Using fNIRS to Characterize Human Perception of TTS System Quality, Comprehension, and Fluency: Preliminary Findings

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

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

Abstract

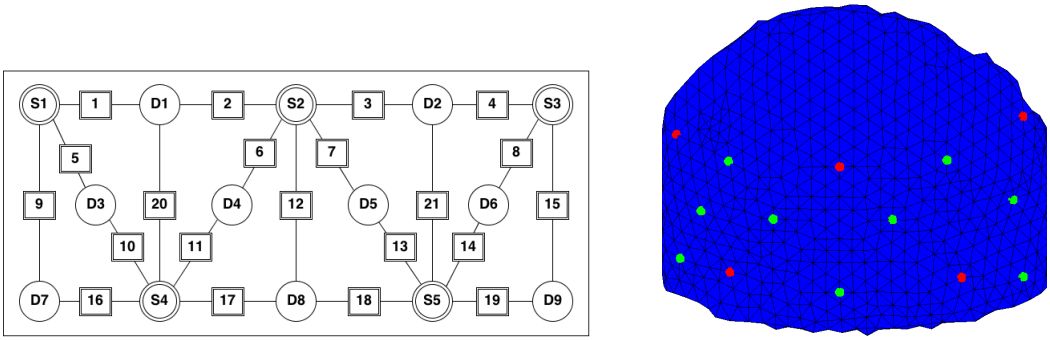
The quality of synthesized speech signals from different Text-to-Speech (TTS) systems is traditionally evaluated using subjective tests based on user ratings. Subjective testing, however, is challenging due to the variability and complexity of human perception. As such, recently there has been a shift towards exploring new objective techniques to evaluate the quality of TTS systems. In this paper, we describe our initial effort of characterizing human TTS quality perception via neurophysiological insights obtained from a neuroimaging technology called functional Near Infrared Spectroscopy (fNIRS). This approach allowed for a link between the human decision making process and the quality of different TTS systems to be established. We showed significant correlations between perceived quality and several fNIRS features related to cerebral haemodynamics. These preliminary results have helped establish the potential of fNIRS as an important tool for evaluating the quality of TTS systems.
Index Terms: fNIRS, TTS, quality measurement, Quality of Experience (QoE), fluency, comprehension

1. Introduction

Text-to-Speech (TTS) systems have seen a tremendous upsurge in recent years with new applications emerging for computers, smartphones, global positioning systems, and assistive technologies (e.g., for the visually impaired), to name a few. In order to develop TTS systems that will be widely accepted, care has to be exercised such that high-quality and intelligible speech is produced. To this end, Quality of Experience (QoE) has emerged as benchmark for user-centric quality evaluation. Usually, subjective tests are performed for QoE based perceptual speech quality evaluation. The International Telecommunications Union (ITU-T), for example, has recommended multi-dimensional subjective listening tests that involve listeners rating the quality of synthesized speech using eight quality dimensions labeled: overall impression, listening effort, comprehension problems, articulation, pronunciation, voice pleasantness, speaking rate, and acceptance. The ITU-T P.85 is the recommended standard for the subjective quality evaluation of synthesized speech [1]. Commonly, the 5-point Absolute Category Rating ('1' = Poor,...,'5' = Excellent quality) is used to obtain a TTS system's Mean Opinion Score (MOS). Recent research, however, has shown that such subjective tests often lead to inconsistent results, likely due to two factors: (1) limitations of the utilized rating scales, and (2) the listener's psychological/behavioural biases to the different acoustic characteristics of the presented TTS stimuli [2]. In order to shed light into these individual biases and their effects on QoE based perceptual speech quality evaluation, recent research has resorted to neuroimaging technologies, such as electro-encephalography (EEG) or functional Near Infrared Spectroscopy (fNIRS) [3].

