COGNITIVE, AFFECTIVE, AND EXPERIENCE CORRELATES OF SPEECH QUALITY PERCEPTION IN COMPLEX LISTENING CONDITIONS

Jan-Niklas Antons[1], *Khalil ur Rehman Laghari*[2], *Sebastian Arndt*[1], *Robert Schleicher*[1], *Sebastian Möller*[1], *Douglas O'Shaughnessy*[2], and *Tiago H. Falk*[2]

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

ABSTRACT

Subjective speech quality assessment depends on listener "quality" opinions after hearing a particular test speech stimulus. Subjective scores are given based on a perception and quality judgment process that is unique to a particular listener. These processes are postulated to be dependent on the listener's internal reference of what good and bad quality sounds like, as well as their mental and emotional states. To overcome this variability, subjective listening tests often average scores over several listeners. In this paper, we use electroencephalography (EEG) and self-assessment tools to investigate the neural and affective correlates of speech quality perception of reverberant speech, with the goal of obtaining new insights into human speech quality perception in complex listening environments. We show that EEG event related potentials (ERP) are a useful tool to monitor the conscious stages of neural-processing during a speech quality assessment task. Significant correlations were obtained between the so-called P300 ERP component and the reverberation time of the room, as well as between the P300 peak amplitude and emotional self-assessment ratings. These insights could lead to more effective ways of characterizing room acoustics for improved speech quality and intelligibility.

Index Terms— Electroencephalography, reverberation, speech quality assessment, Quality-of-Experience, emotions.

1. INTRODUCTION

Following the considerations of Jekosch [1], speech quality assessment comprises a three-step process: perception, judgment, and description. The first step comprises the reception and perception of an auditory event (i.e., speech sound wave reaches the human ear). The judgment process, in turn, is responsible in getting features from the perceived "event" and comparing them to internal reference features of what good and bad quality speech sounds like. This internal reference is unique to each listener and may be influenced by numerous factors, such as the user's expectations, experience, motivation, affective state, and ambient factors, to name a few. Lastly, the final description process involves the pooling of these judgments into a final overall quality rating. To mitigate the negative effects of such inter-subject variability, subjective listening tests, such as the Mean Opinion Score (MOS) test [2], average all listener quality ratings. While auditory perceptual models exist in the literature, cognitive models of the judgment and description processes are widely unknown.

With the advances in neuroimaging technologies, there is growing interest in investigating the neuro-physiological correlates of human speech quality perception. Ultimately, a better understanding of the internal processes could lead to improved subjective testing protocols, objective quality models, and speech-based technologies. To this end, electroencephalography (EEG) has been used to obtain neural correlates of perceptual [3] and higher cognitive processes [4]. Event related potentials (ERP), particularly the so-called P300 component that arises approximately 300 ms after a stimulus onset, have shown to be useful tools in this endeavour [5]. For example, [6] showed that simple degradations in the speech signal (e.g., multiplicative noise) could be processed by humans at an unconscious level; similar findings were also observed for artificial audiovisual stimuli [7]. Recently, P300 peak amplitudes were shown to be significantly correlated with the level of distortions added to audiovisual stimuli [8] and classifiers were developed to discriminate between clean and distorted visual stimuli based only on EEG [9]. Lastly, EEG has also been used to characterize user emotional state while watching Youtube videos [10].

Motivated by these promising recent findings, this paper aims to investigate the use of EEG to obtain neural and affective correlates of speech quality perception in *complex listening environments*. Focus is placed on hands-free speech communications where reverberation can severely degrade the signal timbre [11, 12], cause temporal smearing, and ultimately degrade speech quality and intelligibility [13]. Here, two reverberant environments are considered: a home living room and a large auditorium. Significant correlations were obtained between EEG P300 amplitude and several cognitive, affective, and experiential parameters. The obtained insights have allowed for recommendations to be developed for subjective quality assessment tests using reverberant speech.

The remainder of this paper is organized as follows. Section II describes the study methodology, Section III presents the experimental results, Section IV discusses the obtained findings, and Section V concludes the paper.

2. MATERIALS AND METHODS

2.1. Participants

Twenty two subjects participated in this study (ten female, twelve male; mean age = 23.40 years; SD = 3.80; range = $18-33$); all of them were fluent English speakers. Due to faulty equipment, data from seven subjects had to be discarded. All participants reported normal auditory acuity and no medical problems. Participants gave informed consent and received monetary compensation for their participation. The study protocol was approved by the Research Ethics Office at INRS-EMT and at McGill University (Montreal, Canada).

