SUBJECTIVE QUALITY RATINGS AND PHYSIOLOGICAL CORRELATES OF SYNTHESIZED SPEECH

Sebastian Arndt[1], *Jan-Niklas Antons*[1], *Rishabh Gupta*[2], *Khalil ur Rehman Laghari*[2],
Robert Schleicher[1], *Sebastian Möller*[1], *Tiago H. Falk*[2]

[1]Quality and Usability Lab, Berlin Institute of Technology, Germany
[2]INRS-EMT, University of Quebec, Canada

ABSTRACT

Evaluating the quality of text-to-speech systems (TTS) is usually achieved by subjective methods where participants have to rate the stimulus on multiple scales, such as naturalness, prosody, and overall quality. In the present study, we aim towards evaluating TTS system quality using not only conventional subjective methods, but also via a neurophysiological approach based on obtaining neural correlates of TTS quality perception using electroencephalography (EEG). Such an approach allows for better insight into the perception processes involved during the human quality judgement process, and may open doors to innovative subjective testing methods and/or objective measurement tools. In our experiments, we have shown an inverse relationship between TTS speech quality and the amplitude of an EEG evoked response called the 'P300,' suggesting an increase in cognitive load as TTS quality decreases, likely due to reduction in speech intelligibility.

Index Terms— Quality of Experience, Text-to-Speech, Audio, Electroencephalography, Mean Opinion Score

1. INTRODUCTION

Text-to-speech systems (TTS) are important in everyday life, as they are commonly used in navigation and public announcement systems. Ultimately, high quality synthesized outputs are desired, as low quality systems can lead to frustration (e.g. people do not understand the announcement in the subway) or even lead to dangerous situations if e.g. the driver has to take a closer look to the navigation system while driving. In fact, for the visually impaired who depend on TTS systems for many everyday activities (e.g. read an online newspaper, browse the internet, use a smartphone), high-quality TTS systems are crucial. Characterizing TTS quality however, is not a trivial task as multiple dimensions are involved (e.g. prosody, speaking style, speaking rate).

Traditionally, the quality of TTS systems is assessed by standardized subjective quality testing methods, such as those described by ITU-T Rec. P.85 [1]. It instructs listeners to rate the signal using eight quality dimensions labeled: overall impression, listening effort, comprehension, articulation, pronunciation, speaking rate, pleasantness, and acceptance. Previous studies, however, have shown that these eight dimensions may not capture the full scope of more modern TTS systems [2]. More recently, three perceptual quality dimensions were shown to reliably characterize modern TTS systems: naturalness, temporal distortions and disturbances [3].

In this paper, we take an alternative approach to improving subjective testing of TTS systems. More specifically, we propose to use neurophysiological signal monitoring to characterize human affective and cognitive states and investigate their relationship with perceived quality and quality of experience ratings. Physiological measures such as electroencephalography (EEG) have been shown in the past to reliably augment subjective quality tests of natural speech samples impaired by modern codecs used in the telecommunication industry [4]. One commonly used EEG component is the so-called P300, a positive peak in the EEG signal occurring approximately 300 ms after stimulus onset, shown to be modulated by e.g., cognitive load or stimulus intensity [5].

The EEG has also been used in the past to characterize different cognitive and affective states. In [6], for example, it was used to provide information about listener fatigue during subjective studies. It has also been used to augment subjective testing of visual stimuli [7, 8] and to characterize the emotional states of listeners/viewers of different audio-visual stimuli [9]. Motivated by these recent findings, this study aims to collect neural correlates of human TTS quality perception and affective states using the P300 component.

The remainder of this paper is organized as follows: first we provide an overview of the research methodology followed in our study. Next, we present our experimental results for both the subjective and neurophysiological tests, provid-

This work was co-funded by the German Academic Exchange Service (DAAD), the German Research Foundation (DFG - 1013 Prospective Design of Human-Technology Interaction), the Bernstein Focus: Neurotechnology Berlin (BFNT-B), the Federal Ministry of Education and Research (BMBF) FKZ 01GQ0850, and supported by Telekom Innovation Laboratories. In Canada, it was co-funded by the Canadian Foundation for Innovation, the Ministere de l'Education, du Loisir et du Sport du Quebec, the Ministere du Developpement Economique, Innovation et Exportation du Quebec, and the National Science and Engineering Research Council of Canada

ing a link between the two. Lastly, we discuss the obtained results in light of existing literature and conclude the paper with a brief outlook to future work.