Recently, EEG derived features such as mismatch negativity (MMN) and event related potentials (ERPs) have been shown useful in obtaining important neural correlates of human speech quality perception [4, 5]. These ERP based technologies, however, require the use of short-duration speech stimuli, thus commonly synthesized *vowels* are used. Performing subjective tests with only vowels is not very conclusive, as different levels of quality degrading artifacts may occur throughout synthesized sentences. In this study, we aim to fill this gap and explore the use of a new burgeoning neuroimaging technology, termed functional near-infrared spectroscopy (fNIRS). Unlike EEG, fNIRS does not require short-duration stimuli, thus can be used to assess the neural correlates of human quality, comprehension, and fluency perception of synthesized speech *sentences*.

Functional NIRS determines the properties of the brain tissue by transmitting near-infrared electromagnetic radiation (650-950 nm wavelengths) through the skull and comparing the intensities of the returning and incident light. During neural activation, metabolic demand results in an increase in cerebral blood flow in certain areas of the brain, thus increasing the regional concentration of oxygenated haemoglobin and decreasing the regional concentration of deoxygenated haemoglobin [6,7]. As the fraction of light absorbed versus the fraction transmitted is dependent on the concentrations of these chromophores, fNIRS can be used to assess haemodynamic



(a) (b)

Figure 1: fNIRS headband optode topology where (a) shows the 21 channels, depicted within squares with 'S' showing the source, and 'D' showing the detector positions and (b) presents the 3-D finite element method (FEM) head model with the source and detector shown in red and green, respectively.

responses in different regions of the brain to provide useful insights into human speech quality perception.

Specifically, the areas of the brain responsible for cognition and decision making can provide useful insights into speech quality perception. One such area is the pre-frontal cortex (PFC), situated in the forehead region. It has been found to be actively involved in cognition [8] and decision making [9]. More distinctly, a region of the PFC called orbito-frontal cortex (OFC) has been found to be activated during decision making tasks [10]. According to the so-called neuro-economics literature, the OFC is also responsible for the valuation and outcome evaluation processes involved in decision making [11–15]. Such findings have motivated our research of probing the PFC/OFC regions to obtain insights into the human speech quality perception processes.

The remainder of this paper is organized as follows: Section 2 describes the materials and methods used for the experiment, Section 3 presents the results of this study and Section 4 provides a brief discussion for the obtained results. Finally, Section 5 presents our conclusions.

2. Materials and Methods

2.1. Subjects

Fourteen fluent English speakers (6 Males, 8 Females) with an average age of 21.6 years were recruited to participate in the subjective listening test. None of them reported having any hearing impairments or other health issues. In-ear headphones were used to play the synthesized speech stimuli at their individual preferred volume. The study protocol was approved by the INRS and McGill Research Ethics Offices and participants consented to participate and were compensated monetarily for their time.

2.2. Synthesized speech stimuli

In order to utilize synthesized speech stimuli representative of existing systems, data from the 2009 Blizzard TTS Challenge were used [16]. The challenge was developed to compare existing corpus-based TTS systems on the same development set. The stimuli comprised four English sentences (neutral in content) of duration 8-10 seconds, corresponding to responses of a restaurant recommendation system. Here, we utilized data from two systems, one that obtained a high quality rating (MOS = 3.7) during the Challenge and the other obtained poor quality (MOS = 1.9). For benchmarking purposes we also used the original "Natural speech" development data. All stimuli were presented to listeners at a sampling rate of 16 kHz and a bitrate of 256 kbps.

2.3. Experimental Protocol

Participants were first fitted with a customized fNIRS headband and then placed in front of a computer screen and asked to rate the speech signals heard across multiple dimensions, namely, their perceived comprehension, fluency, and overall quality, on a scale of '1' to '5'. Next, participants were presented with the 12 stimuli (four sentences, three conditions- natural, high quality (HQ) and low quality (LQ) and were instructed to rate the stimulus as 'pleasant' or 'unpleasant', by pressing a button. This part of the experiment was divided into six blocks, each lasting for about 10 minutes with an inter-stimulus interval of around 20s. This gave enough time for changes in cerebral hemodynamics to return to baseline levels. The stimuli were pseudo-randomized within blocks and subjects. Blocks were randomized between subjects.

