dependencies more precisely than conventional RNNs. LSTM and traditional RNNs have been successfully applied to various sequence prediction and sequence labeling tasks [127, 128].

I hypothesize that the ability of LSTMs to model temporal sequence dependencies could be beneficial for noise level estimation. Most noises in the DNLE dataset are not constant (e.g., television, music, radio, street traffic). Therefore, an LSTM can potentially take into account the temporal characteristics of the noise to make predictions.

The LSTM architectures are designed to integrate information in one direction along a particular dimension, e.g., forward in time. On the other hand, it might be beneficial in some applications to integrate data across both directions. This can be achieved by *"Bi-directional recurrent networks"* (BRNNs) by implementing a reduction to the standard one-directional architecture [129]. This approach was successfully adapted in [130] to LSTM architectures, i.e., BiLSTM.

Bi-directional networks and BiLSTM in particular, have been demonstrated to be effective for a wide range of audio analysis tasks, varying from speech recognition [131, 132] beat tracking [133], and event detection [134]. In general, bi-directional networks are more powerful and generally preferred unless forward-sequential processing is required. For instance, research in [134] employed a multilabel bi-directional long- and short-term memory (BiLSTM) recurrent neural network for polyphonic sound event detection from real-life recordings.

In this thesis, our model[6] is composed of three fully connected LSTM layers, plus three fully connected bidirectional LSTM (BiLSTM) layers, plus three fully connected LSTM layers and a linear regression layer as an output. As previously stated, the BiLSTM layers operate on the input sequence in both directions. Table 5.17 summarizes the architecture of our model:

Table 5.17: Architecture of the deep neural network based on LSTM and BiLSTM.

Layer Type	No. Layers
LSTM	3
BiLSTM	3
LSTM	3
Regression	1

5.3.3.1 Results

The dataset was split into 90% for training and 10% for testing. The model was trained with an *"adam"* optimizer, a learning rate of 0.0005, and a batch size of 675. The proposed model achieved an RMSE of 4.58 on average with a minimum of 0.5 and a maximum of 14.4 (scale of 30.6dBA to 81.3dBA), and a standard deviation of 2.72 on the test dataset. The Figure 5.7 presents the resulting scatterplot of the predictions on the test dataset.

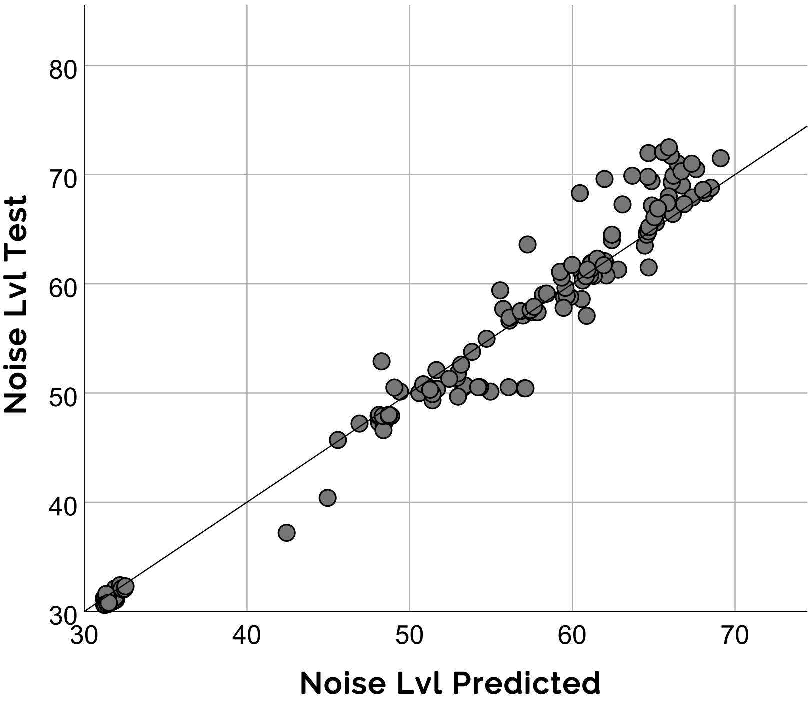


Figure 5.7: The figure shows a scatterplot of the LSTM-based deep model predictions on the test dataset.

[6] https://github.com/zequeira/ENVM-BiLSTM.git

5.3.4 Discussion

This section investigates the use of machine learning for environmental noise classification and noise level estimation from audio recordings collected through the audio-web API in crowdsourcing. Two datasets were created containing recordings of environmental noises that have been shown to affect workers when they perform crowd-work. One experiment was conducted where multiple classifiers were trained with MFCC features varying from 5 to 31 coefficients. The classifier's accuracy increased with the number of MFCC coefficients. However, the computational cost and training time increased with the number of MFCC coefficients, and in some classifiers, it was not possible to collect results for a high number of MFCCs.

Moreover, one of the highest accuracies of 0.64 was achieved with an Ada Boost Classifier with a Random Forest Classifier as a base estimator. A high accuracy was also achieved with a Multi-layer Perceptron Classifier with an *"adam"* solver for weight optimization (0.69). It was also observed that the accuracy did not increase significantly when employing more than 19 or 21 MFCC coefficients. Therefore, it is recommended to use a number of MFCC coefficients around 20 to achieve a good balance between accuracy and computational requirements. All in all, these results indicate that it is possible to classify the type of noise from audio recordings collected through the audio-web API.

In a second experiment, an LSTM-based deep neural network was trained using different spectral and chroma features to estimate the noise level from web-audio recordings. The proposed network architecture was based on fully connected LSTM and BiLSTM layers with a linear regression layer as output. The network achieved an RMSE of 4.58 and a standard deviation of 2.72 on the test dataset. This outcome indicates the validity of the LSTM-based neural network for environmental noise level estimation from web-audio recordings.

5.4 Conclusion

This chapter investigates the influence of environmental background noise on the results of a speech quality assessment test in crowdsourcing. A simulated crowdsourcing study was conducted in the laboratory with three groups of participants. They assessed the overall quality of speech stimuli in the presence of environmental background noise at various levels. The results indicate that the environmental background noise level threshold to achieve reliable speech quality assessment results in crowdsourcing lies between 43dB(A) and 50dB(A). The outcomes suggest that a background noise level of 50dB(A) on average leads to invalid results. In contrast, a level of 43dB(A) yielded reliable speech quality ratings in most cases.

Furthermore, users tolerate more the TV-Show noise when executing the listening test. The participants in Group 2 gave quality scores that were more in line with the quiet test condition than listeners in Group 1. Additionally, the results indicate that the presence of environmental background noise does not provoke that listeners would give consistently lower or higher quality scores among an entire speech quality assessment study. Instead, the effect of the background noise depends on the speech degradation condition in a test.

Moreover, this chapter presents an approach to correct the bias produced by the influence of the environmental background noise in the speech quality ratings given to attenuated speech stimuli in crowdsourcing. To this end, features were derived out of statistics computed from the quality ratings. The imbalanced characteristics of the data were corrected. Then a Gradient Boosting Regressor (GBR) model was fine-tuned and trained employing a 10-fold cross-validation technique. The model achieved a R^2 score of 0.90 and RMSE of 0.416.

Finally, this chapter investigates the feasibility of using audio recordings collected through the audio-web API for environmental noise classification. The goal was to determine if it would be possible to infer information about the workers' environmental characteristics from audio samples gathered when performing the listening test. With this goal in mind, two background noise datasets were created, and standard features were extracted and used for noise classification and noise level estimation.

One experiment was conducted where different classifiers were trained with MFCC coefficients varying from 5 to 31 to classify the type of noise. The classifiers were evaluated in terms of accuracy. The results indicate that accuracy increases with the number of MFCC coefficients but at the expense of the time and resources needed during the training process. The classifiers with the highest accuracy were an Ada Boost Classifier with a Random Forest Classifier as a base estimator (0.64) and a Multi-layer Perceptron Classifier with an *adam* solver (0.69).

The second experiment evaluated the use of deep neural networks for noise level estimation. To this end, different spectral and chroma features were extracted from web audio recordings and used to train a deep model based on a *"Long Short-Term Memory"* (LSTM) architecture. The LSTM-based model achieved an RMSE of 4.58 on average and a standard deviation of 2.72 on the test dataset.

6

Influence of Language

This chapter investigates the feasibility of conducting a speech quality study with listeners of a native language other than the speech dataset to be assessed. The motivation is that evaluating the quality of German speech samples is a cumbersome task. It is challenging to recruit German listeners, as the main crowdsourcing platforms do not have enough active German users. An alternative is to use clickworker, a German-based crowdsourcing platform. However, they lack basic functionalities like audio playback. Thus, it is necessary to implement a system to conduct a listening test, which could be challenging and expensive. Additionally, it is difficult to control the language proficiency of crowd-workers. Consequently, some listeners may end up participating in a listening test of a target language other than their mother tongue.

In this chapter, I determine the influence of assessing the quality of a German speech dataset with native English and Spanish speakers. To this end, three studies were conducted, which are outlined below, as well as the study results.

6.1 Study Setup

6.1.1 Speech Database

The stimuli for this experimentation were taken from the speech dataset SwissQual 501 from the ITU-T Rec. P.863 [9] competition. SwissQual 501 includes 200 speech samples carrying 50 different degradation conditions. Additionally, it incorporates subjective quality assessments from 24 different native German listeners. The resulting Lab-MOS scores are taken as a reference for the analysis presented in this chapter. Further information regarding SwissQual 501 can be found in Section 3.2.1 and in Appendix A.

6.1.2 Method

Three studies were conducted (i.e., E1, E2, and E3) to assess the quality of a German speech dataset. E1 was run with native German listeners while E2 and E3 were executed with native English and Spanish speakers.

The study E1 was conducted using the German-based clickworker crowdsourcing platform. Clickworker reported having 2.2 million users worldwide, 30% from Europe and 12% in total are native German speakers [1]. Then, the study was addressed to crowd-workers from Germany, Austria, and Belgium.

On the other hand, studies E2 and E3 were executed in Amazon Mechanical Turk (AMT). E2 targeted workers from the United States with a Human Intelligence Task (HIT) approval rate greater than 98% and more than 500 HITs approved. In contrast, E3 was addressed to workers from all Spanish speaking countries. Due to the low number of active Spanish workers in AMT, no restriction was set regarding the workers' performance history.

The crowdsourcing studies contained three phases (i.e., Qualification, Training, and Assessment), which are described in Section 3.3.3. In the following, I detail characteristics of the Qualification and the Assessment phase that are particular to the studies E1, E2, and E3.

Since I wanted to recruit workers with no knowledge of the German language in E2 and E3, the Qualification phase included a single choice question where users indicated from zero (no knowledge) to six (native) their understanding of German. Additionally, they listened to a passage in German and answered four content questions about the audio. Workers qualified to participate in the study when they answered zero or one to the German proficiency question and responded wrong to the content questions.

The Assessment phase included 15 speech stimuli plus one trapping question inserted randomly within the first five stimuli and one between the 10th and the 15th sample. All interfaces were presented in the native language of the listener. More details about the Qualification, Training, and Assessment phases can be found in Sections 3.3.3.1, 3.3.3.2, and 3.3.3.3, respectively.

6.2 Results

Overall, 233 workers executed the Qualification phase in E1. Thirty-five failed the included German test, and the remaining 198 workers were invited to participate in the study.

In the case of E2, 190 workers completed the Qualification. Fifteen of them were not invited to participate in the study. They were either not native English speakers or answered the German content questions correctly. On the other hand, 68 workers executed the Qualification to participate in study E3. Twelve were prevented from taking part in the study since they were not native Spanish speakers or were able to understand the German language.

A total of 64 listeners in E1, 53 in E2, and 40 in E3 participated in our study and produced 5400, 4179, and 3800 ratings, respectively. The workers' demographics are shown in Table 6.1. All workers in each group answered the trapping question correctly. Furthermore, I analyzed all of the speech quality scores collected in each group to identify and discard the ratings deemed extreme outliers, i.e., those located at a distance from the median equal or higher than $3.0 \cdot IQR$ (interquartile range) [81]. Then, 137, 154, and 76 ratings were removed in study E1, E2, and E3, respectively. The remaining quality scores were then considered for further analysis.