2.2. Speech Stimuli

As stimulus, a double-sentence utterance commonly used in subjective quality tests was used. The sentence was uttered by a male speaker in an anechoic chamber and digitized at 8 kHz sampling rate with 16-bit resolution. Room impulse responses recorded in a typical home living room environment (reverberation time of 400 ms) and in an auditorium (reverberation time of 1500 ms) were convolved with the clean speech file to generate the reverberant stimuli. For consistency, all files were normalized to -26 dBov using the ITU-T P.56 voltmeter [14]. Unlike typical subjective quality tests, here only one speech file (three stimuli: one clean and two reverberant) is used in order to maintain a controlled environmental setting, as the P300 signals can be sensitive to varying content.

2.3. Experimental Protocol

The experimental protocol followed two parts. The first consisted of a quantitative "pre-test" component where participants i) filled in a demographic questionnaire, ii) performed a subjective quality test using the Absolute Category Rating (ACR) scale [2] (5-point scale with 1 indicating bad quality and 5 excellent), and iii) rated their elicited emotional states after hearing the different speech files. For the emotional self-assessment, modified versions of the Self-Assessment-Manikin (SAM) scales were used [15]. More specifically, listeners rated the arousal, valence and dominance dimensions using 9-point visual anchors. Lastly, in order to gauge the participant's "experience" with the test, they were also asked to rate their "liking" using a 9-point scale ([1 (not at all) to 9 (very much)]) and how familiar they are with the type of degradation using a 5-point scale ([1 (not at all) to 5 (very much)]). Participants listened to the speech files three times.

The second part of the test consisted of the actual EEG experiment following an oddball paradigm. More specifically, the clean speech file served as the so-called standard stimulus (70% of the trials) and the reverberant files served as deviants (30% of the trials). Clean and reverberant speech files were delivered in a pseudo-randomized order, forcing at least one standard to be presented between successive deviants, in sequences of 100 trials. Stimulus sequences were presented with an inter-stimulus-interval varying from 1000 to 1800 ms. Participants were seated comfortably and were instructed to press a button, whether they detected the clean stimulus or one of the deviants. Stimuli were presented binaurally at the individual's preferred listening level through in-ear headphones.

2.4. Extracted "Cognitive" Parameters

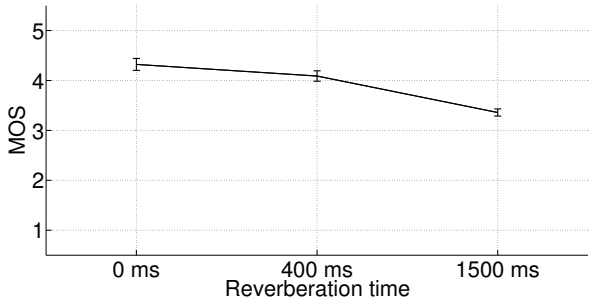
A 128-channel BioSemi EEG system was used but only the following subset was recorded: 64 EEG-electrodes, 4 EOG-electrodes, and two mastoid-electrodes (right and left). Data was recorded at 512 Hz but down-sampled to 200 Hz and band-pass filtered between 1 and 40 Hz for offline analysis. All channels were re-referenced to the average of all EEG-channels. EEG epochs with a length of 2700 ms, time locked to the onset of the stimuli, including a 600 ms pre stimulus baseline, were extracted and averaged separately for each stimulus level and for each participant. To quantify the deviance-related effects of P300, we measured the peak amplitude in a fixed time window relative to the pre-stimulus baseline at electrode Cz. The time window for P300 quantification was set from 200 to 600 ms after stimulus onset. The maximal positive amplitude in this time window was automatically determined and its voltages were extracted for further analysis. Reaction time was also computed for each presented stimulus, and consisted of the time between the stimulus onset and the actual button press.

3. EXPERIMENTAL RESULTS

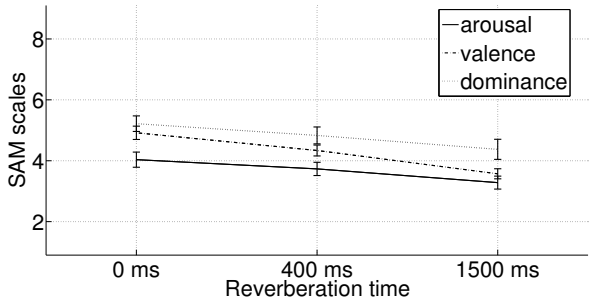
To analyze the data we performed a repeated measure analysis of variance with the independent variable *level of reverberation* and the dependent variables *MOS*, *valence*, *arousal*, *dominance*, *P300 peak amplitude*, and *reaction time*. For the analysis of *liking* and *familiarity*, a Wilcoxon rank sum Paired test was also used as these parameters did not pass a normality test. In the following subsections, test results for the main effects and the Schefé corrected post-hoc comparisons will be reported. Additionally, correlations (Pearson and Spearman) between the quantitative parameters and EEG features will be reported.