2. MATERIALS AND METHODS

In this section, we present the materials and methods used in our experimental study.

2.1. Participants

Fourteen participants (6 male and 8 female) were recruited amongst McGill University students and staff (average age of 21.6 years). All subjects were fluent in English, ten of which were native English speakers. None reported any hearing problems or neurophysiological diseases. Ethics approval was obtained by the Research Ethics Offices at INRS-EMT and McGill. Participants expressed written consent to participate in the study.

2.2. Stimuli

As a source for stimuli we chose the database of the Blizzard Challenge 2009 [10]. Here, three different TTS systems were explored and all stimuli used were generated from the same speaker's voice. The three systems were representative of low- (LQ), medium- (MQ) and high-quality (HQ) systems (MOS-LQ = 1.9; MQ = 2.8; HQ = 3.7). Additionally, the natural voice used to generate the stimuli was used for benchmarking (referred to as natural). For each of the four conditions, four different English sentences (stimuli) were selected. The length of each stimulus was between 8 and 11 seconds. The sentences were a response of a restaurant recommendation system, thus, very neutral in content and also the voice itself was very neutral. All stimuli had a sampling rate of 16 kHz and a bitrate of 256 kbps.

2.3. EEG

For EEG recording we used a 64 channel Biosemi system, with electrodes arranged in the 10-20 standard system, recorded at (AF3-4, FZ, 3-10; FFC1-2, 5-8; FT7-10; FCz, 1-6; CFC5-8;Cz, 3-6; CCP7-8; CP1-2, 5-6; T7-8; TP7-10; P3-4, Pz, 7-8; POz; O1-2)[11]. Impedances of the electrodes were kept below 20 kΩ and recordings were done with a sampling rate of 2000 Hz. No online filtering was applied. For referencing mastoids at both sides were used. Additionally, an EOG (electro-oculogram) was recorded from the participants in order to register eye movement and blinking, thus allowing for such artifacts to be removed from the EEG with independent component analysis. For offline analysis, EEG recordings were down-sampled to 200 Hz and a 0.1 to 60 Hz bandpass filter was applied.

2.4. Test 1: Subjective Part

All 16 stimuli (4 conditions x 4 sentences) were rated in terms of valence and arousal by each listener using the Self-Assessment Manikin scale (SAM) [12]. Subjects rated stimuli on a continuous scale with nine different manikins for each scale. Additionally, three other subjective quality ratings were assessed, namely comprehension problems, fluency, and overall quality, using a 5-point mean opinion score (MOS) like scale. During this stage of the experiment, participants had the chance to repeat the stimuli before providing their ratings. The stimuli were pseudo-randomized between subjects with the only constraint that the low quality sentence was always played before their corresponding high-quality counterparts, as to mitigate the effects of memory recall on stimuli comprehension.

2.5. Test 2: Neurophysiological Part

Once participants were finished with the subjective test, they were set up with the EEG system. Participants sat in front of a 22-inch screen and were given a keyboard to provide a binary response as to whether they found the stimulus presented to them was 'pleasant' or 'unpleasant'. This part of the experiment was divided into six blocks, each lasting for about 10 minutes. One block consisted of 12 stimuli. The inter-stimulus interval was 20 s, which gave the participants enough time to focus on the next stimulus. The stimuli were pseudo-randomized within blocks and subjects. Blocks were randomized between subjects.

3. EXPERIMENTAL RESULTS

In this section, we present experimental results from the subjective and neurophysiological monitoring experiments.