[1] https://www.clickworker.com/clickworker-crowd

Table 6.1: The table presents the demographic information of the 64, 53, and 40 workers that conducted the listening test in Study E1, E2, and E3, respectively. Values are expressed in percentages.

		E1 (64 workers)	E2 (53 workers)	E3 (40 workers)
Age:	18-25	25.0	11.3	45.0
	26-35	34.4	24.5	30.0
	36-45	18.8	30.2	17.5
	>45	21.9	34.0	7.5
Gender:	Female	53.1	52.8	32.5
	Male	46.9	47.2	67.5
Language:		German: 96.9 Russian: 1.6 Hungarian: 1.6	English: 100	Spanish: 100
Country:		Germany: 90.6 Austria: 9.4	US: 100	Spain: 52.5 Mexico: 17.5 Venezuela: 12.5 Argentina: 5.0 Colombia: 5.0 Costa Rica: 5.0 Ecuador: 2.5

6.2.1 Analysis of Laboratory vs. Study E1, E2, and E3

To determine the validity of the mean opinion scores (MOS) collected in the different studies, I compared the Lab-MOS to the E1-, E2-, and E3-MOS, respectively. Then, to analyze the influence of language mismatch in the speech quality ratings, I contrasted the Lab-MOS to the E1-MOS, and then the E1-MOS against the MOS values gathered in studies E2 and E3.

I calculated the Pearson's product-moment correlation and the Root Mean Square Error (RMSE) to assess the relationship between the ratings collected in the laboratory and crowdsourcing in studies E1, E2, and E3. A positive and significant correlation with the Lab-MOS and low RMSE was seen in each study regardless of the listeners' mother tongue. As expected, one of the highest correlations and the lowest RMSE were achieved in E1 with the German listeners, i.e., $r = 0.92 (p < .001); RMSE = 0.319$. Table 6.2 presents a summary of these results. This outcome indicates the validity of the ratings gathered in the different studies.

Table 6.2: The table presents the study sizes, Pearson's product-moment correlation, and the Root Mean Square Error (RMSE) between the Lab-MOS and the ratings collected in the crowdsourcing studies E1, E2, and E3.

Study	# of valid ratings	# of listeners	r	$RMSE$
E1 (native Germans)	5263	64	0.922*	0.319
E2 (native English)	4025	53	0.929*	0.467
E3 (native Spanish)	3724	40	0.862*	0.523

*$p < 0.001$

Additionally, I investigated whether there were significant differences between the speech quality ratings given by the listeners to each of the speech degradation conditions in the laboratory and E1. With this goal in mind, I run a Mann-Whitney U test [101] per condition with Šidák alpha correction to counteract the problem of multiple comparisons. The Šidák correction [135] is conducted under the premise that the executed tests are independent. Since all conditions are considered independently, the adjusted alpha level (α_{SID}) is determined by:

$$\alpha_{SID} = (1 - (1 - \alpha))^{\frac{1}{m}} \tag{6.1}$$

where α equals the unadjusted p-value (i.e., $\alpha = 0.05$) and m represents the number of independent conditions (i.e., 50 in our study). The Šidák correction gives a stronger bond than the Bonferroni correction. It can also be limited by the condition of independence and is less stringent in its control over the type I error [136].

A standard practice in statistical analyses is to accept an alpha level of 0.05 to identify statistical significance from non-significance. However, when conducting a high number of statistical tests, this approach will result in one variable that will appear to be significant when, in reality, it is co-incidental [137]. The occurrence of rejecting the null hypothesis when it is, in fact, true is referred to as a type I error.

I replaced the values of α and m in Eq.(6.1) and resulted in an alpha-corrected value of 0.001. Thus, the statistically significant effects are determined for p-values equal to or less than 0.001.

Out of the 50 conditions that were under test, I found that the median (Mdn) of the rating scores were statistically significantly different for 11 conditions. The results of the Mann-Whitney U test are presented in Table 6.3, together with the mean values per condition. Information about the speech impairment conditions can be found in Appendix A and more details in [9].

Table 6.3: Conditions rated statistically significantly different between listeners in the laboratory and those in the study E1.

Cond. No.	Lab Median	Lab Mean	E1				
			Median	Mean	U	z	p-value
5	3.00	2.938	4.00	3.718	2900.5	-5.262	$< .001$
12	1.00	1.198	1.00	1.590	4763.0	-3.649	$< .001$
22	3.00	2.844	3.00	3.385	3561.0	-3.649	$< .001$
26	1.00	1.479	2.00	2.350	2032.0	-7.690	$< .001$
27	1.00	1.135	1.00	1.574	3906.0	-5.712	$< .001$
29	3.00	2.990	3.00	2.558	6366.0	3.629	$< .001$
31	3.00	2.948	2.00	2.495	6371.0	3.733	$< .001$
34	2.00	1.906	2.00	2.505	3338.5	-4.149	$< .001$
36	2.00	2.250	2.00	1.923	6240.0	3.466	$= .001$
42	3.00	2.865	2.00	2.413	5701.0	3.654	$< .001$
43	3.00	2.844	4.00	3.523	4086.5	-4.413	$< .001$

Additionally, Figure 6.1 shows a scatterplot of the Lab-MOS and the E1-MOS for these 11 conditions. The figure shows that listeners in crowdsourcing overrated conditions 5, 12, 22, 26, 27, 34, and 43 compared to the laboratory participants. These conditions were either wideband

(WB) or super-wideband (SWB), and four of them in combination with the Advanced Audio Coding [2] (AAC) codec. In general, these conditions have in common that they caused a degradation in the speech that would be difficult to perceive if the listener is not completely focused on the test, or if the listening device is of bad quality. For instance, condition 5 had a subtle noise coupled to the speech signal that was apparently unnoticed by listeners in crowdsourcing. Also overlooked was the "electronic" sound of the speech in conditions 26 and 27. Thus, listeners in E1 provided higher quality scores to those speech stimuli.

Moreover, the underrated speech degradation conditions were characterized by being narrowband combined with the Enhanced Full Rate (EFR) or the Adaptive Multirate (AMR) codecs. The speech signal in most of these conditions was attenuated, and the speech sounded a bit artificial in conditions 29 and 31. It seems that listeners in E1 could not properly hear those attenuated stimuli, so they gave lower scores than participants in the laboratory.

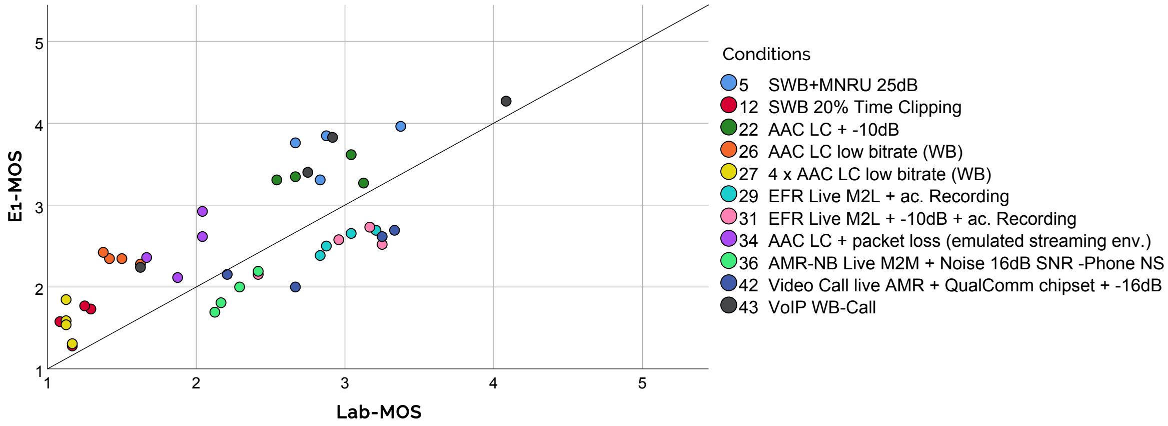


Figure 6.1: Comparison between the Lab-MOS and the E1-MOS. Represented only the conditions that were rated statistically significantly different. More details about the degradation conditions can be found in Appendix A and in [9].

All in all, the MOS differences between the laboratory and E1 were probably due to the hardware employed for the listening test. Laboratory listeners wore professional headphones, while crowdsourcing participants used regular headphones that they employ in their daily lives. Additionally, laboratory listeners had higher discrimination capacity as they assessed the entire speech dataset. In contrast, workers in crowdsourcing rated the quality of 15 speech stimuli per session. Nevertheless, the MOS scores collected in the study E1 are valid and reliable. There were only 11 conditions rated statistically significantly different between the laboratory and E1, and in most cases, the difference was relatively small. Furthermore, Figure 6.2 compares the MOS scores per condition with 95% confidence intervals that were given by the listeners in the laboratory and those in the study E1.

6.2.2 Influence of Language Differences

To investigate the influence of language differences on the collected speech quality ratings, I conducted the studies E2 and E3 with native English and Spanish speakers, respectively. Both groups of listeners assessed the quality of the same German speech dataset evaluated by the participants in E1. Of course, the study setup remained unchanged.

[2] The AAC is an audio coding standard for lossy digital audio compression. It was created to be the MP3 format's successor since it achieves higher sound quality at the same bit rate.

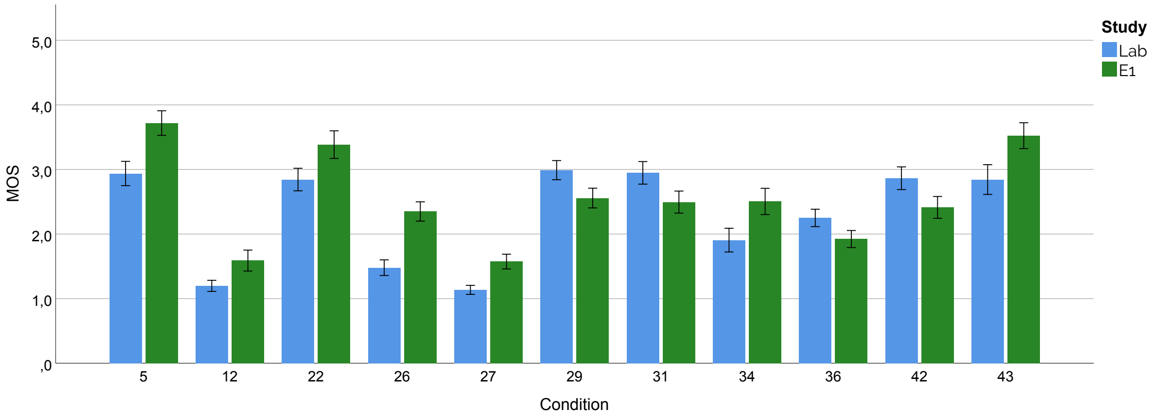


Figure 6.2: Comparison between the Lab-MOS and the E1-MOS with 95% confidence intervals. Represented only the conditions that were rated statistically significantly different. Information about the degradation conditions can be found in Appendix A and in [9].

I computed the Pearson's product-moment correlation and the RMSE to determine the relationship between the ratings collected in the study E1 and the scores collected in E2 and E3. The correlation coefficient was strong and significant, and the RMSE low in both studies with native English and Spanish speakers. These results indicate the validity of the ratings collected in E2 and E3. Table 6.4 summarizes these results.

Table 6.4: The table presents the Pearson's product-moment correlation, and the Root Mean Square Error (RMSE) between the E1-MOS and the crowdsourcing ratings collected in the study E2, and E3.

Study	# of valid ratings	# of listeners	r	$RMSE$
E2 (native English)	4025	53	0.970*	0.342
E3 (native Spanish)	3724	40	0.955*	0.339

$^*p < 0.001$

Additionally, I run a Mann-Whitney U test per condition with Šidák alpha correction to investigate the differences between the speech quality ratings provided by German listeners (E1) and those gathered with native English (E2) and Spanish (E3) listeners. This test revealed that the quality scores' medians were statistically significantly different in 12 of the speech degradation conditions under test in E2 (see Table 6.5) and for 10 conditions in the case of the study E3 (see Table 6.6).

