A Next Step Towards Measuring Perceived Quality of Speech Through Physiology

Sebastian Arndt¹, Markus Wenzel²,³, Jan-Niklas Antons¹, Friedemann Küsters¹, Sebastian Möller¹, Gabriel Curio⁴

¹Quality and Usability Lab, Technische Universität Berlin, Berlin, Germany
²Machine Learning Group, Technische Universität Berlin, Berlin, Germany
³Neurotechnology Group, Technische Universität Berlin, Berlin, Germany
⁴Neurophysics Group, Charité-Universitätsmedizin, Berlin, Germany

sebastian.arndt@telekom.de

Abstract

For service providers, it is important to know the perceived quality of a transmitted speech signal, as this can be an indicator for choosing one operator over the other. To evaluate the perceived quality, several methods have been introduced in the area of Quality of Experience (QoE), such as subjective, instrumental and physiological methods. In the current study, we expose the test participants towards several speech files, while brain activity was recorded with electroencephalography (EEG). Additionally, a subjective quality judgment is obtained after each presentation which is summarized as the mean opinion score (MOS). To validate the selected stimulus corpus, a quality prediction algorithm is used which calculates quality scores based on the speech signal. The recorded EEG data were set in relation to the subjective data and show promising results. In contrast to most previous studies in the domain of QoE using EEG, the study at hand uses a standardized test paradigm proposed by the International Telecommunication Unit (ITU).

Index Terms: Speech Quality, Physiology, Instrumental Quality

1. Introduction

Perceived speech quality is important for users of a speech transmission system, as extra listening effort that results from a degraded signal is not desirable. This is the case, either in the context of an interactive system, such as telephony, or in a scenario where the user is listening passively to speech files, e.g. podcasts.

To determine the quality of a speech signal, several methods can be used. 1) Subjective quality ratings, 2) estimating the quality by analyzing the speech file on the signal level, or 3) directly measuring a physiological response of the user. Each method has its advantages and disadvantages: While the first two methods are already well established in the quality domain and commonly used, the latter one has risen only recently. It was introduced for the audio quality domain by Antons et al. [1] who show that the response of brain signals can be more sensitive towards the identification of distortions than the behavioral answer.

Subjective test methods have their basis in research, performed in the area of quality perception and Quality of Experience (QoE). Test participants listen to an audio file and are asked afterwards to rate the perceived quality on a rating scale. This scale can have the labels: ‘excellent’, ‘good’, ‘fair’, ‘poor’, and ‘bad’, as recommended by ITU-T Recommendation P.800. The overall average quality judgment is summarized in a mean opinion score (MOS). A description of how these tests should be conducted can be found in recommendations from the International Telecommunication Unit (ITU), as e.g. ITU-T Rec. P.800 [2], ITU-T Rec. P.85 [3], etc. These subjective tests are time and money consuming and not always possible to conduct, as the explicit knowledge and capacities to conduct subjective tests are not available. Therefore, quality estimation models have been developed. These models estimate quality values which are based on previously acquired subjective data. Most of these models compare a degraded signal with its reference speech signal and estimate the perceived MOS value.

Recently, a further method has been introduced to the area of quality research, namely physiological measures. Especially, electrical brain signals which are recorded with electroencephalography (EEG). It was shown that they bring valuable supporting and complementary information about the participants cognitive state, such as e.g. cognitive load [1] or fatigue [4]. In the case of cognitive load, the event-related potential (ERP) is of interest: An ERP is being elicited when a test participant is exposed to an external stimulus. In previous research, it could be shown that the P300 component, which is part of an ERP, is being stronger the more distorted an audio signal is [1]. In the domain of video quality, strong correlations between subjectively obtained MOS scores and the P300 amplitude could be shown [5].

Thus, the combination of subjectively obtained judgments and brain signals has proven to be a valid complementary assessment method. However, in these previous studies only very short stimuli, either single syllables or words, have been used. To establish physiological measures in the quality domain, and specifically recordings of brain activity, the test setup has to be moved closer to the test setups as recommended by ITU. E.g. in [8] already longer speech files were used and could confirm previous findings. However, in this study only one sentence was used.

The current study uses longer stimuli which are more consistent with ITU recommendations, as they have a length of 5 s to 10 s. We recorded EEG, collected a subjective quality judgment and additionally generated the predicted quality of the speech signals by a model.

