Neural correlates of Affective States and Speech Perception: Towards Quality-of-Experience Measurement of Reverberant Speech

Jan-Niklas Antons[1], Khalil ur Rehman Laghari [2], Robert Schleicher[1],
Sebastian Arndt[1], Tiago H. Falk[2], and Sebastian Möller[1]
[1] *Quality and Usability Lab, Berlin Institute of Technology, Germany*
[2] *INRS-EMT, University of Quebec, Canada*

Abstract

Speech quality assessment usually depends on subjective judgments after listening to test stimuli. The obtained subjective quality indices are valid and reliable but provide little insight into the underlying perceptual process. Since reverberation is known to influence perceived speech quality and intelligibility, e.g. in conference calls while using a loudspeaker, we analyzed the performance of electroencephalography (EEG), which measures brain activity at the cortical level, for indicating speech stimuli with a high reverberation time as degradation. We collected a database of 22 subjects to test the ability of utilizing EEG-data - especially parameters of event-related-potentials (ERP) - to identify the processing of a stimulus with high reverberation times. The reported preliminary findings provide promising insights: indirect measurements of perceived stimulus quality - without asking for the subjects' opinion - are sensitive to reverberation levels. Correlations between physiological parameters and stimulus features showed that quality degradations can be monitored in conscious stages of stimulus processing. We show that the analysis of ERP is in general a useful and valid tool in quality research. In the case of reverberation, this can actually lead to the indirect measurement of perceived quality with respect to changes of room acoustics.

Introduction

Jekosch [1] suggests that speech quality assessment comprises a three-step process: perception, judgment, and description. The first step includes the reception and perception of an auditory event (i.e., speech sound wave reaches the human ear). In the second step, the judgment process, features from the perceived "event" are extracted and compared to internal reference features of what good and bad quality speech sounds like. This internal reference can possibly be influenced by several factors, such as. user experience and affective states, thus are individual to each listener. During the third and final step, the description process, a comparison of incoming and internal reference features are grouped into a final overall quality rating. The results of the final step are usually captured by subjective listening tests for which ratings are averaged over several subjects. One representative method is the absolute category rating (ACR) test resulting in a Mean Opinion Score (MOS)[2]. For the judgment and description process cognitive models are not yet fully understood. Nevertheless, based on the advances in neuroimaging technologies, there is growing interest in investigating the neuro-physiological correlates of human speech quality perception. The electroencephalography (EEG) has been an effective method 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 particularly a useful method [5] for measuring cognitive processing. For example, [6] showed that simple degradations in the speech signal (e.g., multiplicative noise) could be processed by humans at an unconscious level. Moreover, our recent results have also shown that thejudgment and description processes contribute to the understanding of internal processes [7].

Traditionally, auditory perceptual models are used to emulate the human psychoacoustic properties and the speech perception processes. Commonly, so-called auditory filterbanks (e.g., the gammatone filterbank) are used, as they have been developed based on research performed by physiologists who measured the impulse response of the auditory filter in small mammals. For example, the amplitude characteristic of the gammatone function can predict human masking data [8]. Despite their widespread use in speech and audio signal processing, auditory models as a correlate of speech quality perception have not played a significant role in Quality-of-Experience research. Ultimately, this knowledge of "perception," combined with the "judgement" and "description" processes obtained via neurophysiological monitoring could lead to more accurate quality-assessment process, improved subjective testing protocols, objective quality models, and speech-based technologies.

Motivated by the abovementioned promising findings, this paper aims to investigate the use of EEG to obtain neural correlates of speech quality variations in complex listening environments, as well as compare the obtained cognitive and subjective quality perception insights to psychoacoustic parameters extracted from an auditory-inspired filterbank. This builds on our previous findings where reverberation time, as a single parameter was correlated with subjective and physiological data.

Materials and methods

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).

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.

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. 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.

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. The reaction time was also computed for each presented stimulus. It is defined as the time between the stimulus onset and the actual button press.

Auditory filter

For the computation of the filter coefficients, we used a 1-ERB (equivalent rectangular bandwidth) spaced filter bank based on gammatone filters. This resulted in 25 band-pass filters with the center frequencies at: 84, 119, 159, 203, 253, 307, 369, 437, 513, 597, 692, 797, 913, 1044, 1189, 1350, 1531, 1731, 1955, 2204, 2481, 2790, 3134, 3517 and 3944 Hz. For each stimulus (clean, reverberation time of 400 *ms* and reverberation time of 1500 *ms*), we extracted the power value for the first 200 *ms* of each frequency band, resulting in 25 power values per stimulus. For analysis we calculated the correlations per subject and averaged the values of significant correlations using Fischer z-transformed values.