The number of speech impairment conditions rated statistically significantly different between native Germans and native English, and between the Germans and Spanish speakers was low, considering that there were 50 conditions under test. These results confirm the validity of the speech quality scores collected in the study E2 and E3, which suggests that it is possible to reliably assess the quality of a German speech dataset with native English or Spanish speakers.

Figure 6.3 shows a scatterplot of the MOS scores of the 12 conditions rated statistically significantly different between the German and the English speakers. And, Figure 6.4 presents the scatterplot of the ten conditions that were perceived significantly differently by German and Spanish listeners. It can be seen in both graphs that English and Spanish speakers tended to overrate the quality of the speech stimuli of these impairments conditions. Since the listeners

Table 6.5: Conditions rated statistically significantly different between the German listeners (E1) and the native English (E2) speakers.

Cond. No.	E1	E1	E2				
	Median	Mean	Median	Mean	U	z	p-value
2	2.00	2.036	2.00	2.571	2361.5	-3.983	$< .001$
12	1.00	1.589	2.00	2.012	4043.0	-3.436	$= .001$
23	2.00	2.118	3.00	2.679	2663.5	-4.791	$< .001$
29	3.00	2.558	3.00	3.066	1985.5	-4.520	$< .001$
30	2.00	2.558	3.00	2.929	4065.5	-3.285	$= .001$
32	2.00	2.011	3.00	2.595	2469.5	-4.506	$< .001$
36	2.00	1.923	2.00	2.257	2828.0	-3.399	$= .001$
37	2.00	2.164	3.00	2.747	2396.5	-5.225	$< .001$
42	2.00	2.413	3.00	2.974	2260.5	-4.295	$< .001$
44	2.00	2.442	3.00	2.988	2994.0	-3.842	$< .001$
49	2.00	1.942	3.00	2.548	2859.0	-4.173	$< .001$
50	1.00	1.167	2.00	1.607	2584.0	-5.064	$< .001$

Table 6.6: Conditions rated statistically significantly different between the German listeners (E1) and the native Spanish (E3) speakers.

Cond. No.	E1	E1	E3				
	Median	Mean	Median	Mean	U	z	p-value
1	5.00	4.637	5.00	4.900	2789.5	-3.342	$= .001$
2	2.00	2.036	3.00	2.921	1326.5	-6.765	$< .001$
3	3.00	3.437	4.00	3.826	2499.0	-3.615	$< .001$
11	4.00	4.212	5.00	4.640	2749.0	-3.296	$= .001$
12	1.00	1.589	2.00	2.184	3436.0	-3.903	$< .001$
21	2.00	2.195	3.00	2.956	2273.0	-5.811	$< .001$
28	2.00	2.202	3.00	2.632	2486.5	-3.673	$< .001$
34	2.00	2.505	3.00	3.329	2244.5	-5.041	$< .001$
36	2.00	1.923	2.00	2.447	2412.0	-5.016	$< .001$
49	2.00	1.942	3.00	2.447	2764.5	-3.516	$< .001$

in these two groups were not familiar with the German language, it seems that they did not perceive some impairments in the speech due to the degradation conditions. That was the case of the stimuli from conditions 11 and 12, where the speech was clipped, or conditions 28, 29, and 30 where the speech sounded a bit electronic, or the slight interruptions in the stimuli of conditions 34, 42, and 44, or the robotic-sounding distortion in the samples from conditions 49. These impairments were difficult to perceive for the non-native ear, and thus listeners provided systematically higher quality scores to those speech stimuli.

The bias evidenced in Figures 6.3 and 6.4 can be corrected by applying a first-order mapping [90]. Figures 6.5 and 6.6 presents a scatterplot of the MOS values per condition between the native German (E1-MOS) and English (E2-MOS) and Spanish (E3-MOS) speakers, respectively, after applying the first-order mapping. Additionally, I computed the Pearson's product-moment correlation and RMSE between the E1-MOS and the corrected E2- and E3-MOS. These results are presented in Table 6.7. While the correlation coefficient did not improve significantly, the RMSE decreased to 0.30 and 0.23 in E2 and E3, respectively.

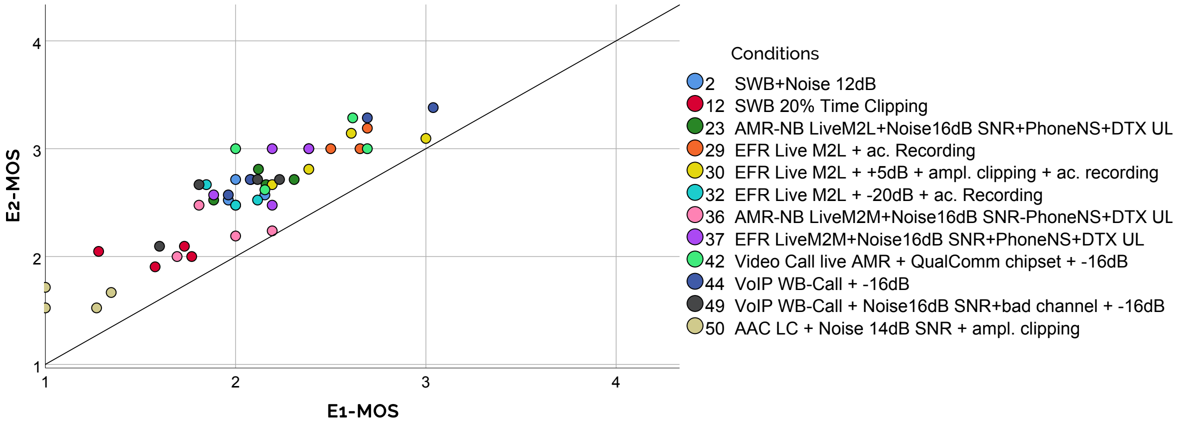


Figure 6.3: Scatterplot of the per-file-MOS for conditions rated statistically significantly different between the native German (E1-MOS) and the native English (E2-MOS) speakers. More details about the speech impairment conditions can be found in [9].

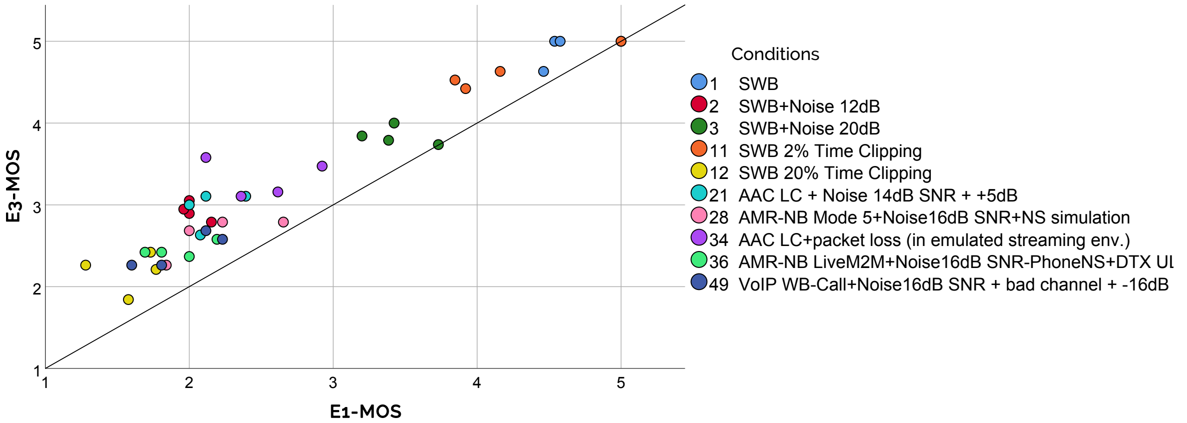


Figure 6.4: Scatterplot of the per-file-MOS for conditions rated statistically significantly different between the native German (E1-MOS) and the native Spanish (E3-MOS) speakers. More details about the speech impairment conditions can be found in [9].

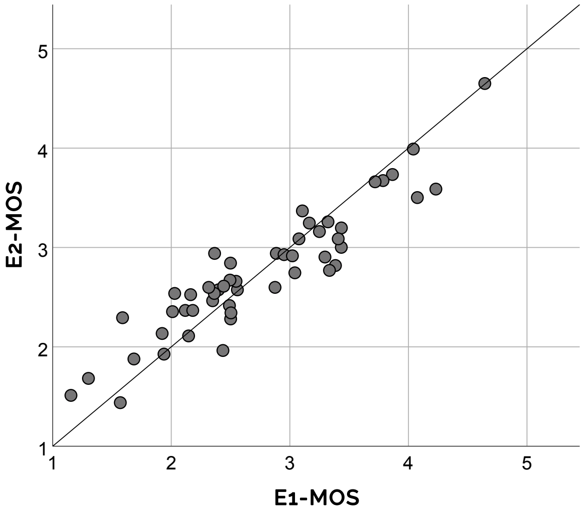


Figure 6.5: Scatterplot of the ratings per condition provided by the native German (E1-MOS) versus the scores provided by the native English (E2-MOS) speakers after applying the first-order mapping. All conditions are represented.

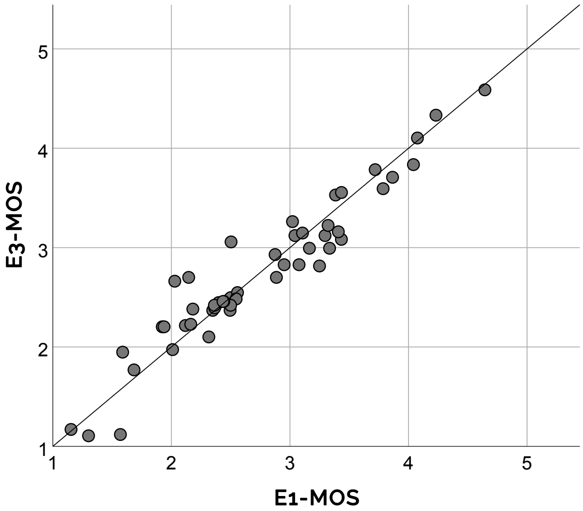


Figure 6.6: Scatterplot of the scores per condition provided by the native German (E1-MOS) versus the ratings provided by the native Spanish (E3-MOS) speakers after applying the first-order mapping. All conditions are represented.

Table 6.7: The table presents the Pearson's product-moment correlation, and the Root Mean Square Error (RMSE) between the E1-MOS and the ratings collected in E2, and E3, after applying a first-order mapping.

Study	r	$RMSE$
E2 (native English)	0.93*	0.30
E3 (native Spanish)	0.96*	0.23

*$p < 0.001$

6.2.3 Analysis of Conditions per Group

One of the main goals of a speech quality assessment test is to understand how users perceive certain speech degradation conditions and also small variations of these impairments. In the following, I analyze the ratings given to the degradation conditions and compare all conditions with each other to determine whether the same conclusion can be drawn in each study group.

I conducted a paired-samples t-test with Šidák alpha correction for multiple comparisons between all speech degradation conditions within each study group. Considering that there were 50 conditions under test, then there were 1176 t-tests executed in total. First, I compare the results between the laboratory and the native German listeners. And afterward, I contrast the results from the native English and Spanish speakers to the German crowd-workers.

Out of the 1176 t-tests, 986 yielded the same results with the German listeners as with the laboratory participants, representing 83.84%. This result also confirms the validity of conducting speech quality assessments with German listeners in crowdsourcing. The same conclusion was reached in most cases as if the test were performed in the laboratory.

Furthermore, when contrasting the results from the native English and Spanish speakers to the Germans, 1021 and 990 t-tests produced the same results, which represent 86.21% and 84.18%, respectively. These outcomes also demonstrate the validity of evaluating the quality of German speech stimuli with native English and Spanish listeners in crowdsourcing.