3.1. Quality, Emotion, and Experience Correlates

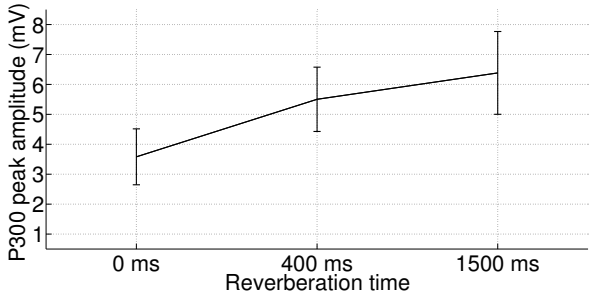
For the *MOS* parameter, a significant main effect for reverberation level ($F(2,16) = 128.89$, $p < .01$, $\eta^2 = .94$) was observed. The plots in Fig. 1 (a) depict the subjective MOS versus reverberation time curve obtained. As can be seen, a monotonic de-



(a)



(b)



(c)

Fig. 1. Plots of subjective (a) MOS, (b) SAM (arousal, valence, and dominance), and (c) P300 peak amplitudes versus reverberation levels averaged over all subjects. Whiskers denote standard errors.

crease in MOS was observed as reverberation time increased. Figure 1 (b), in turn, depicts the three emotional SAM dimensions (arousal, valence, and dominance) versus reverberation time. The *arousal* dimension achieved a main effect for reverberation only at the 95% level ($F(2,16) = 5.45$, $p < .05$, $\eta^2 = .40$), whereas a significant main effect was found for the dimension *valence* ($F(2,16) = 91.85$, $p < .01$, $\eta^2 = .86$) and *dominance* ($F(2,16) = 9.00$, $p < .01$, $\eta^2 = .52$). As can be seen, a monotonic decrease across all three emotion dimensions was observed with an increase in reverberation time.

Moreover, significant main effects were also observed for the *liking* ($F(2,16) = 45.88$, $p < .01$, $\eta^2 = .85$) and *familiarity* experience scales ($F(2,16) = 22.07$, $p < .01$, $\eta^2 = .73$); plots are omitted for brevity but monotonically decreasing curves

Table 1. Schefé corrected post-hoc comparison. Column labels correspond to: RT0 vs. RT1 = clean vs. RT=400 ms; RT0 vs. RT2 = clean vs. RT=1500 ms; RT1 vs. RT2 = RT=400ms vs. RT=1500ms; ns = not significant.

Parameter	RT0 vs. RT1	RT0 vs. RT2	RT1 vs. RT2
MOS	$p < .05$	$p < .05$	$p < .05$
Arousal	$p < .05$	ns	ns
Valence	$p < .05$	$p < .05$	$p < .05$
Dominance	$p < .05$	$p < .05$	ns
Liking	$p < .05$	$p < .05$	$p < .05$
Familiarity	$p < .05$	$p < .05$	ns

Table 2. Correlation matrix of different quantitative parameters. [**: $p < 0.01$ and *: $p < 0.05$].

Parameter	A	V	D	L	F
MOS	0.43**	0.81**	0.19	0.71*	0.37**
Arousal (A)	1	0.65**	0.14	0.38**	0.41**
Valence (V)	-	1	0.34*	0.79**	0.51**
Dominance (D)	-	-	1	0.43**	0.13
Liking (L)	-	-	-	1	0.57**
Familiarity (F)	-	-	-	-	1

were also observed with increasing reverberation time. The Wilcoxon test also showed significant effects for both parameters. Results of the post-hoc comparisons are reported in Table 1. As observed, significant differences were seen for all parameters for the clean vs. reverberation time (RT) = 400 ms scenario (column labelled 'RT0 vs RT1' in the Table) and for the clean vs. RT=1500 ms ('RT0 vs RT2') scenario, with the exception of the arousal dimension. In the RT=400 ms vs RT=1500 ms ('RT1 vs RT2') scenario, only the *MOS*, *valence*, and *liking* scales were significantly different. Lastly, Table 2 reports the correlation matrix for all the collected subjective parameters. As can be seen, the *dominance* dimension is only significantly correlated with the *valence* dimension. Particularly interesting are the high correlations obtained between *MOS* and *valence*, *MOS* and *liking*, and *valence* and *liking*, thus indicating that affective states, quality perception, and Quality-of-Experience (QoE) are inter-related parameters.

