# The Effects of Text-to-Speech System Quality on Emotional States and Frontal Alpha Band Power

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]

*Abstract*— **The tolerance limit for acceptable multimedia quality is changing as more and more high quality services approach the market. Thus, negative emotional reactions towards low quality services may cause user disappointment and are likely to increase churn rate. The current study analyzes how different levels of synthetic speech quality, obtained from different text-to-speech (TTS) systems, affect the emotional response of a user. This is achieved using two methods: subjective, by means of user reports; and neurophysiological by means of electroencephalography (EEG) analysis. More specifically, we analyzed the frontal alpha band power and correlated this with the subjective ratings based on the Self-Assessment Manikin scale. We found an increase in neuronal activity in the left frontal area with decreasing quality and argue that this is due to user disappointment with low quality TTS systems as they become harder to understand.**

## I. Introduction

As synthesized speech is being used more and more in everyday's life and now also is an inherent part of smartphones, the quality of these systems needs to be on an acceptable level in order to understand the text-to-speech (TTS) systems easily on the one hand and to have a pleasant interaction with the device on the other hand. A TTS system in the context of smartphones acts often as an information system and thus conveys mostly neutral content. This paper will examine whether the quality of an information giving system has an influence on the users' perception.

In order to better understand neuronal information processing Electroencephalography (EEG) is being used. While event-related potentials (ERP) give information about short term changes, and are often used on short stimuli, frequency analysis of EEG signals can give insights to long term effects in relation with longer lasting stimuli. However, these extracted features are still being more specific to information processing than e.g. skin conductance measurements. With help of ERPs not only obvious changes in the presented material but also subtle changes can be quantified. This relationship was deployed to investigate differences in stimulus material concerning their quality. A few studies using short stimuli and varying them in their quality evoked ERPs of different amplitudes, more specifically the worse the quality was the higher the P300 amplitude was [1] [2]. Additionally, longer stimuli were used and showed a change in different cognitive state information, e.g. fatigue while subjects were presented with different levels of quality [3].

Besides indicating these effects, EEG also is a possible method to assess information about the emotional state of users in general, and the change of emotional arousal between two conditions in particular. As shown in several studies where subjects were exposed to audio and video stimuli their emotional state was captured using EEG [4]. Often the frontal asymmetry in alpha band activity can be used to assess this. The basic principle here is an inverse relationship between the alpha band power and the underlying neural activity (i.e. the higher the alpha band power is, the less active the underlying neuronal network becomes) [5]. In general it is assumed that more neuronal processing in the left frontal area, hence less alpha band power, is associated with positive emotions and more activity in the right frontal hemisphere with negative emotions [6]. This is supported by several studies, e.g. [7] who used musical excerpts with different levels of valence which in the end correspond to the measured alpha portions in the EEG signal. However, two approaches to interpret this asymmetry still hold: the emotional valence and the motivational direction, these may in parts result in opposite results. Especially anger and frustration respectively, by itself a negatively valanced emotion, results in larger activity in the left frontal area [8].

A number of studies used EEG signals to confirm and predict emotional content from auditory signals. These studies often rather use music than speech, because music elicits strong emotional responses [9]. To the best of the authors knowledge, no research has been done on the effects of synthesized speech quality on affective states and EEG activity.

The remaining of the paper is structured as follows, in Section II the experimental setup will be described followed by reporting the obtained data in Section III. In Section IV a discussion based on the results will be done before closing with a conclusion paragraph in Section V.

## II. Experimental Setup

In this section, a brief description of the used stimuli, paradigm, and equipment will be given.

[1] Quality and Usability Lab, TU Berlin, Germany sebastian.arndt@telekom.de

[2] INRS-EMT, University of Quebec, Canada

This work was co-funded by the German Academic Exchange Service (DAAD) and the German Research Foundation (DFG - 1013 Prospective Design of Human-Technology Interaction). Thanks goes also to the Canadian Foundation for Innovation and the Ministere de lEducation, du Loisir et du Sport du Quebec for funding the experimental facility and the Ministere du Developpement Economique, Innovation et Exportation du Quebec and the National Science and Engineering Research Council of Canada for funding the pilot studies. This was also co-funded by the the Bernstein Focus: Neurotechnology Berlin (BFNT-B) which was supported by the Federal Ministry of Education and Research (BMBF) FKZ 01GQ0850.