Moreover, I analyzed the t-test results of the comparisons involving similar degradation conditions. For example, conditions 2 and 3 that where SWB with two different levels of noise (i.e., 12dB and 20dB, respectively), or conditions 6 and 7 also SWB but with different attenuations (i.e., -10dB and -20dB, respectively). Then there were 48 of these comparisons. Forty-one of these t-tests produced the same outcome with the German listeners as with the laboratory participants, representing 85.42%. Also, the results of 45 and 41 t-tests were the same between the German listeners and the English and Spanish speakers, respectively. These findings reinforce our claim that reliable speech quality scores of German speech stimuli can be gathered with native English and Spanish speakers. The table below shows the executed comparisons between similar degradation conditions for all groups.

Table 6.8: The table presents the outcomes of the executed paired-samples t-tests between similar degradation conditions within each study group. Represented are the results for the laboratory participants and the native German, English, and Spanish speakers, respectively. Depicted with an "X" the comparisons deemed significant and in blank where no significance was seen.

Comparison of conditions	Laboratory Participants	German Listeners	English Listeners	Spanish Listeners
2 vs 3				
4 vs 5		X	X	
6 vs 7	X			X
9 vs 10	X		X	X
13 vs 14				
15 vs 16				
15 vs 17				
16 vs 17				
18 vs 28	X		X	
19 vs 20				
21 vs 22		X		
21 vs 50				
23 vs 24				
23 vs 25	X			
24 vs 25				
24 vs 36				
26 vs 27				X
26 vs 34				
27 vs 34				X
29 vs 30				
29 vs 31				
29 vs 32				X
29 vs 33				
31 vs 32				
31 vs 33				
32 vs 33				
33 vs 35				

Comparison of conditions	Laboratory Participants	German Listeners	English Listeners	Spanish Listeners
33 vs 37				
35 vs 37				
38 vs 39				
38 vs 40				
38 vs 41				
38 vs 42				
39 vs 40	X			
39 vs 41				
39 vs 42				
40 vs 41				
40 vs 42				
41 vs 42				
43 vs 44				
43 vs 45				
43 vs 46				
44 vs 45				
44 vs 46				
45 vs 46				
47 vs 48				
47 vs 49				
48 vs 49				

6.3 Conclusion

The studies presented in this chapter aim to assess whether it is possible to gather reliable speech quality scores for German stimuli with native English and Spanish speakers in a crowdsourcing environment. The reason being that there are not enough active German workers in the main crowdsourcing platform (i.e., Amazon Mechanical Turk (AMT) and microWorkers (MW)). An alternative is to use clickworker, which is a crowdsourcing platform based in Germany. However, they lack basic functionalities like audio-playback, quality control, or task repetition. Thus, it is mandatory to implement a system to carry listening tests in clickworker, which might be cumbersome. Therefore, the question of whether assessments of the speech quality of a German dataset could be made at AMT with native English and Spanish workers.

Three studies (i.e., E1, E2, and E3) were conducted with different listeners to assess the quality of the same German speech stimuli. E1 was carried out in the clickworker crowdsourcing platform with native German speakers. And studies E2 and E3 were executed in AMT with native English and Spanish speakers, respectively. First, results gathered in E1 were contrasted to ratings collected in a previous laboratory experiment with a panel of German listeners. Then, the MOS scores from the English and Spanish listeners were contrasted to the MOS scores given by the Germans participants in E1.

Overall, a high and significant Pearson correlation and low RMSE was achieved between the laboratory ratings and the scores collected in all crowdsourcing studies, despite the listeners' mother tongue (see Table 6.2).

Furthermore, a scatterplot of the mean opinion scores from the German crowd-workers and the native English and Spanish speakers revealed that the non-German participants tended to overestimate the quality of the speech stimuli. Then, this bias was corrected with a first-order mapping.

A similar effect was also seen in the speech quality scores gathered by the researchers of the study performed in [73]. In that experiment, non-native English listeners evaluated the quality of English speech stimuli, and the ratings collected from listeners with low English knowledge were biased. Likewise, native and non-native English listeners evaluated the quality of an American-English speech dataset in an experiment in [75]. The results showed that the non-native subjects rated the stimuli with additional noise lower than the native English participants. The trend observed in these studies was different from that perceived in my experiments with non-native English and Spanish listeners. I do not have an explanation for these differences, as further research would be necessary. However, what is important is that biases can be expected when performing the speech listening test with non-native speakers (relative to the speech dataset). Therefore, the bias perceived in the speech quality ratings provided by the non-German participants of our study is not surprising.

Additionally, an analysis was executed per group and per speech degradation condition to determine if the same conclusion regarding listener perception of the speech impairments could be reached in each study group in crowdsourcing. This analysis showed that more than 85% of the results obtained with native English and Spanish speakers were equivalent to the results gathered with German crowd-workers. These outcomes indicate the feasibility of conducting speech quality assessment studies of a German dataset with native English or Spanish speaking crowd-workers in Amazon Mechanical Turk.

7 Conclusion

The quality of transmitted speech is a vital indicator for telecommunication network providers used to evaluate their systems and services. Traditional laboratory methods to estimate speech quality are expensive and time-consuming. Therefore, the need for instrumental models to predict the overall quality of transmitted speech and alternative test methods has risen in recent years [9, 25].

Crowdsourcing is a valid approach for the rapid collection of speech quality evaluations. However, due to the remote nature of crowdsourcing studies and the lack of supervision of participants, multiple challenges arise that need to be addressed to collect valid and reliable outcomes. This thesis investigates the influence of different factors in the speech quality assessments executed in crowdsourcing. Specifically, I address important questions concerning the test structure, environment background noise conditions, and language mismatch between the speech dataset and listeners' pool.

Chapter 2 provides a review of related work that contributes to motivating the research carried out in this dissertation. Chapter 3 presents the speech material used in the studies and the test procedures fundamentals followed in the experiments.

Chapter 4 addresses the research questions regarding the test structure. Particularly, I investigate the optimum number of speech stimuli to include in a single task, and also the influence of performing the evaluation task multiple times, on different worker reliability metrics. The first research question was addressed by conducting a study in crowdsourcing with three non-overlapping groups. Workers were confronted with tasks consisting of a different number of stimuli, i.e., 10, 20, or 40. The speech quality ratings were highly correlated to previously collected laboratory scores despite the number of employed stimuli.

The highest correlation and lowest RMSE was achieved in the group assessing 40 speech samples per task. However, a decrease in the correlation coefficient was perceived in the second half of these tests. Also, a significant number of workers in this group reported being exhausted at the end of the task. Consequently, most of them participated in the study only once, and it was challenging to collect the desired number of votes per file compared to the other groups. Therefore, it is encouraging to offer tasks containing a reduced number of speech stimuli, i.e., between 10 and 15.

Additionally, the results indicate that crowd-workers perceived higher the quality of wideband (WB) and super-wideband (SWB) speech stimuli than the listeners in the laboratory. On the contrary, the overall quality of narrowband (NB) speech samples was perceived lower by crowdsourcing listeners than those in the laboratory. I hypothesize that these differences were due to the hardware employed to perform the listening test. Listeners in the laboratory used professional equipment, whereas crowd-workers employed regular headphones. Nevertheless, there is no clear explanation for these differences. Hence, further investigation would be needed to determine the reasons for the discrepancy in the quality evaluations for certain WB, SWB, and NB speech files.

The second research question was addressed by carrying out two studies, one in the laboratory and the other in crowdsourcing. In both, listeners had the chance to evaluate four times the quality of the speech stimuli in the dataset. The results showed that listeners as individuals were very consistent with their ratings. However, this consistency did not lead to an increase in the agreement of all listeners as a group. As a result, the correlation and RMSE to the laboratory results remained almost constant from the first to the last time workers conducted the listening test. Finally, a model is proposed to predict wokers' performance based on intra-rater reliability, root-mean-squared-deviation, and listeners' age. Such a model would be convenient for estimating the validity of the speech quality scores collected in crowdsourcing when there are no laboratory results to compare to.

It is worth noticing that the proposed model was built with data collected in a crowdsourcing experiment in which listeners evaluated the same set of stimuli several times. Such a study setup is relatively uncommon due to financial limitations and experimental design constraints in crowdsourcing.

Frequently, in speech quality assessment experiments, the database to be evaluated comprises multiple samples that are coded with very similar speech impairments. Then, a listener would provide quality scores to a condition by assessing multiple speech stimuli. To circumvent the limitations mentioned above and to be able to apply the proposed model to evaluate the listener performance, an idea is to consider the ratings per condition from a single worker. This way, an intra-rater reliability score can be computed that could be used as an input feature. Nonetheless, further empirical studies would need to be carried out to validate this assumption.

Chapter 5 answers questions related to the impact of environmental background noise in speech quality experiments carried out in crowdsourcing. To this end, a simulated crowdsourcing study was executed in the laboratory with three groups of participants. They evaluated the overall quality of speech files in the presence of environmental background noise at different levels.

The results indicate that the environmental background noise's threshold to collect reliable speech quality assessments in crowdsourcing lies between 43dB(A) and 50dB(A). A background noise level of 50dB(A) on average led to invalid results, whereas a 43dB(A) level yielded reliable speech quality ratings in most cases.

Additionally, listeners were more tolerant towards the TV-Show noise than to the street-traffic noise. Participants performing the test under the TV-Show noise condition provided quality scores that were more in line with the quiet test condition. Moreover, the results indicate that the presence of background environmental noise does not cause listeners to consistently give lower or higher quality scores in an entire speech quality assessment study. Instead, the effect of the background noise depends on the speech degradation condition in a test.

Furthermore, a Gradient Boosting Regressor model is proposed for correcting the bias found in the quality scores given to attenuated speech samples. Features were derived out of different statistics computed from the quality ratings and noise level. The regressor model achieved an R^2 score of 0.90 and an RMSE of 0.416.

The proposed model can be applied under the premise that information about workers' environmental background noise characteristics in crowdsourcing is available. An idea to fulfill this requirement is to include a step in the training phase of the speech quality assessment study, asking crowd-workers to assess the environmental background noise with one of the available sound meter apps for smartphones. However, further research would be needed to determine the accuracy of such sound meter apps and whether workers would be willing to install a mobile application as part of the study.

Another approach to infer information about the workers' environmental characteristics is to automatically record audio samples through the audio-web API while performing the speech quality listening test. Of course, this would only be possible after informing them about the recording and their subsequent consent. The collected recordings could then be used to make automatic predictions about the listeners' environment background noise characteristics.

However, it remains to be explored in future work whether good environmental recordings can be collected from the audio-web API from users in crowdsourcing. And also, if it is possible to use machine learning to derive information about the characteristics of the noise from those audio samples, considering that audio data collected through a web browser can be very diverse due to the diversity of computers, operating systems, and software versions.

Finally, Chapter 5 takes the first steps to address the question of using machine learning for noise classification from web-audio recordings. Two environmental background noise datasets were created, taking into account the noises affecting workers when they participate in crowdsourcing tasks. These datasets were used for testing the performance of different machine learning algorithms for noise classification and noise level estimation.

One experiment was carried out where multiple state-of-the-art machine learning classifiers were trained with a different number of MFCC coefficients for noise classification. The classifiers' accuracy increased with the number of MFCC coefficients and also increased the computational requirements. The highest accuracy was achieved with a Multi-layer Perceptron Classifier with an *adam* solver for weight optimization. It is recommended to employ a number of MFCC coefficients around 20 to accomplish a fair balance between accuracy and computational requirements.

Additionally, another study was conducted to verify the performance of a deep learning model based on a *"Long- Short-Term Memory"* (LSTM) architecture for noise level estimation. The proposed LSTM-based model achieved an RMSE of 4.58 and a standard deviation of 2.72 on the test dataset.

All in all, these outcomes confirm the validity of using web audio recordings to infer information about the environmental background noise characteristics. However, additional empirical studies would be needed to collect background audio data from crowdsourcing and further validate the proposed models.

Chapter 6 investigates whether non-native German listeners could provide reliable speech quality evaluations to a German speech dataset. Three studies were conducted in crowdsourcing with native German, English, and Spanish speakers. All participants evaluated the quality of the same German speech stimuli. The correlation to the laboratory results was strong and significant in all groups regardless of the listeners' mother tongue. Additionally, a scatterplot of the mean opinion scores revealed that the English and Spanish crowd-workers tended to overrate the quality of the speech files. However, a first-order mapping permitted to correct such a bias. Moreover, an analysis was made per group to determine if the same conclusion regarding listener perception of the speech impairments could be drawn in each study group. This analysis revealed that more than 85% of the results obtained with native English and Spanish speakers are equivalent to the results gathered with the German crowd-workers.