3.1. Subjective Ratings

Figure 1 depicts the boxplots for the subjective ratings obtained during the first test. As expected, MOS, comprehension and fluency ratings decrease as TTS quality decreases. For the valence scale a similar trend can be observed, but with medium and low quality stimuli achieving similar ratings. Interestingly, for the arousal ratings, natural speech showed to arouse participants the most. For the TTS systems, higher arousal scores were obtained for low-quality systems. Testing for statistical significance with a repeated analysis of variance (ANOVA) yielded a statistical main effect for each subjective score. Table 1 shows the obtained F and eta squared results. Also pair-wise comparisons, Bonferroni-corrected, yield significance on a level of $\alpha \leq 0.05$ for each pair, but valence between MQ and LQ, and arousal for HQ-MQ, HQ-LQ and MQ-LQ. Lastly, Figure 2 depicts the subjective pleasantness ratings obtained during the neurophysiological test. As can

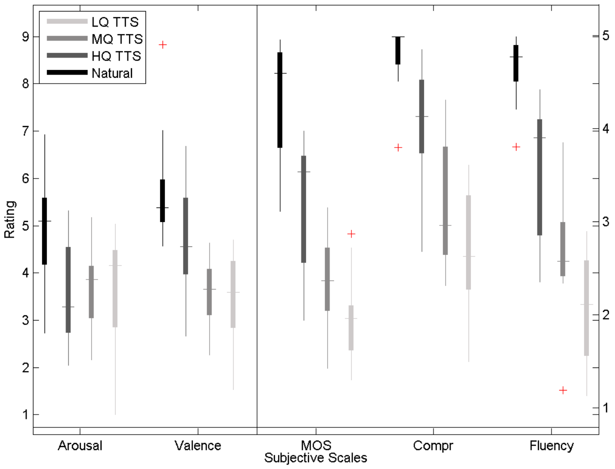

Fig. 1. Boxplots for subjective scores over all subjects. Line in the box represents the median. Edges of the boxes are the 25th and 75th percentiles. Whiskers denote most extreme points.

	p	F(3, 33)	η^2
Arousal	≤ 0.01	10.92	0.5
Valence	≤ 0.01	24.34	0.69
MOS	≤ 0.01	48.15	0.81
Comprehension	≤ 0.01	33.62	0.75
Fluency	≤ 0.01	46.82	0.81

Table 1. Overview of repeated ANOVA measurements for compiled subjective scores. With quality levels as independent variable and the corresponding subjective scores as dependent variable.

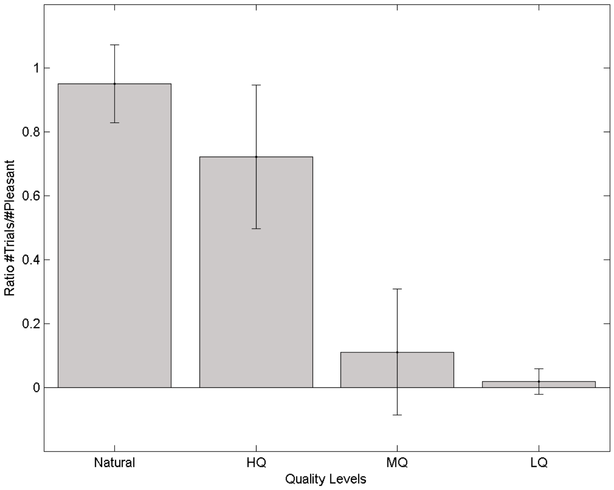

Fig. 2. Response distribution averaged over all subjects with number of pleasant trials divided by number of trials for each condition.

	AF3	AF4	Δ(AF4, AF3)
unpleasant	2.82	3.71	0.89
pleasant	1.26	1.34	0.08

Table 2. P300 Amplitudes at electrodes AF3 and AF4 for different responses (pleasant vs. unpleasant) for the stimuli. Plus difference between AF4 and AF3 in amplitude. Data is based on grand average.

be seen, pleasantness ratings sharply decrease as TTS quality decreases.

3.2. Neurophysiological Insights

Figure 3 depicts a representative P300 signal measured from electrode CPz. As can be seen, the P300 amplitude increases with a decrease in TTS quality. This suggests that more neuronal networks are involved in the comprehension of low quality text-to-speech systems. The EEG response for natural speech in the cue of synthetic speech is resulting into the highest ERP. Calculating an ANOVA with the P300 peak amplitude as dependent variable and the quality level as independent yields statistical significance (F(3, 33) = 4.39, $p \leq 0.05$, $\eta^2 = 0.29$).