3.2. Neural/Cognitive Correlates

Lastly, we explore the neural/cognitive correlates of speech quality perception. It was observed that a significant main effect was present for the *P300 peak amplitude* against reverberation time ($F(2,16) = 8.15$, $p < .01$, $\eta^2 = .50$). The plots in Fig 1 (c) depict the average *P300 peak amplitude* versus reverberation time. As can be seen, P300 amplitude increases with an increase in reverberation time. Table 3 reports the correlations obtained between P300 peaks and all subjective parameters. As can be seen, significant negative correlation was attained with *MOS* and the *valence* dimensions. Lastly,

Table 3. Correlation between P300 amplitude and quantitative parameters [**: $p < 0.01$ and *: $p < 0.05$]. A = arousal, V = valence, D = dominance, L = liking, F = familiarity.

Parameter	MOS	A	V	D	L	F
P300	-0.44*	-0.15	-0.40*	0.15	-0.27	0.01

a significant main effect with reverberation time was also observed for *reaction time* ($F(2,16) = 11.73$, $p < .01$, $\eta^2 = .59$).

4. DISCUSSION

This study investigated the effects of increasing reverberation levels on human self-assessed quality, affective, and experience scores. Inherent human cognitive/neural effects were also observed via EEG P300 amplitudes and reaction times. As expected, subjective quality (*MOS*), experience (e.g., *liking*), and valence ratings decreased as reverberation levels increased. Interestingly, arousal levels also decreased as reverberation times increased. Given the significant positive correlations observed between *arousal* and *liking*, it is conjectured that as reverberation times increased, *listening* quality decreased and participants became less engaged in the task, thus were less aroused. In practical *conversational* situations where reverberation can affect intelligibility, it is expected that increased arousal would be observed with increasing RT.

Moreover, participants felt more dominant in their judgments for the clean stimuli compared to the stimuli with reverberation. With higher reverberation time more temporal smearing occurs and resulted in less dominant judgments. As also expected, participants were more familiar with the quality of the clean stimulus as none of the participants were accustomed to communicating "hands-free" in an environment with such high reverberation levels. As such, the listener's "internal reference" could not account for such distortions. Perhaps if lower reverberation time values were explored (e.g., between 200-500ms) the listeners would have been more familiar with the introduced distortions.

Regarding the observed cognitive/neural correlates observed, P300 peak amplitudes were seen to be significantly correlated with the *MOS* and *valence* parameters, thus shedding light into the human quality judgment and descriptive processes. Moreover, increased P300 amplitudes were observed as reverberation levels increased, suggesting that participants found the listening task to be less demanding as reverberation levels increased. This corroborates with the decrease in *arousal* levels as quality decreased. It is believed that an inverse relationship would have been observed if the test were either an intelligibility or a conversational task, as participants would require greater attentional resources (lower P300 amplitudes) as quality decreased - thus more in line with practical situations. It is recommended, when performing subjective listening quality tests with reverberant speech, that a relevant task be given to the participants such that they remain attentive to the spoken content (e.g., what time will the bus arrive?); this is similar to what is done with listening quality assessment of text-to-speech systems. As in previous literature, we have shown that EEG can be used to gather cognitive, distortion, and quality-of-experience insights ([6], [8], [9], [16]).

Lastly, it was observed that the response time vs. reverberation time curve was non-monotonic. More specifically, the average response times over all participants was 604ms, 739ms, and 691ms for the clean, RT=400ms, and RT=1500ms stimuli, respectively. This behaviour may have been different if a lower range of RT were used and/or if participants were given a task to perform while listening to the speech files. As quality decreased to less acceptable values, participants were quicker in judging the listening quality. For intermediate quality levels, judgment took longer, perhaps because participants were hesitant to describe the final quality score.