All in all, these outcomes indicate that it is possible to evaluate the quality of a German speech dataset with native English or Spanish speakers in crowdsourcing. Yet, a bias can be expected, but such deviations could be corrected with a first-order mapping.

A direction for future research is to investigate if reliable annotations of speech quality to a German dataset could be gathered with native Indian crowd-workers. Workers from India represent one of the most prominent user groups in the crowdsourcing platforms Amazon Mechanical Work and microWorkers [67]. Therefore, it is important to determine the validity of the speech quality evaluations from Indian workers. Addressing speech quality experiments to Indian users would reduce study costs and study turn-around time as they are very active in doing crowd-work. Furthermore, future research should address whether reliable speech quality annotations to a German dataset could be gathered with listeners from a language group different than German.

References

[1] Tobias Hoßfeld, Matthias Hirth, Judith Redi, Filippo Mazza, Pavel Korshunov, Babak Naderi, Michael Seufert, Bruno Gardlo, Sebastian Egger and Christian Keimel. *Best Practices and Recommendations for Crowdsourced QoE - Lessons learned from the Qualinet Task Force "Crowdsourcing"*. Oct. 2014. URL: https://hal.archives-ouvertes.fr/hal-01078761.

[2] Judith Redi, Ernestasia Siahaan, Pavel Korshunov, Julian Habigt and Tobias Hoßfeld. "When the Crowd Challenges the Lab: Lessons Learnt from Subjective Studies on Image Aesthetic Appeal". In: *Fourth International Workshop on Crowdsourcing for Multimedia*. CrowdMM '15 (2015), pp. 33–38. DOI: 10.1145/2810188.2810194.

[3] Sebastian Egger-Lampl, Judith Redi, Tobias Hoßfeld, Matthias Hirth, Sebastian Möller, Babak Naderi, Christian Keimel and Dietmar Saupe. "Crowdsourcing Quality of Experience Experiments". In: *Evaluation in the Crowd. Crowdsourcing and Human-Centered Experiments*. Ed. by Daniel Archambault, Helen Purchase, and Tobias Hoßfeld. Cham: Springer International Publishing, 2017, pp. 154–190. ISBN: 978-3-319-66435-4.

[4] Alexander Raake. *Speech Quality of VoIP: Assessment and Prediction*. John Wiley & Sons, 2007.

[5] Kjell Brunnström, Sergio Ariel Beker, Katrien De Moor, Ann Dooms, Sebastian Egger, Marie-Neige Garcia, Tobias Hossfeld, Satu Jumisko-Pyykkö, Christian Keimel, Mohamed-Chaker Larabi, Bob Lawlor, Patrick Le Callet, Sebastian Möller, Fernando Pereira, Manuela Pereira, Andrew Perkis, Jesenka Pibernik, Antonio Pinheiro, Alexander Raake, Peter Reichl, Ulrich Reiter, Raimund Schatz, Peter Schelkens, Lea Skorin-Kapov, Dominik Strohmeier, Christian Timmerer, Martin Varela, Ina Wechsung, Junyong You and Andrej Zgank. *Qualinet White Paper on Definitions of Quality of Experience*. 2013. URL: https://hal.archives-ouvertes.fr/hal-00977812.

[6] Alexander Raake and Sebastian Egger. "Quality and Quality of Experience". In: *Quality of Experience: Advanced Concepts, Applications and Methods*. Ed. by Sebastian Möller and Alexander Raake. Cham: Springer International Publishing, 2014, pp. 11–33. ISBN: 978-3-319-02681-7. DOI: 10.1007/978-3-319-02681-7_2.

[7] Ulrich Reiter, Kjell Brunnström, Katrien De Moor, Mohamed-Chaker Larabi, Manuela Pereira, Antonio Pinheiro, Junyong You and Andrej Zgank. "Factors Influencing Quality of Experience". In: *Quality of Experience: Advanced Concepts, Applications and Methods*. Ed. by Sebastian Möller and Alexander Raake. Cham: Springer International Publishing, 2014, pp. 55–72. ISBN: 978-3-319-02681-7. DOI: 10.1007/978-3-319-02681-7_4.

[8] Sebastian Möller, Wai-Yip. Chan, Nicolas Côté, Tiago H. Falk, Alexander Raake and Marcel Wältermann. "Speech Quality Estimation: Models and Trends". In: *IEEE Signal Processing Magazine* 28.6 (Nov. 2011), pp. 18–28. ISSN: 1053-5888. DOI: 10.1109/MSP.2011.942469.

[9] ITU-T Recommendation P.863. *Perceptual objective listening quality assessment.* Geneva: International Telecommunication Union, 2014.

[10] Jeff Howe. "The Rise of Crowdsourcing". In: *Wired Magazine* 14.6 (2006), pp. 1–4.

[11] Anhai Doan, Raghu Ramakrishnan, and Alon Y Halevy. "Crowdsourcing Systems on the World-Wide Web". In: *Communications of the ACM* 54.4 (2011), pp. 86–96.

[12] Flávio Ribeiro, Dinei Florencio, and Vítor Nascimento. "Crowdsourcing subjective image quality evaluation". In: *18th IEEE International Conference on Image Processing.* Sept. 2011, pp. 3097–3100. DOI: 10.1109/ICIP.2011.6116320.

[13] Ernestasia Siahaan, Alan Hanjalic, and Judith Redi. "A Reliable Methodology to Collect Ground Truth Data of Image Aesthetic Appeal". In: *IEEE Transactions on Multimedia* 18.7 (July 2016), pp. 1338–1350. ISSN: 1520-9210. DOI: 10.1109/TMM.2016.2559942.

[14] M. Shahid, J. Søgaard, J. Pokhrel, K. Brunnström, K. Wang, S. Tavakoli and N. Gracia. "Crowdsourcing based subjective quality assessment of adaptive video streaming". In: *Sixth International Workshop on Quality of Multimedia Experience (QoMEX).* Sept. 2014, pp. 53–54. DOI: 10.1109/QoMEX.2014.6982289.

[15] Jacob Søgaard, Muhammad Shahid, Jeevan Pokhrel and Kjell Brunnström. "On subjective quality assessment of adaptive video streaming via crowdsourcing and laboratory based experiments". In: *Multimedia Tools and Applications* (2016), pp. 1–22. ISSN: 1573-7721. DOI: 10.1007/s11042-016-3948-3. URL: http://dx.doi.org/10.1007/s11042-016-3948-3.

[16] Dietmar Saupe, Franz Hahn, Vlad Hosu, Igor Zingman, Masud Rana and Shujun Li. "Crowd workers proven useful: A comparative study of subjective video quality assessment". In: *8th International Conference on Quality of Multimedia Experience (QoMEX).* 2016.

[17] Steven Schmidt, Babak Naderi, Saeed Shafiee Sabet, Saman Zadtootaghaj and Sebastian Möller. "Assessing Interactive Gaming Quality of Experience using a Crowdsourcing Approach". In: *2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX).* QoMEX '20. IEEE, May 2020, pp. 1–6. ISBN: 978-1-7281-5965-2. DOI: 10.1109/QoMEX48832.2020.9123122.

[18] Jeanne Parson, Daniela Braga, Michael Tjalve and Jieun Oh. "Evaluating Voice Quality and Speech Synthesis Using Crowdsourcing". In: *Text, Speech, and Dialogue: 16th International Conference, TSD 2013, Pilsen, Czech Republic, September 1-5, 2013. Proceedings.* Ed. by Ivan Habernal and Václav Matoušek. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 233–240. ISBN: 978-3-642-40585-3. DOI: 10.1007/978-3-642-40585-3_30.

[19] Mark Cartwright, Bryan Pardo, Gautham J. Mysore and Matt Hoffman. "Fast and Easy Crowdsourced Perceptual Audio Evaluation". In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* Mar. 2016, pp. 619–623. DOI: 10.1109/ICASSP.2016.7471749.

[20] Tobias Hoßfeld, Christian Keimel, Matthias Hirth, Bruno Gardlo, Julian Habigt, Klaus Diepold and Phuoc Tran-Gia. "Best Practices for QoE Crowdtesting: QoE Assessment With Crowdsourcing". In: *IEEE Transactions on Multimedia* 16.2 (Feb. 2014), pp. 541–558. ISSN: 1520-9210. DOI: 10.1109/TMM.2013.2291663. URL: http://dx.doi.org/10.1109/TMM.2013.2291663.

[21] Daniel Archambault, Helen Purchase, and Tobias Hoßfeld. *Evaluation in the Crowd. Crowdsourcing and Human-Centered Experiments: Dagstuhl Seminar 15481, Dagstuhl Castle, Germany, November 22–27, 2015, Revised Contributions.* Vol. 10264. Springer, 2017.

[22] Flavio P Ribeiro, Dinei A F Florêncio, Cha Zhang and Michael L Seltzer. "CROWDMOS: An Approach for Crowdsourcing Mean Opinion Score Studies". In: *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* May 2011, pp. 2416–2419. DOI: 10.1109/ICASSP.2011.5946971.

[23] Tim Polzehl, Babak Naderi, Friedemann Köster and Sebastian Möller. "Robustness in speech quality assessment and temporal training expiry in mobile crowdsourcing environments". In: *INTERSPEECH.* 2015, pp. 2794–2798.

[24] Babak Naderi, Tim Polzehl, Ina Wechsung, Friedemann Köster and Sebastian Möller. "Effect of Trapping Questions on the Reliability of Speech Quality Judgments in a Crowdsourcing Paradigm". In: *Interspeech.* ISCA, 2015, pp. 2799–2803.

[25] ITU-T Recommendation P.808. *Subjective evaluation of speech quality with a crowdsourcing approach.* Geneva: International Telecommunication Union, 2018.

[26] ITU-T Technical Report PSTR-CROWDS. *Subjective evaluation of media quality using a crowdsourcing approach.* Geneva: International Telecommunication Union, 2019.

[27] Martin Cooke, Jon Barker, Maria Luisa Garcia Lecumberri and Krzysztof Wasilewski. "Crowdsourcing in Speech Perception". In: *Crowdsourcing for Speech Processing: Applications to Data Collection, Transcription and Assessment. Hoboken, NJ: John Wiley & Sons* (2013), pp. 137–172.

[28] Cas Smits, Theo S Kapteyn, and Tammo Houtgast. "Development and validation of an automatic speech-in-noise screening test by telephone". In: *International Journal of Audiology* 43.1 (2004), pp. 15–28. DOI: 10.1080/14992020400050004. URL: https://doi.org/10.1080/14992020400050004.

[29] M. Buschermöhle, K. C. Wagener, D. Berg, M. Meis and B. Kollmeier. "The German digit triplets test (Part II): validation and pass/fail criteria". In: *Zeitschrift für Audiologie* 54.1 (2015), pp. 6–13.

[30] Tobias Hoßfeld, Matthias Hirth, Pavel Korshunov, Philippe Hanhart, Bruno Gardlo, Christian Keimel and Christian Timmerer. "Survey of Web-based Crowdsourcing Frameworks for Subjective Quality Assessment". In: *16th International Workshop on Multimedia Signal Processing.* Jakarta, Indonesia, 2014.

[31] Babak Naderi. *Motivation of Workers on Microtask Crowdsourcing Platforms.* T-Labs Series in Telecommunication Services. Heiderlberg: Springer, 2018. ISBN: 978-3-319-72699-1. DOI: 10.1007/978-3-319-72700-4.

[32] Babak Naderi, Sebastian Möller, and Gabriel Mittag. "Speech Quality Assessment in Crowdsourcing: Influence of Environmental Noise". In: *44. Deutsche Jahrestagung für Akustik (DAGA).* Alte Jakobstraße 88, 10179 Berlin: Deutsche Gesellschaft für Akustik DEGA e.V., 2018, pp. 229–302. ISBN: 978-3-939296-13-3.