### A. Synthesized Speech Stimuli

The synthesized speech stimuli used in this study were taken from the Blizzard Challenge 2009 [10], a competition for TTS systems. More specifically, two different synthetic speech systems plus the corresponding natural voice were chosen. The data used in our experiment comprised four sentences, spoken in English, and synthesized by the two systems (plus the natural speech counterpart). The two systems chosen correspond to two levels of speech quality, as obtained via a mean opinion score (MOS) test, where lower ratings indicate lower quality. The high-quality (HQ) TTS system obtained an average MOS of 3.7 (out of 5) in the Blizzard Challenge, whereas the low-quality (LQ) systems obtained average MOS ratings of 1.9. The speech stimuli had a duration between 8-10 seconds and were digitized using 16-bit resolution and 16 kHz sampling rate.

### B. EEG Signal Acquisition and Analysis

A 64 channel ActiveII EEG system from Biosemi was used with electrodes positioned 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 according to the 10-20 international standard system [11]. For the reference electrode we used both mastoids. Recordings were done with a sampling rate of 2000 Hz, no online filtering was applied.

For the EEG analysis, the MATLAB toolbox EEGLAB [12] was used. The data was down-sampled to 200 Hz and a band-pass filter from 0.1 to 60 Hz was applied before further analysis. Trials with large artifacts were manually removed. For the frequency analysis a fast Fourier transform algorithm was applied to the EEG data. Then the spectral power density estimate was calculated by using Welch's method. For the analysis in this paper the range of the alpha band, from 8 to 13 Hz, was taken into consideration. In order to avoid any movement artifacts in the analyzed data, only the timespan between 2 and 7 s after stimulus onset was considered.

### C. Subjective Quality and Affective State

Prior to the EEG recording, a subjective quality and affective state test was performed. During this test, subjects had to rate all stimuli on several subjective scales. Among them the Self-Assessment Manikin scale (SAM) [13] was used. Here the arousal and valence dimensions were asked on a nine point continuous scale. Additionally, a rating for the overall quality of the stimulus was obtained using the 5-point MOS scale [1-bad, 2-poor, 3-fair, 4-good, 5-excellent] as suggested by ITU-T Recommendation P.85 [14] and the level of comprehension was assessed (higher values indicate better comprehension).

### D. Participants

Fourteen subjects (6 male, 8 female) with an average age of 21.6 years conducted the test. All of them were fluent English speakers. None of them reported any hearing impairment or other health issues. In addition to the subjective test above, participants were also asked to rate after each stimulus whether they felt the voice was pleasant to them or not (Buttons labeled 'Pleasant', 'Unpleasant') while donning the EEG cap. Stimuli were played via in-ear headphones at a preferred individual volume. The test protocol was approved by the Research Ethics Office.

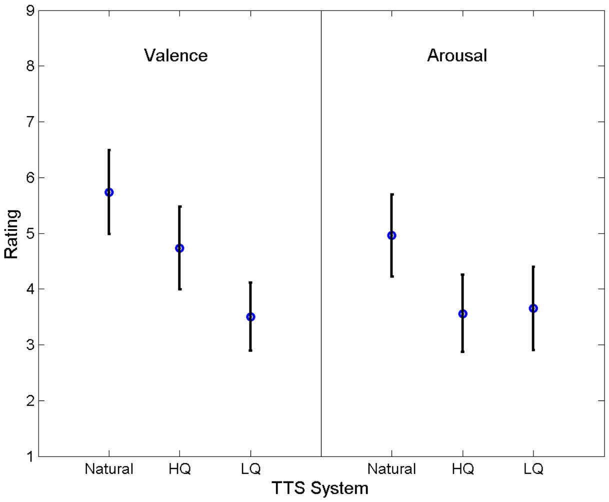

Fig. 1. Obtained SAM ratings of valence and arousal averaged over all subjects.