Moreover, in order to to further investigate the effects of emotional response on quality of experience perception, we examined the frontal lateralization index, as suggested in previous studies [13]. Here, we explored if differences in EEG patterns for the left versus right frontal hemisphere were related to the pleasantness ratings. As such, EEG measured from electrodes AF3 and AF4 were used. It was observed that the P300 amplitude differences between AF3 and AF4 are greater for unpleasant than for pleasant stimuli, as shown in Table 2. However, this effect is not statistically significant but from the data at least a trend can be observed (F(1, 9) = 1.4, n.s.).

3.3. Correlating Subjective and Physiological Findings

In order to evaluate the correctness of EEG as a complementary method for subjective testing, the results of both methodologies were correlated. An overview of correlation values between all the gathered scores can be seen in Table 3. As was to be expected, subjective scores correlate very well with each other. Overall, neurophysiological insights (namely P300 peak amplitude) were inversely correlated with the subjective scores. Larger correlation coefficients were obtained with the MOS and comprehension scales. Since the chosen

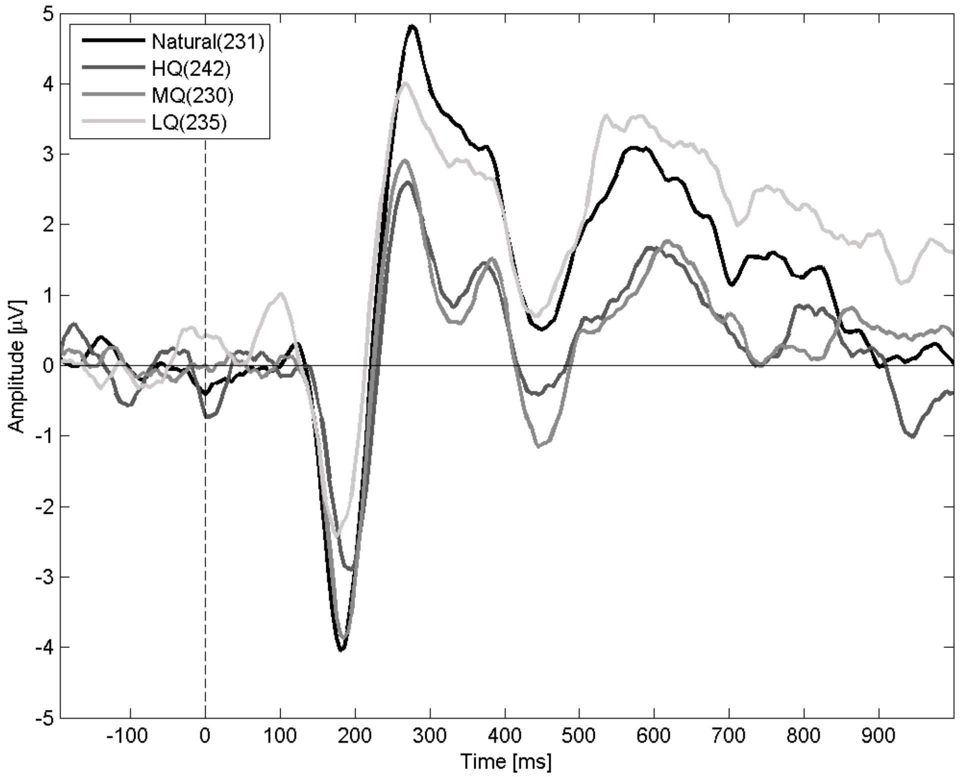

Fig. 3. Grand average EEG data from electrode CPz. Time point 0 depicts stimulus onset. Times from -200 ms to 0 ms were taken as baseline. Abbreviations: Natural: Natural Speech; HQ: high quality TTS; MQ: medium quality TTS; LQ: low quality TTS.