5. CONCLUSIONS

This study has explored cognitive, affective, and experiential factors inherent to humans when asked to perform a listening speech quality assessment task; these insights are non-existent in the quality assessment literature. Focus was placed on quality-of-experience (QoE) assessment of reverberant speech, thus simulating burgeoning hands-free communications. Based on the obtained insights, recommendations were given on how to conduct subjective listening quality assessment tests of reverberant speech. It is expected that the obtained results may lead to improved room acoustic characterization algorithms and subjective listening tests.

6. ACKNOWLEDGMENT

The authors are grateful to V. Gracco, L. Coady, H. Cheang, F.-X. Brajot and colleagues from the Centre for Research on Brain, Language and Music for sharing their EEG-hardware, expertise, and discussions, as well as the Bernstein Focus: Neurotechnology - Berlin (BFNT-B), the Federal Ministry of Education and Research (Grant FKZ 01GQ0850), the Ministere du Developpement Economique, Innovation et Exportation du Quebec, and the National Science and Engineering Research Council of Canada for funding this work.

7. REFERENCES

[1] U. Jekosch, "Voice and Speech Quality Perception: Assessment and Evaluation", Berlin, Springer, 2005.

[2] "Methods for Subjective Determination of Transmission Quality", ITU-T Recommendation P.800, International Telecommunication Union, Geneva, 1996.

[3] C. Duncan, R. Barry, J. Connolly, C. Fischer, P. Michie, R. Näätänen, J. Polich, I. Reinvang, C. Petten, "Event-related potentials in clinical research: Guidelines for eliciting, recording, and quantifying mismatch negativity, P300, and N400", *Clinical Neurophysiology*, vol. 120, pp.1883-1903, 2009.

[4] M. S. Coles, M. Rugg, "Event-related brain potentials: an introduction", in Electrophysiology of Mind: Event-Related Brain Potentials and Cognition, Oxford University Press, 1995.

[5] J. Polich, "Updating P300: an integrative theory of P3a and P3b", *Clinical Neurophysiology*, vol. 118(10), pp.2128-2148, 2007.

[6] J.-N. Antons, R. Schleicher, S. Arndt, S. Möller. A.K. Porbadnigk, G. Curio , "Analyzing Speech Quality Perception Using Electroencephalography," *IEEE J. Select. Topics Signal Proc.*, vol. 6(6), pp.721-731, 2012.

[7] S. Scholler, S. Bosse, M. S. Treder, B. Blankertz, G. Curio, K. R. Müller, T. Wiegand, "Towards a Direct Measure of Video Quality Perception using EEG" *IEEE Trans Image Process*, vol. 21(5), pp.2619-2629, 2012.

[8] S. Arndt, J.-N. Antons, R. Schleicher, S. Möller, G. Curio. "Perception of low-quality Videos analyzed by means of Electroencephalography", Fourth International Workshop on Quality of Multimedia Experience (QoMEX), AUS-Yarra Valley, pp.284-289, 2012.

[9] M. Mustafa, S. Guthe, M. Magnor, "Single Trial EEG Classification of Artifacts in Videos" *ACM Trans on Applied Perception (TAP)*, vol. 9(3), pp.1201-1215, 2012.

[10] S. Koelstra, C. Muehl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, I. Patras, "DEAP: A Database for Emotion Analysis using Physiological Signals" *IEEE Transaction on Affective Computing*, vol. 3(1), pp. 18–31, Jan.-March 2012.

[11] T. Halmrast, "Sound coloration from (very) early reflections", in Proc. Meeting Acoust. Soc. Amer., 2001.

[12] P. Rubak, "Coloration in room impulse responses", in Proc. Joint Baltic-Nordic Acoust. Meeting, 2004.

[13] T. Falk, C. Zheng, W.-Y. Chan, "A Non-Intrusive Quality and Intelligibility Measure of Reverberant and Dereverberated Speech", *IEEE Transaction on Audio, Speech and Language Processing*, vol. 18(7), 2010.

[14] "Objective Measurement of Active Speech Level", ITU-T Recommendation P.56, International Telecommunication Union, Geneva, 2011.

[15] P. J. Lang, "Behavioral treatment and bio-behavioral assessment: computer applications", in J. Sidowski, J. Johnson, and T. Williams (Eds.), "Technology in mental health care delivery systems", pp. 119-l37, NJ, 1980.

[16] T. Falk, Y. Pomerantz, K. Laghari, S. Möller, T. Chau, "Preliminary Findings on Image Preference Characterization based on Neurophysiological Signal Analysis: Towards Objective QoE Modelling", Fourth International Workshop on Quality of Multimedia Experience (QoMEX), AUS-Yarra Valley, pp.146-147, 2012.