## III. Experimental Results

In this section the effect on the subjectively obtained emotional ratings with respect to the selected stimuli will be reported. Also if and how these were affecting the EEG recordings will be analyzed in detail. Additionally, possible relations between the two methodologies will be examined.

### A. Subjective Data Analysis

As shown in Figure 1 the subjectively obtained valence score is decreasing as the quality of the TTS systems decreases. The data of the valence scale has a significant main effect when calculating an ANOVA (F(2,22) = 28.37, $p \leq 0.01$) with quality as independent variable. A post-hoc analysis also yields significance for all quality levels. Next, we analyzed the arousal ratings. As can also be seen from the right-most plot in Figure 1, arousal ratings were higher for the natural speech stimuli. There was a significant drop in arousal levels from natural to HQ an then only slight changes between the two different TTS conditions. Overall, a statistical main effect for the factor arousal with the independent variable quality (F(2,22) = 14.48, $p \leq 0.01$) was found. A post-hoc test with Bonferoni corrected pair-wise comparison showed statistical significance between natural speech and the two TTS systems, however no significance between the TTS quality levels. For the qualitative part higher MOS as well as higher comprehension scores were achieved for better qualities of TTS systems.

Next, we explored the ratio of trials rated as pleasant by the participants to the total number of presented trials. This ratio parameter serves as indicator of the listener's perceived experience and acceptance with the different TTS systems. As can be seen from the plot in Figure 2, the ratio

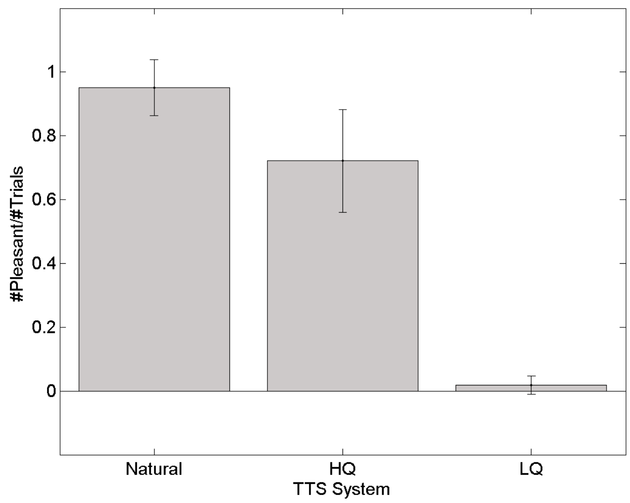

Fig. 2. Ratio of pleasant and unpleasant marked trials for each condition, averaged over all subjects.

|  | AF3 | AF4 | Index |
|---|---|---|---|
| **Unpleasant** | 1.74 | 1.98 | 0.084 |
| **Pleasant** | 1.84 | 1.97 | 0.054 |

TABLE I

GRAND AVERAGE OF FRONTAL ALPHA ACTIVITY WHILE DIVIDING EEG DATA INTO PLEASANT AND UNPLEASANT TRIALS.

decreases monotonically as the quality level decreases. An ANOVA with quality as independent and ratio as dependent variable resulted in significant differences (F(2,18) = 101.32, $p \leq 0.01$). Due to technical difficulties the result of this additional rating could not be obtained from two participants.

### B. *Physiological Data Analysis*

As literature suggests, a frontal asymmetry due to emotions can be expected thus we examined this possible effect. Therefore, the alpha band power for each stimulus level was extracted from electrodes AF3 and AF4. Then an asymmetry index was computed by subtracting the natural logarithm of the alpha power of the two electrodes ($ln(\alpha_{AF4}) - ln(\alpha_{AF3})$) as suggested in [6]. This index describes the ratio of the alpha band power distribution between the two frontal electrodes (AF3 and AF4). As can be seen from Figure 3, the index increases as quality decreases. This effect is also statistically significant when calculating a repeated measured ANOVA (F(2,24) = 4.91, $p \leq 0.05$), thus suggesting increased neuronal activity in the left frontal region for lower-quality TTS systems.

Lastly, we computed the asymmetry index separately for trials classified as pleasant and unpleasant by the participants. As shown in Table I, the left frontal hemisphere (AF3) has a higher power of alpha for the pleasant trials than for the unpleasant trials, whereas no significant changes are observed in the right frontal hemisphere. This is corroborating the findings depicted by Figure 3 as TTS quality decreases (and becomes more unpleasant) the index is increasing.