Experimental Results

First we briefly mention the already reported results from [7] and based on these findings, we will present the results using the auditory filter features. To analyze this data we performed a Pearson's linear correlation between the power within the frequency bands of the auditory filter and the P300 peak amplitude, as well as the MOS and SAM scales (valence, arousal, dominance) and the liking and familiarity ratings.

Qualitative, 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. A monotonic decrease in MOS was observed as reverberation time increased. 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 ($F_{(2,16)} = 9.00$, $p < .01$, $Eta^2 = .52$). 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 ($F_{(2,16)} = 22.07$, $p < .01$, $Eta^2 = .73$); monotonically decreasing curves were also observed with increasing reverberation time. 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 which indicate that affective states, quality perception, and Quality-of-Experience (QoE) are strongly related parameters.

Neural/Cognitive Correlates

A significant main effect was observed between P300 peak amplitude and reverberation time ($F_{(2,16)} = 8.15$, $p < .01$, $Eta^2 = .50$). P300 amplitude increases with an increase in reverberation time. Significant negative correlation was attained with MOS and the valence dimensions. Lastly, a significant main effect with reverberation time was also observed for reaction time ($F_{(2,16)} = 11.73$, $p < .01$, $Eta^2 = .59$).

Auditory Filter Correlates

It was observed that for almost all filter bands significant correlations with the P300 peak amplitude could be found, at least for one subject each (see Table 1). For MOS, arousal, valence, liking and familiarity significant correlations could be found mainly for frequency bands 9 to 16, a mid-frequency range. For the dominance scale only a few correlations could be found for frequency band 1, 2, 13, and 16.

Auditory filter bands	#	Center frequency in Hz	Cognitive/subjective variables													
			P300		MOS		Arousal		Valence		Dominance		Liking		Familiarity	
			R	N	R	N	R	N	R	N	R	N	R	N	R	N
	1	84	-0,99*	3							0,87**	1				
	2	119	-0,99*	3							0,86**	1				
	3	159	-0,99*	3												
	4	203	-0,99*	3												
	5	253	-0,99*	2												
	6	307	-0,99*	2												
	7	369	-0,99*	3												
	8	437	-0,99*	3												
	9	513	-0,99*	3			-0,99**	1							0,99**	3
	10	597											0,99**	6	0,99**	2
	11	692	-0,99*	1	0,99**	15			0,99**	3					0,99**	3
	12	797					0,99**	1	0,99**	3			0,99**	4	0,99**	4
	13	913	0,99*	1	0,99**	13	0,99**	9	0,99**	4	-0,8**	1	0,99**	5	0,99**	4
	14	1044	-0,99*	1	0,99**	15	-0,99**	1	0,99**	5					0,99**	1
	15	1189	-0,99*	1			-0,99**	1	0,99**	2					0,99**	2
	16	1350	-0,99*	1							0,96**	1	0,99**	3		
	17	1531	-0,99*	2												
	18	1731	0,99*	1												
	19	1955	0,99*	1												
	20	2204	0,99*	2												
	21	2481	-0,99*	2												
	22	2790	-0,99*	2												
	23	3134	-0,99*	2												

Table 1: Correlation matrix of the power of auditory filter bands (0-200 ms of stimulus) with P300 peak amplitude and subjective ratings (MOS, SAM scales, liking and familiarity). With N as number of subjects considered for each correlation (**: $p < 0.05$ and *: $p < 0.10$). Auditory filter bands without significant correlation are not displayed.

Discussion and Conclusions

This study investigated the effects of increasing reverberation levels (time) 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.

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.

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.

Lastly, it was observed that for almost all auditory filter bands correlations with the P300 peak amplitude were existent at least on single subject basis. In contrast, we could show that for the qualitative and emotional ratings the correlations were dominant within the mid-range frequency bands. This could be caused by the fact that neural activation is based on a variety of features spread over all frequency bands and in addition this frequency bands are not similar for all subjects. For the subjective ratings the most dominant features are in the frequency range of 500 to 1350 Hz.

This study has explored cognitive, affective, and experiential factors inherent to humans when asked to perform a listening speech quality assessment task. Focus was placed on quality-of-experience (QoE) assessment of reverberant speech and the inter-correlation with auditory bands. It is expected that the obtained results may lead to improved room acoustic characterization algorithms and subjective listening tests.

Acknowledgement

The authors are grateful to the colleagues from the Centre for Research on Brain, Language and Music (CRBLM) 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.

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, "Eventrelated 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] J.-N. Antons, K. Laghari, S. Arndt, R. Schleicher, S. Möller, D. O'Shaughnessy, and T. H. Falk, „Cognitive, affective, and experience correlates of speech quality perception in complex listening conditions," in ICASSP, 2013.

[8] R. Patterson, I. Nimmo-Smith, J. Holdsworth, P. Rice. "An efficient auditory filterbank based on the gammatone function." *APU report*, 2341, 1988.

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