2.4. fNIRS Signal Acquisition and Analysis

The NIRScout system from NIRx Medical Technologies was used (probed wavelengths were 760 and 850 nm) with a customized headband. It comprised 5 transmitters

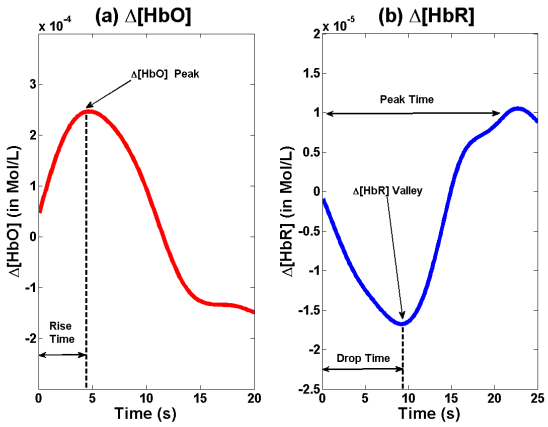


Figure 2: Visual representation of the features computed from the $\Delta[HbO]$ and $\Delta[HbR]$ waveforms.

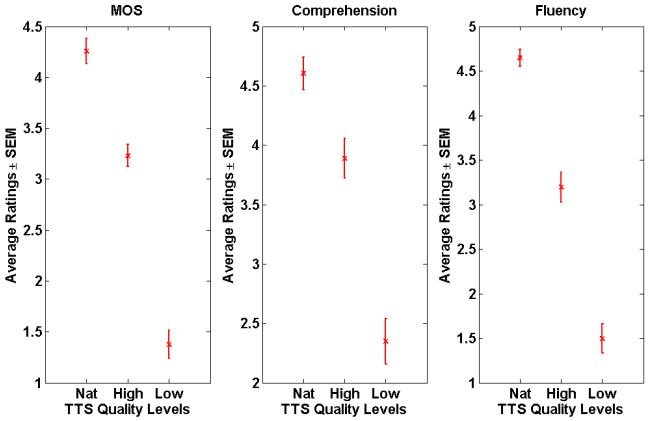


Figure 3: Summary of obtained subjective ratings.

and 9 detectors with a minimum of 2.5 cm and a maximum of 3.4 cm inter-optode distance, thus resulting in 21 functional channels as shown in the optode topology and 3-D finite element head model in Figs. 1(a) and 1(b), respectively. Recordings were made at a sampling frequency of 10.42 Hz. Note that channels 10-11; 13-14, and 16-19 correspond to the OFC region of the PFC.

NIRS data were preprocessed and analyzed using the NIRS-SPM toolbox for MATLAB [17]. The raw intensity signals from each channel were detrended using a discrete cosine transform based algorithm and converted into concentration levels of oxygenated ($\Delta[HbO]$) and deoxygenated haemoglobins ($\Delta[HbR]$) using the well-known modified Beer-Lambert law (MBLL) [18].

2.5. fNIRS Features

In order to characterize the observed changes in the $\Delta[HbO]$ and $\Delta[HbR]$ patterns, five features were extracted from the two detrended waveforms, as depicted by Fig. 2. The features included: peak amplitude of the $\Delta[HbO]$ curve and its corresponding rise time, the amplitude of the $\Delta[HbR]$ curve valley and its corresponding drop time, as well as the curve peak time. The five features were extracted from each of the 21 functional channels for each participant. The $\Delta[HbO]$ peak and $\Delta[HbR]$ valley have been found to be correlated with the Blood Oxygenation Level Dependent (BOLD) signal measured via magnetic resonance imaging, which in turn is positively correlated with regional neural activation [6, 7, 19]. Moreover, to account for the variation in these features, the coefficient of variation was calculated, which is the ratio between the standard deviation and the mean for a particular feature.