In the following, the experimental test setup and corre-
sponding measures are described, and the results and analyses are detailed. The paper concludes with a general discussion and an outlook for future work.

2. Experiment

2.1. Participants

Eleven subjects participated in the study, two had to be excluded due to technical problems, resulting into nine subjects to be analyzed (4 female, 5 male) with an average age of 27.2 years (range from 24 to 30 years). All were native German speakers. None reported any medical problems or hearing problems. All gave confirmed consent and received monetary compensation.

2.2. Stimuli

The stimuli used in this study were spoken by a male and a female speaker. 20 different sentences were chosen, each was spoken by both speakers. The original recorded signal was processed by an internal toolbox for adding signal correlated noise (SNR) to the speech signal. This was done by using a MNRU (modulated noise reference unit) according to ITU-T P.810 [6]. The SNR levels were chosen to be at 5 dB, 10 dB, 14 dB, 16 dB, 18 dB and between 20 dB to 35 dB in 1 dB steps, thus 22 levels. Resulting in a total corpus of 880 stimuli. Stimuli were presented via in-ear headphones.

2.3. Procedure

Table 1: Individual chosen MNRU levels for each participant.

<table>
<thead>
<tr>
<th>Participant</th>
<th>HQ</th>
<th>LQI</th>
<th>LQII</th>
<th>LQmax</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>100</td>
<td>28</td>
<td>32</td>
<td>10</td>
</tr>
<tr>
<td>2</td>
<td>100</td>
<td>24</td>
<td>32</td>
<td>10</td>
</tr>
<tr>
<td>3</td>
<td>100</td>
<td>22</td>
<td>26</td>
<td>10</td>
</tr>
<tr>
<td>4</td>
<td>100</td>
<td>26</td>
<td>30</td>
<td>10</td>
</tr>
<tr>
<td>5</td>
<td>100</td>
<td>26</td>
<td>30</td>
<td>10</td>
</tr>
<tr>
<td>6</td>
<td>100</td>
<td>26</td>
<td>32</td>
<td>10</td>
</tr>
<tr>
<td>7</td>
<td>100</td>
<td>28</td>
<td>34</td>
<td>10</td>
</tr>
<tr>
<td>8</td>
<td>100</td>
<td>22</td>
<td>26</td>
<td>10</td>
</tr>
<tr>
<td>9</td>
<td>100</td>
<td>26</td>
<td>30</td>
<td>10</td>
</tr>
</tbody>
</table>

During a short training phase, subjects could adjust the loudness to their preferred level. In the subsequent calibration, the individual noise levels were chosen for each subject. Therefore, one sentence, uttered by both speakers, was presented to the participant in all 22 available quality levels ten times each. During this calibration phase, the task of the test participant was to tell whether they perceived a distortion or not. Analyzing the behavioral data, the levels which were closest to a detection rate of 50% and 75% were chosen; and labeled from here on as LQI (50% detection rate) and LQII (75% detection rate). Additionally, the reference condition, i.e. 100 dB, (named HQ) and the most distorted condition with 10 dB SNR level (LQmax) built the stimulus set for each subject. An overview of the chosen levels for each subject can be seen in Table 1. The selection of a small subset of possible degradations is essential when performing EEG studies as a large number of repetitions is necessary for each condition [7].

During the main test, subjects listened to the 20 different selected sentences which were spoken by the two speakers in the predetermined four quality levels, resulting in a total of 160 trials per test participant. After each stimulus presentation, participants had to rate the quality of the audio file on a MOS scale. Therefore, a modified external keyboard was used with the corresponding labels for the quality levels: Namely, 'Excellent', 'Good', 'Fair', 'Poor', 'Bad', as recommended by the ITU [2].

Two-way repeated measure ANOVA, quality level as independent variable, F(3, 24) = 259.01, p ≤ 0.01, η² = 0.97). This finding proves that the selected quality levels for lower SNR values. Calculating a Pearson correlation between PESQ values and quality levels results into a significance (r = 0.96, p ≤ 0.01).