[33] Adrien Leman, Julien Faure, and Etienne Parizet. "Influence of informational content of background noise on speech quality evaluation for VoIP application". In: *Journal of the Acoustical Society of America* 123.5 (2008), p. 3066. DOI: 10.1121/1.2932822.

[34] Adrien Leman, Julien Faure, and Etienne Parizet. "A non-intrusive signal-based model for speech quality evaluation using automatic classification of background noises". In: *10th Annual Conference of the International Speech Communication Association (INTERSPEECH).* 2009, pp. 1139–1142.

[35] Rafael Zequeira Jiménez, Laura Fernández Gallardo, and Sebastian Möller. "Outliers Detection vs. Control Questions to Ensure Reliable Results in Crowdsourcing. A Speech Quality Assessment Case Study". In: *Companion Proceedings of the The Web Conference 2018.* WWW '18. Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee, 2018, pp. 1127–1130. ISBN: 978-1-4503-5640-4. DOI: 10.1145/3184558.3191545. URL: https://doi.org/10.1145/3184558.3191545.

[36] Rafael Zequeira Jiménez, Laura Fernández Gallardo, and Sebastian Möller. "Influence of Number of Stimuli for Subjective Speech Quality Assessment in Crowdsourcing". In: *2018 Tenth International Conference on Quality of Multimedia Experience (QoMEX).* May 2018, pp. 1–6. DOI: 10.1109/QoMEX.2018.8463298.

[37] Rafael Zequeira Jiménez, Laura Fernández Gallardo, and Sebastian Möller. "Environmental Noise Recording as a Quality Control for Crowdsourcing Speech Quality Assessments". In: *44. Deutsche Jahrestagung für Akustik (DAGA).* Deutsche Gesellschaft für Akustik DEGA e.V., Mar. 2018, pp. 303–306. ISBN: 978-3-939296-13-3.

[38] Rafael Zequeira Jiménez, Gabriel Mittag, and Sebastian Möller. "Effect of Number of Stimuli on Users Perception of Different Speech Degradations. A Crowdsourcing Case Study". In: *2018 IEEE International Symposium on Multimedia (ISM).* 2018, pp. 175–179. DOI: 10.1109/ISM.2018.00-16.

[39] Rafael Zequeira Jiménez, Babak Naderi, and Sebastian Möller. "Evaluating Acoustic Features from Environmental Audio Recordings via Web. A Crowdsourcing Survey on Background Noise Characteristics". In: *45. Deutsche Jahrestagung für Akustik (DAGA*

2019). Deutsche Gesellschaft für Akustik DEGA e.V., Mar. 2019, pp. 1190–1193. ISBN: 978-3-939296-14-0.

[40] Rafael Zequeira Jiménez, Babak Naderi, and Sebastian Möller. "Background Environment Characteristics of Crowd-Workers from German Speaking Countries Experimental Survey on User Environment Characteristics". In: *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX).* 2019, pp. 1–3. DOI: 10.1109/QoMEX.2019.8743208.

[41] Rafael Zequeira Jiménez, Babak Naderi, and Sebastian Möller. "Effect of Environmental Noise in Speech Quality Assessment Studies using Crowdsourcing". In: *2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX).* May 2020, pp. 1–6. DOI: 10.1109/QoMEX48832.2020.9123144.

[42] Rafael Zequeira Jiménez, Babak Naderi, and Sebastian Möller. "Effect of Environment in Speech Quality Assessment in Crowdsourcing". In: *Proceedings of Forum Acusticum.* European Acoustics Association. 2020.

[43] Rafael Zequeira Jiménez, Sebastian Möller, and Gabriel Mittag. "Removing the Bias in Speech Quality Scores Collected in Noisy Crowdsourcing Environments". In: *submitted to: 13th International Conference on Quality of Multimedia Experience (QoMEX).* 2021.

[44] Rafael Zequeira Jiménez, Babak Naderi, and Sebastian Möller. "Influence of Language in Speech Quality Studies in Crowdsourcing". In: *submitted to: 13th International Conference on Quality of Multimedia Experience (QoMEX).* 2021.

[45] Rafael Zequeira Jiménez, Anna Llagostera, Babak Naderi, Sebastian Möller and Jens Berger. "Modeling Worker Performance Based on Intra-rater Reliability in Crowdsourcing : A Case Study of Speech Quality Assessment". In: *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX).* 2019, pp. 1–6. DOI: 10.1109/QoMEX.2019.8743148.

[46] Rafael Zequeira Jiménez, Anna Llagostera, Babak Naderi, Sebastian Möller and Jens Berger. "Intra- and Inter-rater Agreement in a Subjective Speech Quality Assessment Task in Crowdsourcing". In: *Companion Proceedings of The 2019 World Wide Web Conference.* WWW '19. New York, NY, USA: ACM, 2019, pp. 1138–1143. ISBN: 978-1-4503-6675-5. DOI: 10.1145/3308560.3317084. URL: http://doi.acm.org/10.1145/3308560.3317084.

[47] Rafael Zequeira Jiménez and Sebastian Möller. *Investigating the Influence of Number of Stimuli in Speech Quality Assessments in Crowdsourcing.* ITU-T Contribution SG12-C.290. CH-Geneva: International Telecommunication Union, Nov. 2018, pp. 1–8.

[48] Babak Naderi, Sebastian Möller, and Rafael Zequeira Jiménez. *Evaluation of the Draft of P.CROWD Recommendation.* ITU-T Contribution SG12-C.290. CH-Geneva: International Telecommunication Union, Nov. 2018, pp. 1–8.

[49] Rafael Zequeira Jiménez, Babak Naderi, and Sebastian Möller. *Influence of environmental background noise on Speech Quality Assessment in a simulated Crowdsourcing scenario.* ITU-T Contribution SG12-C.0425. CH-Geneva: International Telecommunication Union, Nov. 2019, pp. 1–9.

[50] Tobias Hoßfeld and Christian Keimel. “Crowdsourcing in QoE Evaluation”. In: *Quality of Experience*. Springer, 2014, pp. 315–327.

[51] Maria K Wolters, Karl B Isaac, and Steve Renals. “Evaluating Speech Synthesis Intelligibility using Amazon Mechanical Turk”. In: *7th Speech Synthesis Workshop (SSW7)*. 2010, pp. 136–141.

[52] Martin Cooke, Jon Barker, Maria Luisa Garcia Lecumberri and Krzysztof Wasilewski. “Crowdsourcing for word recognition in noise”. In: *Interspeech*. 2011, pp. 3049–3052.

[53] Catherine Mayo, Vincent Aubanel, and Martin Cooke. “Effect of prosodic changes on speech intelligibility”. In: *Interspeech*. 2012, pp. 1708–1711.

[54] Benjamin Munson. “Assessing the Utility of Judgments of Children’s Speech Production Made by Untrained Listeners in Uncontrolled Listening Environments”. In: *Interspeech*. 2013, pp. 2147–2151.

[55] Tara McAllister Byun, Peter F Halpin, and Daniel Szeredi. “Online crowdsourcing for efficient rating of speech: A validation study”. In: *Journal of Communication Disorders* 53 (2015), pp. 70–83. ISSN: 0021-9924. DOI: 10.1016/j.jcomdis.2014.11.003.

[56] Joseph Slote and Julia F Strand. “Conducting spoken word recognition research online: Validation and a new timing method”. In: *Behavior Research Methods* 48.2 (2016), pp. 553–566. DOI: 10.3758/s13428-015-0599-7. URL: http://dx.doi.org/10.3758/s13428-015-0599-7.

[57] ITU-R Recommendation BS.1534-3. *Method for the subjective assessment of intermediate quality level of audio systems*. Geneva: International Telecommunication Union, 2014.

[58] John Le, Andy Edmonds, Vaughn Hester and Lukas Biewald. “Ensuring quality in crowdsourced search relevance evaluation: The effects of training question distribution”. In: *SIGIR 2010 Workshop on Crowdsourcing for Search Evaluation (CES 2010)*. Vol. 2126. 2010.

[59] Ujwal Gadiraju, Besnik Fetahu, and Ricardo Kawase. “Training Workers for Improving Performance in Crowdsourcing Microtasks”. In: *Design for Teaching and Learning in a Networked World*. Ed. by Gráinne Conole et al. Cham: Springer International Publishing, 2015, pp. 100–114. ISBN: 978-3-319-24258-3.

[60] Tobias Hoßfeld, Raimund Schatz, and Sebastian Egger. “SOS: The MOS is not Enough!” In: *Third International Workshop on Quality of Multimedia Experience (QoMEX)*. 2011, pp. 131–136. DOI: 10.1109/QoMEX.2011.6065690.

[61] Jan Holub, Hakob Avetisyan, and Scott Isabelle. “Subjective speech quality measurement repeatability: comparison of laboratory test results”. In: *International Journal of Speech Technology* 20.1 (2017), pp. 69–74.

[62] ITU-T Recommendation P.835. *Subjective Test Methodology for Evaluating Speech Communication Systems that Include Noise Supression Algorithm*. Geneva: International Telecommunication Union, 2003.

[63] Raimund Schatz, Sebastian Egger, and Kathrin Masuch. "The Impact of Test Duration on User Fatigue and Reliability of Subjective Quality Ratings". In: *Journal of the Audio Engineering Society* 60.1/2 (2012), pp. 63–73. URL: http://www.aes.org/e-lib/browse.cfm?elib=16167.

[64] Umair ul Hassan and Edward Curry. "A Capability Requirements Approach for Predicting Worker Performance in Crowdsourcing". In: *9th IEEE International Conference on Collaborative Computing: Networking, Applications and Worksharing.* Oct. 2013, pp. 429–437. DOI: 10.4108/icst.collaboratecom.2013.254181.

[65] Jakob Rogstadius, Vassilis Kostakos, Aniket Kittur, Boris Smus, Jim Laredo and Maja Vukovic. "An Assessment of Intrinsic and Extrinsic Motivation on Task Performance in Crowdsourcing Markets". In: *Proceedings of the Fifth International Conference on Weblogs and Social Media.* 2011. URL: http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2778.

[66] ITU-T Recommendation P.800. *Methods for subjective determination of transmission quality.* Geneva: International Telecommunication Union, 1996.

[67] David Martin, Sheelagh Carpendale, Neha Gupta, Tobias Hoßfeld, Babak Naderi, Judith Redi, Ernestasia Siahaan and Ina Wechsung. "Understanding the Crowd: Ethical and Practical Matters in the Academic Use of Crowdsourcing". In: *Evaluation in the Crowd. Crowdsourcing and Human-Centered Experiments.* Ed. by Daniel Archambault, Helen Purchase, and Tobias Hoßfeld. Cham: Springer International Publishing, 2017, pp. 27–69. ISBN: 978-3-319-66435-4.

[68] ITU-T Recommendation P.Sup23. *ITU-T coded-speech database.* Geneva: International Telecommunication Union, 2004.

[69] Julian Marscheider, Michał Sołoducha, and Janto Skowronek. "Environmental noise classification in context of speech transmission quality". In: *7th Forum Acusticum.* 2014.

[70] Julien Meyer, Laure Dentel, and Fanny Meunier. "Speech Recognition in Natural Background Noise". In: *PLOS ONE* 8.11 (2013), pp. 1–14. DOI: 10.1371/journal.pone.0079279. URL: https://doi.org/10.1371/journal.pone.0079279.

[71] Emma Jokinen, Jérémie Lecomte, Nadja Schinkel-Bielefeld and Tom Bäckström. "Intelligibility evaluation of speech coding standards in severe background noise and packet loss conditions". In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* 2015, pp. 5152–5156. DOI: 10.1109/ICASSP.2015.7178953. URL: https://doi.org/10.1109/ICASSP.2015.7178953.

[72] P Bazilinskyy and J C F De Winter. "Analyzing crowdsourced ratings of speech-based take-over requests for automated driving". In: *Applied Ergonomics* 64 (2017), pp. 56–64. ISSN: 0003-6870. DOI: https://doi.org/10.1016/j.apergo.2017.05.001. URL: http://www.sciencedirect.com/science/article/pii/S0003687017301047.