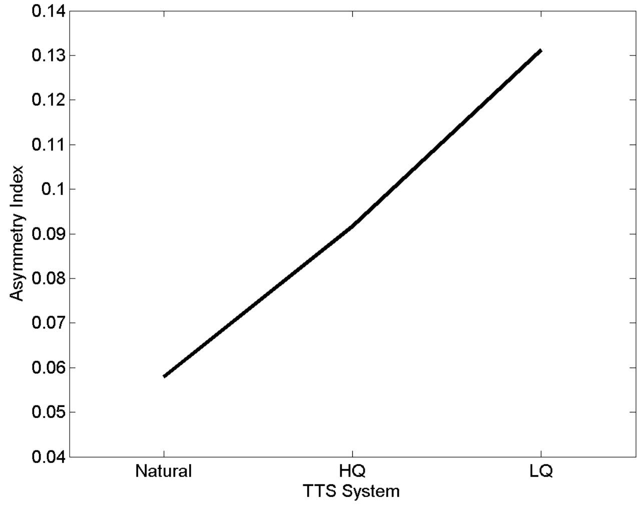

Fig. 3. Asymmetry index, computed by $ln(\alpha_{right}) - ln(\alpha_{left})$ averaged over all subjects for each stimulus level.

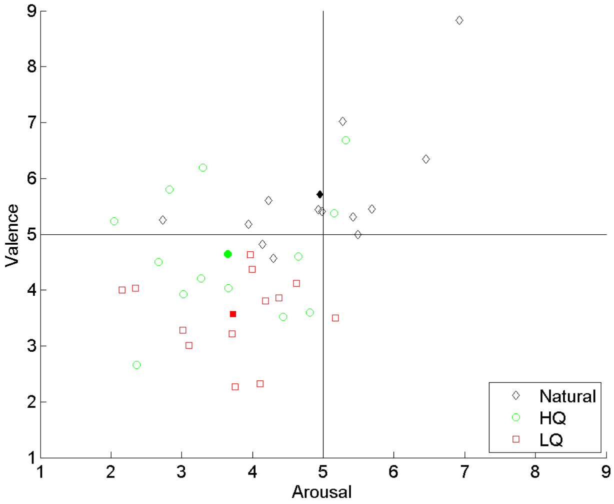

Fig. 4. Arousal Valence Space. Scores plotted for each subject, filled-out markers depict averaged values over all subjects.

### C. *Relationship between Subjective Ratings and Physiological Data*

To investigate the relationship between the subjective ratings and neurophysiological data, correlations were computed between the EEG alpha band powers and the arousal from the subjective ratings. This yielded a correlation of $r = -0.29$ ($p = 0.07$) for the alpha power of electrode AF4.

## IV. DISCUSSION

When inspecting the subjective data we found that there is no significant difference between the TTS systems on the arousal scale. This may be due to several facts: The played stimuli were all neutral in content (since it was a response of a restaurant recommendation system). The intonation of the speaker also was neutral. This leads to the suggestion that purely on these impressions no level of arousal could have been determined. Thus, the difference in rating was only due to the quality of the TTS system. Another fact which might contribute towards this is that all invited subjects were naïve listeners. For them a rating on an arousal scale using common emotional stimuli might be even difficult enough, but having only these neutral stimuli certainly affected their insecure rating procedure. This was different for the rating of the valence scale since this is a less abstract scale. This is confirmed by that the results from this scale are nicely in line with the pleasantness ratings during the EEG experiment. When spanning a 2 dimensional space for the arousal valence rating a clustering of the different qualities can be observed (see Figure 4). The mean of these clusters corresponds to the emotional dimensions mentioned in [15]. Here, natural speech is mapped to contentment and LQ to disappointment; HQ is situated right in the middle of the two. This suggests that the emotion space mapping is working as well when we manipulate the quality of the stimulus which has to be rated and do not change any obvious criteria concerning emotions.