	MOS	C	F	A	V	P300
MOS	1	0.93*	0.99*	0.56	0.93*	-0.45
C	-	1	0.93*	0.52	0.86	-0.47
F	-	-	1	0.58	0.92*	-0.36
A	-	-	-	1	0.54	0.18
V	-	-	-	-	1	-0.18

Table 3. Overview of calculated correlations between subjective scores and P300. * Indicates statistical significance on a $\alpha \leq 0.1$ level. Abbreviations stand for: MOS: Mean Opinion Score; C: Comprehension; F: Fluency; A: Arousal; V: Valence; P300: P300 Amplitude

alpha value for statistical significance was chosen on a 0.1 level, rather a trend than a strong correlation can be claimed.

4. DISCUSSION

This study investigated the feasibility of using physiological measurements to complement subjective quality studies using synthetic speech. We showed that different quality levels are processed in the brain differently. More specifically, a reverse relationship between perceived quality (i.e., MOS) and the P300 amplitudes could be observed, suggesting an increased mental load with low-quality TTS systems. This finding suggests that listener fatigue may play a crucial role in subjective listening tests using varying-quality TTS stimuli, as was shown in other tests [6].

The fact that the natural speech and LQ P300 amplitudes have a similar trend can be due to several reasons. One of them being that natural speech was practically a deviant in a cue of synthesized speech. As such, natural speech could be considered to be a so-called “oddball” stimulus, which is known to generate an increase in P300 amplitudes. Another possible explanation could be that the LQ TTS and natural speech stimuli are arranged on opposite ends on a scale describing the naturalness of a stimulus. For both stimulus classes participants had a very distinct rating concerning pleasantness during the EEG experiment (almost 100% pleasantness for natural speech versus almost 0% for LQ, see Figure 2). Bearing this in mind and investigating the ERP curve, it can be seen that the curves for the two conditions have a similar course. The P3a which is the standard physiological response to (auditory) stimuli is present for all conditions. But the subsequent decline is not as steep as for the HQ and MQ condition. Thus, the developing in the LQ and natural condition might be a P3b which is rather task related [14]. This is supported furthermore by [15] since the answers for the two conditions were very clear. This means subjects could classify the two stimuli conditions very easily into the categories which were given for responding (pleasant versus not pleasant). This was not the case for the other two conditions. Thus, a clearer categorizing in this scenario leads to a more distinct P3b.

This leads to the suggestion that a bipolar rating like this may not be assessed easily using ERPs, at least not in this setup. Nevertheless, assessing the graduation of a particular distortion or effect ERPs have again proven to be an appropriate tool. In any case, for ERP analysis it might be not of any advantage to analyze natural speech and synthetic speech from the same experiment if their odds are distributed like this. Nevertheless, the natural speech condition gives participants a ‘ground truth’ condition and make them feel more comfortable. Thus, it might be beneficial to still include this condition in future studies.

Lastly, compared to standard quality tests this experiment was conducted with far fewer subjects. This is due to the fact that a more complex setup and preparation is needed, such as placing electrodes and checking impedances. Moreover, since EEG P300 components need to be averaged over several trials in order to obtain reliable signal-to-noise ratios, subjective tests using neurophysiological monitoring tend to last much longer than conventional subjective tests. While these factors represent disadvantages of physiological measurements, the type of data collected may provide complementary information not obtained with subjective testing, such as fatigue indicators, affective states, or even comprehension.

Another aspect of this study was to explore whether the pleasantness rating had an effect on the recorded EEG. As such, the frontal electrodes were looked at more closely. To explore pleasantness connections between the quality levels, the alpha band from the measured EEG was used. Since there is an inverse relationship between alpha activity and P300 amplitude, ERPs can be used instead for brief analysis of this aspect [16]. A general trend for the differences between the amplitudes of AF3 and AF4 can be seen; as this value is lower for pleasant stimuli than for unpleasant ones, although this is not statistically significant. As seen in Figure 2, the MQ and LQ conditions interestingly result into similar ratings for arousal and valence. Since the used stimuli was neutral in content, it is expected that the emotional ratings achieve this plateau. To avoid this issue, future studies could explore varying emotional content in addition to varying TTS quality.