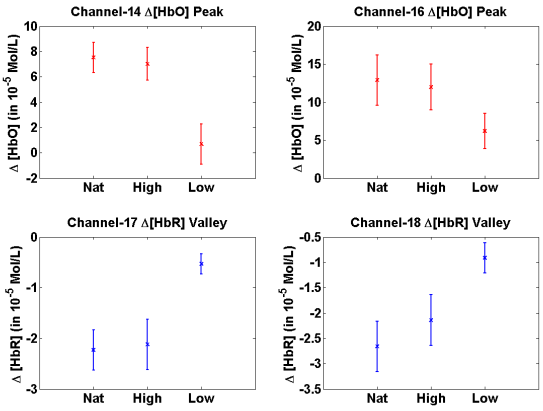


Figure 4: Physiological features: Mean± Standard Error of Mean (SEM).

3. Experimental Results

3.1. Subjective Data Analysis

As shown in Fig. 3, MOS and fluency ratings decreased linearly as speech quality decreased; a non-linear decrease was observed with the comprehension scale. To measure the significance of these differences, a repeated measure within subjects ANOVA was computed using the predictive analytics software SPSS. It compares the effects of three different quality conditions (natural quality, high quality TTS and low quality TTS) on each response variable: MOS, comprehension and fluency. As a prerequisite, Mauchly's test was performed and confirmed the sphericity for all three subjective response variables across three different conditions ($p > 0.05$). The ANOVA results, as reported in Table 1, show that there is a significant main effect in subjective response variables across three different quality conditions. Effect size η^2 shows the strength of the association between subjective factors across three different quality conditions.

3.2. Physiological Data Analysis

As the literature suggested, increased activation of the OFC was expected with increasing perceptual quality of

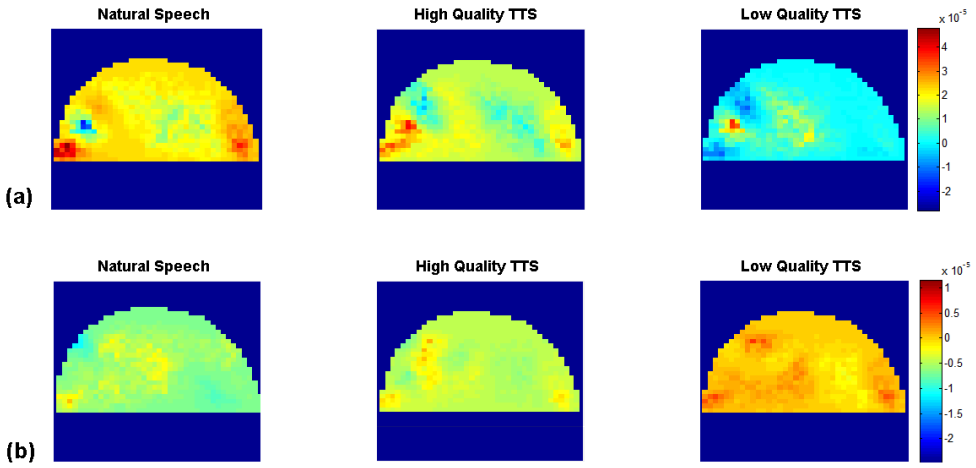


Figure 5: Prefrontal cortex (coronal view) with: (a) $\Delta[HbO]$ peak and (b) $\Delta[HbR]$ valley.

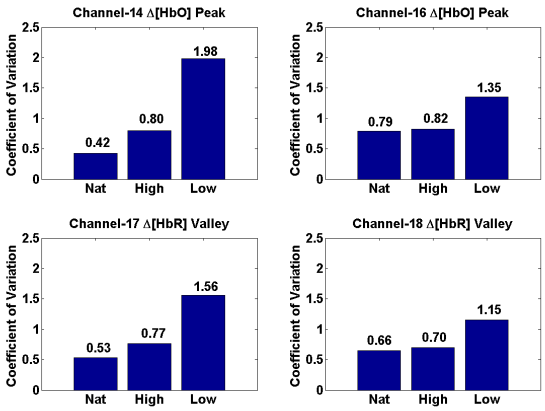


Figure 6: Coefficient of Variation for the Physiological Features.

Table 1: ANOVA for Subjective Measures.