[73] Lubica Blaskova and Jan Holub. "How do non-native listeners perceive quality of transmitted voice?" In: *Communications-Scientific letters of the University of Zilina* 10.4 (2008), pp. 11–14.

[74] Alexander Raake. “Does the Content of Speech Influence its Perceived Sound Quality?” In: *Sign* 1 (2002), p. 2.

[75] Deborah U. Ebem, John G. Beerends, Jereon Van Vugt, Christian Schmidmer, Robert E. Kooij and Joy O. Uguru. “The Impact of Tone Language and Non-Native Language Listening on Measuring Speech Quality”. In: *Journal of the Audio Engineering Society* 59.9 (2011), pp. 647–655.

[76] Nadja Schinkel-Bielefeld, Zhang Jiandong, Qin Yili, Anna Katharina Leschanowsky and Fu Shanshan. “Is it Harder to Perceive Coding Artifact in Foreign Language Items?–A Study with Mandarin Chinese and German Speaking Listeners”. In: *Audio Engineering Society Convention 142*. Audio Engineering Society. 2017.

[77] ITU-T Recommendation P.501. *Test signals for use in telephonometry.* Geneva: International Telecommunication Union, 2017.

[78] Jens Berger and Anna Llagostera. *A subjective ACR LOT testing super-wideband speech coding in real field measurements and prediction by P.863.* ITU-T Contribution SG12-C.286. CH-Geneva: International Telecommunication Union, 2018, pp. 1–11.

[79] Bruno Gardlo, Sebastian Egger, and Tobias Hoßfeld. “Do Scale-Design and Training Matter for Video QoE Assessments Through Crowdsourcing?” In: *Proceedings of the Fourth International Workshop on Crowdsourcing for Multimedia.* CrowdMM ’15. New York, NY, USA: ACM, 2015, pp. 15–20. ISBN: 978-1-4503-3746-5. DOI: 10.1145/2810188.2810193.

[80] ETSI E G 202 396-1. *Speech Processing, Transmission and Quality Aspects (STQ); Speech quality performance in the presence of background noise; Part 1: Background noise simulation technique and background noise database.* Sophia-Antipolis, France: European Telecommunications Standards Institute, 2011.

[81] David C Hoaglin and Boris Iglewicz. “Fine-tuning some resistant rules for outlier labeling”. In: *Journal of the American Statistical Association* 82.400 (1987), pp. 1147–1149.

[82] Barbara G Tabachnick and Linda S Fidell. *Using Multivariate Statistics.* Allyn & Bacon/Pearson Education, 2007.

[83] William H Kruskal and W Allen Wallis. “Use of Ranks in One-Criterion Variance Analysis”. In: *Journal of the American Statistical Association* 47.260 (1952), pp. 583–621. ISSN: 01621459. URL: http://www.jstor.org/stable/2280779.

[84] Olive Jean Dunn. “Multiple Comparisons Using Rank Sums”. In: *Technometrics* 6.3 (1964), pp. 241–252. DOI: 10.1080/00401706.1964.10490181. URL: http://www.tandfonline.com/doi/abs/10.1080/00401706.1964.10490181.

[85] Yosef Hochberg. “A sharper Bonferroni procedure for multiple tests of significance”. In: *Biometrika* 75.4 (1988), pp. 800–802. DOI: 10.1093/biomet/75.4.800. URL: http://dx.doi.org/10.1093/biomet/75.4.800.

[86] 3GPP T S 26.441. *Codec for Enhanced Voice Services (EVS); General overview.*

[87] 3GPP T S 26.070. *Mandatory speech CODEC speech processing functions; AMR speech Codec; General description.*

[88] 3GPP T S 26.171. *Speech codec speech processing functions; Adaptive Multi-Rate - Wideband (AMR-WB) speech codec; General description.*

[89] Maurice George Kendall. *Rank Correlation Methods.* 4th. Charles Griffin, 1970.

[90] ITU-T Recommendation P.1401. *Methods, metrics and procedures for statistical evaluation, qualification and comparison of objective quality prediction models.* Geneva: International Telecommunication Union, 2012.

[91] Wayne W Daniel. *Applied nonparametric statistics (2nd ed.)* Boston, MA: Cengage Learning, 1980.

[92] Leonard A Marascuilo and Maryellen McSweeney. *Nonparametric and distribution-free methods for the social sciences.* Belmont, CA: Wadsworth Publishing Company, 1977.

[93] Kevin P Weinfurt. *Repeated measures analysis: ANOVA, MANOVA, and HLM. In L. G. Grimm & P. R. Yarnold (Eds.), Reading and understanding MORE multivariate statistics*, pp. 317–361.

[94] Scott E Maxwell. "Pairwise Multiple Comparisons in Repeated Measures Designs". In: *Journal of Educational Statistics* 5.3 (1980), pp. 269–287. DOI: 10 . 3102 / 10769986005003269. URL: https://doi.org/10.3102/10769986005003269.

[95] Patrick E Shrout and Joseph L Fleiss. "Intraclass Correlations: Uses in Assessing Rater Reliability". In: *Psychological Bulletin* 86.2 (1979), p. 420.

[96] Kevin A. Hallgren. "Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial". In: *Tutorials in quantitative methods for psychology* 8.1 (2012), p. 23.

[97] Richard Landers. "Computing Intraclass Correlations (ICC) as Estimates of Interrater Reliability in SPSS". In: *The Winnower* (2015). DOI: 10.15200/winn.143518.81744. URL: https://dx.doi.org/10.15200/winn.143518.81744.

[98] Isabel Cristina Bolaños Villalobos, Gabriela Cerdas Ramírez, and Jimmy Ramírez Acosta. "Intra-rater Reliability and the Role of Experience: A Comparative Case". In: *Revista de Lenguas Modernas* 20 (2014).

[99] R Dennis Cook and Sanford Weisberg. *Residuals and influence in regression.* New York: Chapman and Hall, 1982.

[100] Joseph F. Hair, William C. Black, Barry J. Babin and Rolph E. Anderson. *Multivariate Data Analysis (7th Edition).* 7th ed. Prentice Hall, 2009. ISBN: 0138132631.

[101] H B Mann and D R Whitney. "On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other". In: *The Annals of Mathematical Statistics* 18.1 (1947), pp. 50–60. ISSN: 00034851. URL: http://www.jstor.org/stable/2236101.

[102] Leo Breiman. "Bagging Predictors". In: *Machine Learning* 24.2 (1996), pp. 123–140.

[103] Leo Breiman. "Random Forests". In: *Machine Learning* 45.1 (2001), pp. 5–32.

[104] Pierre Geurts, Damien Ernst, and Louis Wehenkel. "Extremely Randomized Trees". In: *Machine Learning* 63.1 (2006), pp. 3–42.

[105] Harris Drucker. "Improving Regressors using Boosting Techniques". In: *ICML*. Vol. 97. 1997, pp. 107–115.

[106] Jerome H Friedman. “Greedy Function Approximation: A Gradient Boosting Machine”. In: *Annals of Statistics* (2001), pp. 1189–1232.

[107] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot and Édouard Duchesnay. “Scikit-learn: Machine Learning in Python”. In: *Journal of Machine Learning Research* 12.85 (2011), pp. 2825–2830. URL: http://jmlr.org/papers/v12/pedregosa11a.html.

[108] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, W. Philip Kegelmeyer. “SMOTE: Synthetic Minority Over-sampling Technique”. In: *Journal of Artificial Intelligence Research* 16 (2002), pp. 321–357.

[109] Guillaume Lemaître, Fernando Nogueira, and Christos K Aridas. “Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning”. In: *Journal of Machine Learning Research* 18.17 (2017), pp. 1–5. URL: http://jmlr.org/papers/v18/16-365.

[110] Scikit-Learn. *Gradient Boosting Regressor*. URL: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingRegressor.html.

[111] Rafael Zequeira Jiménez. *DNC: Dataset for Noise Classification*. 2021. DOI: 10.14279/depositonce-11645. URL: http://dx.doi.org/10.14279/depositonce-11645.

[112] Rafael Zequeira Jiménez. *DNLE: Dataset for Noise Level Estimation*. 2021. DOI: 10.14279/depositonce-11588. URL: http://dx.doi.org/10.14279/depositonce-11588.

[113] George Tzanetakis and Perry Cook. “Musical Genre Classification of Audio Signals”. In: *IEEE Transactions on Speech and Audio Processing* 10.5 (2002), pp. 293–302. DOI: 10.1109/TSA.2002.800560.

[114] Karol J Piczak. “ESC: Dataset for Environmental Sound Classification”. In: *Proceedings of the 23rd ACM International Conference on Multimedia*. MM ’15. New York, NY, USA: ACM, 2015, pp. 1015–1018. ISBN: 978-1-4503-3459-4. DOI: 10.1145/2733373.2806390. URL: http://doi.acm.org/10.1145/2733373.2806390.

[115] Karol J Piczak. “Environmental sound classification with convolutional neural networks”. In: *2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP)*. Sept. 2015, pp. 1–6. DOI: 10.1109/MLSP.2015.7324337.

[116] Romain Serizel, Victor Bisot, Slim Essid and Gaël Richard. “Acoustic Features for Environmental Sound Analysis”. In: *Computational Analysis of Sound Scenes and Events*. Ed. by Tuomas Virtanen, Mark D Plumbley, and Dan Ellis. Cham: Springer International Publishing, 2018, pp. 71–101. ISBN: 978-3-319-63450-0. DOI: 10.1007/978-3-319-63450-0_4.

[117] Brian McFee, Colin Raffel, Dawen Liang, Daniel P. W. Ellis, Matt McVicar, Eric Battenberg and Oriol Nieto. “librosa: Audio and Music Signal Analysis in Python”. In: *Proceedings of the 14th Python in Science Conference*. Vol. 8. 2015, pp. 18–25.

[118] Jacob Goldberger, Sam Roweis, Geoffrey Hinton and Ruslan Salakhutdinov. “Neighbourhood Components Analysis”. In: *Advances in Neural Information Processing Systems* 17 (May 2004), pp. 513–520.

[119] Ji Zhu, Hui Zou, Trevor Hastie and Saharon Rosset. “Multi-class AdaBoost”. In: *Statistics and its Interface* 2.3 (2009), pp. 349–360.

[120] Geoffrey E Hinton. “Connectionist learning procedures. Artificial Intelligence”. In: *Machine Learning: Paradigms and Methods, MIT Press* (1989), pp. 185–234.

[121] Dharmesh M. Agrawal, Hardik B. Sailor, Meet H. Soni and Hemant A. Patil. “Novel TEO-based Gammatone Features for Environmental Sound Classification”. In: *2017 25th European Signal Processing Conference (EUSIPCO)*. 2017, pp. 1809–1813. DOI: 10.23919/EUSIPCO.2017.8081521.

[122] Anurag Kumar and Bhiksha Raj. “Features and Kernels for Audio Event Recognition”. In: *arXiv preprint arXiv:1607.05765* (2016).

[123] Anssi Klapuri and Manuel Davy. “Signal Processing Methods for Music Transcription”. In: (2007).

[124] Daniel Ellis. “Chroma feature analysis and synthesis”. In: *Resources of Laboratory for the Recognition and Organization of Speech and Audio-LabROSA* (2007).

[125] Meinard Müller and Sebastian Ewert. “Chroma Toolbox: MATLAB implementations for extracting variants of chroma-based audio features”. In: *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR), 2011. hal-00727791, version 2-22 Oct 2012*. Citeseer. 2011.

[126] Sepp Hochreiter and Jürgen Schmidhuber. “Long Short-Term Memory”. In: *Neural Computation* 9.8 (1997), pp. 1735–1780. DOI: 10.1162/neco.1997.9.8.1735. URL: https://doi.org/10.1162/neco.1997.9.8.1735.

[127] Benjamin Cauchi, Kai Siedenburg, João F. Santos, Tiago H. Falk, Simon Doclo and Stefan Goetze. “Non-Intrusive Speech Quality Prediction Using Modulation Energies and LSTM-Network”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 27.7 (2019), pp. 1151–1163. ISSN: 2329-9304. DOI: 10.1109/TASLP.2019.2912123.