Analyzing the frontal hemispheres with respect to the alpha band power an increase of the corresponding index could be observed. This increase is due to more alpha band power in the right frontal area, thus less activity in the right hemisphere but more activity in the left hemisphere. Still, current research argues whether frontal asymmetry is due to emotional valence or motivational direction [8] since studies show also greater left activity can be due to anger [16] [17]. The results in the current study indicate that rather the motivational approach is the appropriate one for this study, because the lower quality makes subjects rather disappointed because they have difficulties to understand what is being said by the synthetic voice, thus greater left frontal activity is the result. The index increase with lower quality also is present when dividing the EEG data into pleasant and unpleasant trials, here the index for unpleasant trials is higher, which corresponds to the former finding. Another reason for increasing activity in the left hemisphere with decreasing quality is that TTS systems with lower quality need more processing capacity due to more complex speech processing. This is supported by the fact that language processing happens predominately in the left hemisphere for right handed people (all subjects were right handed).

When combining the data of the two measurements a negative correlation between the alpha band power and scored arousal was shown. While arousal in general elicits neural activation, there should be a negative relationship between arousal and the obtained power of the alpha frequency band, so that the scored arousal level is higher with lower portions of alpha waves in the EEG signal. We tried to replicate this with our obtained data.

## V. Conclusion

Using EEG for assessing Quality of Experience (QoE) already turned out to be a valid methodology in previous experimental setups. This study examined whether EEG can be used to specifically assess the emotional aspects of QoE. Therefore the overlap between quality, emotion and EEG research paradigms was used. This study used synthetic speech samples of different quality levels and examined the emotional affect on the listener. Hereby it could be observed that lower quality TTS systems result in higher left frontal activity which in this case might be due to disappointment, i.e. the listeners get more frustrated about the presented quality and have difficulties to understand it. This was confirmed while using the subjectively obtained arousal and valence dimension. To confirm this argumentation future work should additionally conduct a scale for the level of frustration or disappointment.

## References

[1] J. Antons, R. Schleicher, S. Arndt, S. Möller, and G. Curio, *Selected Topics in Signal Processing, IEEE Journal of, title=Analyzing Speech Quality Perception Using Electroencephalography*, vol. 6, no. 6, pp. 721 –731, 2012.

[2] S. Arndt, J. 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.

[3] J. 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.

[4] 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.

[5] A. Gevins, M. Smith, L. McEvoy, and D. Yu, "High-resolution EEG mapping of cortical activation related to working memory: effects of task difficulty, type of processing, and practice." *Cerebral Cortex*, vol. 7, no. 4, pp. 374–385, 1997.

[6] J. A. Coan and J. J. Allen, "Frontal EEG asymmetry as a moderator and mediator of emotion," *Biological psychology*, vol. 67, no. 1, pp. 7–50, 2004.

[7] L. A. Schmidt and L. J. Trainor, "Frontal brain electrical activity (EEG) distinguishes valence and intensity of musical emotions," *Cognition & Emotion*, vol. 15, no. 4, pp. 487–500, 2001.

[8] E. Harmon-Jones and J. Sigelman, "State anger and prefrontal brain activity: evidence that insult-related relative left-prefrontal activation is associated with experienced anger and aggression." *Journal of personality and social psychology*, vol. 80, no. 5, p. 797, 2001.

[9] A. Gabrielsson, *Emotions in strong experiences with music*. New York Oxford University Press, 2001.

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

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

[12] A. Delorme and S. Makeig, "EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis," *Journal of neuroscience methods*, vol. 134, no. 1, pp. 9–21, 2004.

[13] M. Bradley and P. 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.

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

[15] R. Plutchik, "The Nature of Emotions Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice," *American Scientist*, vol. 89, no. 4, pp. 344–350, 2001.

[16] E. Harmon-Jones, "On the relationship of frontal brain activity and anger: Examining the role of attitude toward anger," *Cognition and Emotion*, vol. 18, no. 3, pp. 337–361, 2004.

[17] E. Harmon-Jones, P. A. Gable, and C. K. Peterson, "The role of asymmetric frontal cortical activity in emotion-related phenomena: A review and update," *Biological psychology*, vol. 84, no. 3, pp. 451–462, 2010.