Subjective Measures	p	F(2, 108)	η^2
MOS	< 0.05	120.6	0.70
Comprehension	< 0.05	64.96	0.55
Fluency	< 0.05	112.1	0.67

the stimuli [11]. This effect is clearly visible in Fig. 4, where the amplitudes of the $\Delta[HbO]$ peaks increase and $\Delta[HbR]$ valleys decrease with increasing perceptual quality in the OFC region (channels 14,16,17 and 18), suggesting increased activation of the region. To provide a substantial visualization of this effect, a representative sample of fNIRS data was used to develop a reconstructed image using a Matlab based toolbox called NAVI. The change in two major chromophores in the PFC region of the brain with different quality TTS stimuli is clearly visible in Fig. 5 which shows the coronal view of the PFC.

To provide statistical evidence for the findings from physiological data, a within subjects repeated measures ANOVA followed after the Mauchly's test was carried out. Sphericity was found to be valid for the fNIRS features from the OFC ($p > 0.05$). Under this assumption, the results of ANOVA and linear trend analysis were evaluated. A significant effect of different quality levels of stimuli on the amplitudes of fNIRS features was observed, as reported in Table 2. The 'η^2' was maximum (0.44) for $\Delta[HbO]$ peak for channel 14 located on the right hemisphere of the OFC whereas all other channels showing significant differences had $\eta^2 > 0.30$. This suggests that more than 30% of the variability in the features can be accounted for by stimuli quality.

Table 2: ANOVA for Physiological Features.

Feature	Channel	p	F(df1,df2)	η^2
$\Delta[HbO]$	14	0.01	6.23(2,16)	0.44
Peak	16	0.02	4.70(2,16)	0.37
$\Delta[HbR]$	17	0.01	5.28(2,20)	0.35
Valley	18	0.02	4.53(2,20)	0.31

Moreover, a significant ($p < 0.05$) F-statistic with $\eta^2 > 0.50$ was found for all the fNIRS features from the linear trend analysis as shown in Table 3. This confirmed the linear relationship between the fNIRS features and subjective quality ratings of the stimuli. Also, the centrality of these features was determined by measuring their Mean $\pm$ its Standard Error of Mean (SEM) as shown in Fig. 4, and the spread of the data was measured using the coefficient of variation in the data.

Lastly, the coefficient of variation computed, for all the features, increased with the decrease in quality of the

stimuli as seen in Fig. 6. In order to test the significance of this trend in the coefficient of variation, Levene's test was carried out [20]. The trend in the coefficient of variation for $\Delta[HbO]$ peak for channel 14 was found to be significant with $p < 0.05$. The post-hoc analysis was done using the Tukey's honestly significant difference (HSD) test. The $\Delta[HbO]$ peak for channels 14 and 16 and the $\Delta[HbR]$ valley for channel 17 showed significant ($p < 0.05$) differences between the natural-low and high-low qualities of TTS whereas, $\Delta[HbR]$ valley for channel 18 only showed significant difference between natural-low qualities of TTS.

Table 3: Linear Trend Analysis for Physiological Features.

Feature	Channel	p	F(df1,df2)	η^2
$\Delta[HbO]$	14	0.015	9.43(1,8)	0.54
Peak	16	0.013	9.98(1,8)	0.56
$\Delta[HbR]$	17	0.006	11.80(1,10)	0.54
Valley	18	0.005	13.05(1,10)	0.57

3.3. Relationship between Subjective Measures and Physiological Features

To evaluate the relationship between the subjective scores and fNIRS features, Pearson and Spearman correlation coefficients, denoted by ρ and ρ_{spear} respectively, were used. Moderately high and significant correlations ($p < 0.05$) were found with all the fNIRS features and at least one of the three subjective ratings (MOS, comprehension, and fluency). The $\Delta[HbR]$ valley of channel 17 showed the highest correlation of -0.54 with MOS. $\Delta[HbO]$ Peak of channel 14 and $\Delta[HbR]$ valley of channel 17 showed the highest correlation of 0.58 and -0.59 with comprehension, respectively. These two features were also well correlated with fluency. The features that were significantly correlated with all the three subjective ratings are reported in Table 4.