3. Results

3.1. Subjective Quality Ratings

Figure 1a summarizes the averaged results of the subjective judgment process at the end of each trial. High quality stimuli were rated significantly better than low quality stimuli (repeated measure ANOVA, quality level as independent variable, MOS as dependent variable, F(3, 24) = 259.01, p ≤ 0.01, η² = 0.97).

3.2. Quality Prediction

To compare the subjective results, predictions of the overall quality of the speech material are incorporated. The long established standard WB-PESQ (wideband perceptual evaluation of speech quality) is used for the estimations, which has demonstrated acceptable accuracy for measuring the effects of one-way speech distortion and noise on the overall speech quality [10]. The measurement algorithm is a full reference model that operates by performing a comparison between a known reference signal and a captured degraded signal that was sent through a transmission system, addressing NB (narrowband 300 - 3400 Hz) and WB (wideband 50 - 7000 Hz) channels. To generate the estimations for the data in the WB-mode of PESQ, the samples were preprocessed (band-pass-filter 50Hz - 7kHz) and level adjusted.

The analyzed P300 amplitude was determined by identifying the maximum amplitude of the ERP signal in the range of 250 ms to 550 ms for each subject and condition.
Subjective MOS rating.

(b) Predicted MOS by PESQ.

Figure 1: Overview of (1a) subjective quality MOS, and (1b) estimated quality MOS by PESQ. Error-bars indicate 95% confidence interval.

3.3. Combination

Analyzing the relationship between the calculated PESQ MOS and the subjectively obtained MOS, a Pearson correlation was calculated and resulted into a significant correlation of $r = 0.86$ ($p \leq 0.01$). This shows that PESQ is able to correctly estimate the subjective MOS ratings for speech distorted by signal-correlated noise and validates our results. However, it can also be seen, that PESQ slightly over-predicts the ratings for the lowest quality level $LQ_{max}$. A possible explanation could be wide variety of signal-correlated noise ratios and that in comparison to the other levels $LQ_{max}$ was subjectively rated much worse, which cannot be considered by PESQ.

3.4. EEG

Figure 2 shows, the grand average of the elicited ERP averaged over all participants for all rated MOS level. The $P300$ is larger with better MOS rating, except for the MOS rating ‘fair’ which has the lowest amplitude. We found a tendency using the $P300$ amplitude as the dependent variable and the MOS value as independent variable. A repeated measure ANOVA ($F(4, 32) = 2.97, p = 0.06, \eta^2 = 0.27$) did not show a significant outcome.

4. Discussion

The results suggest a meaningful dependency of all three analyzed methods. The instrumental quality estimation by PESQ gives enough reason that the different audio files already differ in quality on a signal level. The obtained subjective scores give the same trend, that a lower signal to noise ratio was perceived worse and led to a lower subjective MOS score. Furthermore, this relationship is supported by a high correlation between the subjective and instrumental MOS. Lastly, we analyzed the recorded physiological data which can be seen as an intermediate step between subjective MOS and instrumental MOS. Using this measure, no conscious quality rating is necessary and might impact good quality estimations in the future. The subjective quality rating appears to be related to the $P300$ amplitude of the test participants, even though the effect showed only a tendency and was not statistically significant.

In the current study we used ecologically valid stimuli as well as a test setup that was closer to the recommended test setups by ITU, which are commonly used in the domain of quality research. The presented data give reason enough to see this study as a further step to developing EEG paradigms in this research domain. When analyzing the predicted MOS of $LQ_{I}$ and $LQ_{II}$, it is obvious that they have a very similar MOS value. Thus, as a first study using this design, it might have been better to choose these levels not only on the basis of the detection rate but also on e.g. PESQ values, and therefor have a linear stepwise increase of MOS.

Future work could additionally inspect the continuous EEG and analyze whether low quality conditions lead to more fatigue than high quality stimuli, as suggested by [4]. A follow-up study could use quality levels which are more apart and therefore show the effect more obviously. However, the presented study shows successful a next step towards measuring quality using brain signals.

5. Acknowledgment

This work was co-funded by the German Research Foundation (DFG - 1013 Prospective Design of Human-Technology Interaction) and the Bernstein Focus: Neurotechnology Berlin (BFNT-B) which was supported by the Bundesministerium für Bildung und Forschung (BMBF) FKZ 01GQ0850/851.

6. References

Figure 2: Grand Average ERP at electrode Cz separated for each answer category.

