Rafael Zequeira Jiménez

Influencing Factors in Speech Quality Assessment using Crowdsourcing

Springer

Influencing Factors in Speech Quality Assessment using Crowdsourcing

Rafael Zequeira Jiménez

Influencing Factors in Speech Quality Assessment using Crowdsourcing

Springer

Rafael Zequeira Jiménez
Quality and Usability Lab
Technische Universität Berlin
Berlin, Germany

ISBN 978-3-030-93309-8 ISBN 978-3-030-93310-4 (eBook)
https://doi.org/10.1007/978-3-030-93310-4

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

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, who I love very much. Thanks for teaching me the multiplication tables, which was the genesis of all this. .

.

.

.

.

.

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 whether 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 [1–3].

The quality of the transmitted speech signal is essential for telecommunication network providers. It is an important indicator used to evaluate their systems and 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 and 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 book 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 43 dB(A) and 50 dB(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 30.6 dBA to 81.3 dBA) 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 Pearson correlation to the laboratory results was strong and significant, and the RMSE was 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 [1–3].

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 43 dB(A) und 50 dB(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 30,6 dBA bis 81,3 dBA) 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.

Acknowledgments

During my time working on this book 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, and 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, and 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.

Contents