5. CONCLUSION AND FUTURE WORK

In this paper, we proposed the use of neurophysiological monitoring to complement subjective quality tests for text-to-speech (TTS) systems. By measuring electroencephalography data during a subjective quality perception test, we found some relationships between EEG P300 responses and perceived quality as well as emotional states. Future studies will focus on frequency band power analysis, which has also been linked to affective state characterization. Moreover, future studies will focus on TTS stimuli of varying content, thus potentially covering a wider arousal-valence space.

Acknowledgements

We would like to thank the Centre for Research on Language, Brain and Mind (McGill University, Canada), led by Dr. Vincent Gracco, for letting us use their facility and equipment for this study.

6. REFERENCES

[1] ITU-R Recommendation P.85, “A method for subjective performance assessment of the quality of speech voice output devices,” *International Telecommunication Union, Geneva*, 1994.

[2] D. Sityaev, K. Knill, and T. Burrows, “Comparison of the itu-t p. 85 standard to other methods for the evaluation of text-to-speech systems,” in *Ninth International Conference on Spoken Language Processing*, 2006.

[3] F. Hinterleitner, S. Möller, C. Norrenbrock, and U. Heute, “Perceptual quality dimensions of text-to-speech systems,” in *Proc. Interspeech*, 2011, pp. 2177–2180.

[4] J.N. Antons, R. Schleicher, S. Arndt, S. Möller, and G. Curio, “Analyzing speech quality perception using electroencephalography,” *Selected Topics in Signal Processing, IEEE Journal of*, vol. 6, no. 6, pp. 721 –731, 2012.

[5] M.D. Rugg and M.G.H. Coles, *Electrophysiology of mind: Event-related brain potentials and cognition.*, Oxford University Press, 1995.

[6] J.N. Antons, R. Schleicher, S. Arndt, S. Möller, and G. Curio, “Too tired for calling? a physiological measure of fatigue caused by bandwidth limitations,” in *Quality of Multimedia Experience (QoMEX), 2012 Fourth International Workshop on*. IEEE, 2012, pp. 63–67.

[7] L. Lindemann, S. Wenger, and M. Magnor, "Evaluation of video artifact perception using event-related potentials," in *Proc. ACM Applied Perception in Computer Graphics and Visualization (APGV) 2011*, 2011.

[8] S. Arndt, J.N. Antons, R. Schleicher, S. Möller, and G. Curio, "Perception of low-quality videos analyzed by means of electroencephalography," in *Quality of Multimedia Experience (QoMEX), 2012 Fourth International Workshop on*. IEEE, 2012, pp. 284–289.

[9] S. Koelstra, C: Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "Deap: A database for emotion analysis; using physiological signals," *Affective Computing, IEEE Transactions on*, vol. 3, no. 1, pp. 18–31, 2012.

[10] A.W. Black, S. King, and K. Tokuda, "The blizzard challenge 2009," 2009.

[11] H.H. Jasper, "The ten-twenty electrode system of the international federation," *Electroencephalography and clinical neurophysiology*, vol. 10, no. 2, pp. 371–375, 1958.

[12] M.M. Bradley and P.J. Lang, "Measuring emotion: the self-assessment manikin and the semantic differential," *Journal of behavior therapy and experimental psychiatry*, vol. 25, no. 1, pp. 49–59, 1994.

[13] R.E. Wheeler, R.J. Davidson, and A.J. Tomarken, "Frontal brain asymmetry and emotional reactivity: A biological substrate of affective style," *Psychophysiology*, vol. 30, no. 1, pp. 82–89, 1993.

[14] John Polich, "Updating p300: an integrative theory of p3a and p3b," *Clinical neurophysiology: official journal of the International Federation of Clinical Neurophysiology*, vol. 118, no. 10, pp. 2128, 2007.

[15] Albert Kok, "On the utility of p3 amplitude as a measure of processing capacity," *Psychophysiology*, vol. 38, no. 3, pp. 557–577, 2001.

[16] M.A. Conroy and J. Polich, "Affective valence and p300 when stimulus arousal level is controlled," *Cognition and emotion*, vol. 21, no. 4, pp. 891–901, 2007.