[128] Szu-wei Fu, Yu Tsao, Hsin-Te Hwang and Hsin-Min Wang. “Quality-Net: An End-to-End Non-intrusive Speech Quality Assessment Model Based on BLSTM”. In: *Proc. Interspeech 2018*. 2018, pp. 1873–1877. DOI: 10.21437/Interspeech.2018-1802. URL: http://dx.doi.org/10.21437/Interspeech.2018-1802.

[129] Mike Schuster and Kuldip K Paliwal. “Bidirectional Recurrent Neural Networks”. In: *IEEE Transactions on Signal Processing* 45.11 (Nov. 1997), pp. 2673–2681. ISSN: 1941-0476. DOI: 10.1109/78.650093.

[130] Alex Graves and Jürgen Schmidhuber. “Framewise phoneme classification with bidirectional LSTM and other neural network architectures”. In: *Neural networks* 18.5-6 (2005), pp. 602–610.

[131] Alex Graves and Navdeep Jaitly. “Towards end-to-end speech recognition with recurrent neural networks”. In: *International conference on machine learning*. 2014, pp. 1764–1772.

[132] Grégoire Mesnil, Xiaodong He, Li Deng and Yoshua Bengio. "Investigation of recurrent-neural-network architectures and learning methods for spoken language understanding." In: *Interspeech*. 2013, pp. 3771–3775.

[133] Sebastian Böck and Markus Schedl. "Enhanced Beat Tracking with Context-Aware Neural Networks". In: *Proc. Int. Conf. Digital Audio Effects*. 2011, pp. 135–139.

[134] Giambattista Parascandolo, Heikki Huttunen, and Tuomas Virtanen. "Recurrent Neural Networks for Polyphonic Sound Event Detection in Real Life Recordings". In: *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Mar. 2016, pp. 6440–6444. DOI: 10.1109/ICASSP.2016.7472917.

[135] Zbyněk Šidák. "Rectangular Confidence Regions for the Means of Multivariate Normal Distributions". In: *Journal of the American Statistical Association* 62.318 (1967), pp. 626–633. DOI: 10.1080/01621459.1967.10482935. URL: https://doi.org/10.1080/01621459.1967.10482935.

[136] J Sinclair, Paul J Taylor, and Sarah Jane Hobbs. "Alpha Level Adjustments for Multiple Dependent Variable Analyses and Their Applicability – A Review". In: *International Journal of Sports Science and Engineering* 7.1 (2013), pp. 17–20.

[137] D A Williams. "A Test for Differences between Treatment Means When Several Dose Levels are Compared with a Zero Dose Control". In: *Biometrics* 27.1 (1971), pp. 103–117. ISSN: 0006341X, 15410420. URL: http://www.jstor.org/stable/2528930.

A

Appendix A

A.1 Speech Database SwissQual 501

Table A.1: The table presents information about the 50 degradation conditions of the SwissQual 501 speech database. The Mean Opinion Scores (MOS) per condition is also displayed. More details can be found in [9].

Condition Number	Degradation Description	MOS per Condition
1	SWB	4.67
2	SWB+Noise 12dB	2.25
3	SWB+Noise 20dB	3.36
4	SWB+MNRU 10dB	1.23
5	SWB+MNRU 25dB	2.94
6	SWB Level -10dB	4.17
7	SWB Level -20dB	3.10
8	SWB mIRSsend+IRSrcv	3.79
9	SWB 500-2500Hz	2.57
10	SWB 100-5000Hz	3.89
11	SWB 2% Time Clipping	3.86
12	SWB 20% Time Clipping	1.20
13	AMR-WB Mode 0 (6.6 kbps)	2.94
14	AMR-WB Mode 2 (12.65 kbps)	3.95
15	AMR-WB Mode 0 (6.6 kbps) + Noise 16dB SNR	2.30
16	AMR-WB Mode 2 (12.65 kbps) + Noise 16dB SNR	2.84
17	AMR-WB Mode 2 (12.65 kbps) + Noise 16dB SNR +- 16dB	2.49
18	AMR-NB Mode 5 (7.95 kbps)	3.40
19	EVRC-B OP 0	3.57
20	EVRC-WB OP 3	3.61

Condition Number	Degradation Description	MOS per Condition
21	AAC LC + Noise 14dB SNR + +5dB	2.05
22	AAC LC + -10dB	2.84
23	AMR-NB Live M2L + Noise 16dB SNR + Phone NS + DTX UL + Nokia chipset	2.31
24	AMR-NB Live M2M + Noise 16dB SNR - Phone NS + DTX UL + bad channel + QualComm chipset	1.85
25	AMR-NB Live L2M + Bad channel + No DTX DL + Nokia chipset	3.58
26	AAC LC low bitrate (WB)	1.48
27	4 x AAC LC low bitrate (WB)	1.14
28	AMR-NB Mode 5 (7.95 kbps) + Noise 16dB SNR + NS simulation + AMR-NB Mode 5(7.95 kbps)	2.28
29	EFR Live M2L + ac. Recording	2.99
30	EFR Live M2L + +5dB + ampl. clipping + ac. recording	2.78
31	EFR Live M2L + -10dB + ac. Recording	2.95
32	EFR Live M2L + -20dB + ac. Recording	2.31
33	EFR Live M2M + Noise 16dB SNR + Phone NS + DTX UL + Nokia chipset	2.28
34	AAC LC + packet loss (in emulated streaming environment)	1.91
35	EFR Live M2M + bad channel	2.23
36	AMR-NB Live M2M + Noise 16dB SNR - Phone NS + DTX UL + QC chipset	2.25
37	EFR Live M2M + bad channel + Noise 16dB SNR + Phone NS + DTX UL + Nokia chipset	2.17
38	Video Call live AMR + QualComm chipset	3.05
39	Video Call live AMR + Nokia chipset	2.78
40	Video Call live AMR + QualComm chipset + +5dB	3.31
41	Video Call live AMR + QualComm chipset + -8dB	3.16
42	Video Call live AMR + QualComm chipset + -16dB	2.86
43	VoIP WB-Call	2.84
44	VoIP WB-Call + -16dB	2.34
45	VoIP WB-Call + -8dB	2.88
46	VoIP WB-Call + +5dB	2.73
47	VoIP WB-Call + Noise 16dB SNR + bad channel + +5dB	1.93
48	VoIP WB-Call + Noise 16dB SNR + bad channel + -8dB	2.20
49	VoIP WB-Call + Noise 16dB SNR + bad channel + -16dB	1.64
50	AAC LC + Noise 14dB SNR + ampl. clipping	1.19

B

Appendix B

B.1 Speech Database SwissQual 502

Table B.1: The table presents information about the 15 degradation conditions taken from the SwissQual 502 speech database. The Mean Opinion Scores (MOS) per condition is also displayed. More details can be found in [9].

Condition Number	Degradation Description	MOS per Condition
1	SWB	4.625
2	SWB+Noise 12dB	1.594
3	SWB+Noise 20dB	3.365
6	SWB Level -10dB	4.031
7	SWB Level -20dB	2.792
32	EVRC-A	3.698
33	EVRC-A + Noise 18dB SNR + Codec NS	1.990
43	VoIP WB-Call + acoust. send	4.042
44	VoIP WB-Call + -16dB + acoust. send	2.297
45	VoIP WB-Call + -8dB + acoust. send	3.677
46	VoIP WB-Call + +5dB + acoust. send	3.615
47	VoIP WB-Call + acoust. noise (rcv)	2.688
48	VoIP WB-Call + acoust. noise (rcv) + -8dB	2.740
49	VoIP WB-Call + acoust. noise (rcv) + -16dB	2.594
50	VoIP WB-Call + ampl. clipping + acoust. send	2.146

C

Appendix C

C.1 Speech Database SwissQual P.501 Annex D

Table C.1: The table presents information about the 53 degradation conditions of the SwissQual P.501 Annex D speech database. This database was used in the studies conducted in Subsection 4.2. The table also shows the Mean Opinion Scores (MOS) per condition. More details in [78].

Condition Number	Degradation Description	MOS per Condition
1	FB reference	4.615
2	WB (P.341 filtered, 7kHz LP)	4.292
3	100-5000Hz BP	3.438
4	FB 2% Time Clipping	3.177
5	FB 20% Time Clipping	1.281
6	EVS 24.4 kbps SWB	4.469
7	EVS 13.2 kbps SWB	4.292
8	OPUS CBR 20kbps WB	4.208
9	AMR-WB 23.85 kbps	3.667
10	AMR-WB 12.65 kbps	3.458
11	AMR-NB 12.2 kbps	2.604
12	EVS 24.4 kbps SWB + losses	2.448
13	AMR-NB 12.2 kbps + losses	1.875
14	4 x EVS 9600 SWB	1.615
15	2 x AMR-WB 6.6kbps	2.010
16	M2M VoLTE call EVS 24.4kbps SWB good network conditions 1	4.219
17	M2M VoLTE call EVS 24.4kbps SWB good network conditions 2	4.344
18	M2M VoLTE call EVS 24.4kbps SWB packet loss 1	2.531
19	M2M VoLTE call EVS 24.4kbps SWB packet loss 2	1.656
20	M2M VoLTE call EVS 24.4kbps SWB variable delay + packet loss 1	3.281

Condition Number	Degradation Description	MOS per Condition
21	M2M VoLTE call EVS 24.4kbps SWB variable delay + packet loss 2	2.229
22	M2M VoLTE call EVS 24.4kbps SWB variable delay + packet loss 3	2.000
23	M2M VoLTE call EVS 24.4kbps SWB variable delay 1	2.844
24	WhatsApp call good network conditions 1	1.146
25	WhatsApp call good network conditions 2	4.250
26	WhatsApp call packet loss 1	3.646
27	WhatsApp call packet loss 2	2.719
28	WhatsApp call packet loss 3	2.854
29	WhatsApp call variable delay 1	2.167
30	WhatsApp call variable delay 2	2.656
31	WhatsApp call variable delay + interruptions 1	1.146
32	WhatsApp call variable delay + interruptions 2	2.635
33	M2M UMTS call AMR-WB 23.85kbps good network conditions 1	3.979
34	M2M UMTS call AMR-WB 23.85kbps good network conditions 2	4.010
35	M2M UMTS call AMR-WB 23.85kbps avg. network conditions 1	2.198
36	M2M UMTS call AMR-WB 23.85kbps avg. network conditions 2	1.823
37	M2M UMTS call AMR-WB 23.85kbps avg. network conditions 3	3.052
38	M2M UMTS call AMR-WB 23.85kbps bad network conditions 1	1.260
39	M2M UMTS call AMR-WB 23.85kbps bad network conditions 2	1.208
40	M2M UMTS call AMR-WB 12.65kbps good network conditions 1	3.427
41	M2M UMTS to GSM call AMR-WB 12.65kbps interruption	1.260
42	M2M UMTS call AMR-WB 12.65kbps good network conditions 1	3.406
43	M2M UMTS call AMR-WB 12.65kbps good network conditions 2	3.563
44	M2M UMTS call AMR-WB 12.65kbps average network conditions 1	2.719
45	M2M UMTS call AMR-WB 12.65kbps average network conditions 2	3.135
46	M2M UMTS call AMR-NB 12.2kbps good network conditions 1	2.813
47	M2M UMTS call AMR-NB 12.2kbps good network conditions 2	2.802
48	M2M UMTS call AMR-NB 12.2kbps average network conditions 1	2.344
49	M2M UMTS call AMR-NB 12.2kbps average network conditions 2	2.563
50	M2M UMTS call AMR-WB 12.65kbps to AMR-NB 12.2 transcoding 1	2.802
51	M2M UMTS call AMR-WB 12.65kbps to AMR-NB 12.2 transcoding 2	2.656
52	M2M UMTS call AMR-WB 12.65kbps to AMR-NB 12.2 transcoding + interruption 1	1.958
53	M2M UMTS call AMR-WB 12.65kbps to AMR-NB 12.2 transcoding + interruption 2	1.156