Table 4: fNIRS Correlates of Subjective Quality Metrics.

Feature	Type	Channel	MOS	Comp.	Fluency
$\Delta[HbO]$ Peak	ρ	14	0.45	0.51	0.42
		16	0.40	0.37	0.42
	ρ_{spear}	14	0.52	0.58	0.43
		16	0.37	0.33	0.35
$\Delta[HbR]$ Valley	ρ	17	-0.54	-0.55	-0.50
		18	-0.32	-0.43	-0.42
	ρ_{spear}	17	-0.54	-0.59	-0.44
		18	-0.27	-0.41	-0.36

4. Discussion

It is quite intuitive that, as the quality of audio stimulus changes from natural to low quality TTS stimuli, subjective ratings for MOS, fluency and comprehension tend to decrease significantly. However, comprehension shows less steep of a decrease as compared to MOS and fluency, probably because subjects could comprehend even low quality speech stimuli, but did not approve of its quality and fluency. To understand the neural basis of this trend, the pre-frontal cortex, more specifically the orbito-frontal cortex was investigated.

In light of the results obtained from the subjective and physiological data, our study has shown a linear increase in the activation of the OFC with a linear increase in TTS quality. This differential activation of the OFC could be attributed to the valuation based on the perceived quality of speech stimuli in the brain, thus corroborating results from previous studies [13]. But no significant difference could be found between the natural and high quality synthetic speech stimuli in the post-hoc analysis. This observation can be attributed to the proximity of the perceptual quality ratings of the two speech stimuli.

Furthermore, the increasing coefficient of variation of fNIRS features with decreasing quality of TTS stimuli, suggests an increase in variability in the value assessment process. Owing to this observation it can be argued that, it becomes more difficult/confusing to assign the lower quality TTS to a particular category (pleasant or unpleasant). This difficulty in decision making or value assessment can also be attributed to the low comprehensibility and fluency of the lower quality TTS stimuli.

Also, a high correlation between the activation of the OFC and the MOS ratings of the speech stimuli provides evidence to the existence of the underlying neurophysiological basis for speech quality perception. In addition, a moderate level of correlation between the neurophysiological features and comprehension as well as fluency ratings of the stimuli indicates that these dimensions contribute significantly towards its valuation. However, a larger correlation of comprehension in comparison to fluency indicates relatively higher contribution of comprehension in the decision making valuation process.

Among the accepted valuation systems, as reported in [11], a goal-directed system could be the one responsible for the valuation of speech stimuli in the human brain. This can be attributed to the fact that it assigns values to the responses based on the action-outcome associations and activates the OFC region of the brain [13]. However, there are indications of existence of deeper located neural systems for valuation [21]. However, due to the inaccessibility of the deeper regions of the brain, which is one of the major limitations of fNIRS, it is not possible to conclusively reject the possibility of a different neural system working in tandem for the valuation process.

5. Conclusion

We have successfully explored the use of functional near-infrared spectroscopy (fNIRS) to obtain insights into the

neural processes involved in TTS system quality, comprehension, and fluency perception. The findings of the study point towards significant correlations between the fNIRS features with the above mentioned percepts. It was found that the OFC located in the PFC of the brain was primarily involved in speech quality perception via value based decision making processes. However, in order to understand the complete neural basis of the human perception of TTS quality, other regions of the brain still need to be investigated.

6. Acknowledgments

The authors are grateful to staff from the Centre for Research on Brain, Language and Music (Montreal) for fruitful discussions and to NIRx Medical Technologies for loaning the NIRS equipment. They also acknowledge funding from the Ministere du Developpement Economique, Innovation et Exportation du Quebec, the National Science and Engineering Research Council of Canada, the German Federal Ministry of Education and Research (Grant FKZ 01GQ0850) and the German Research Foundation (DFG-1013 'Prospective Design of Human-Technology Interaction').

7. References

[1] ITU-T, "P. 85. A Method for Subjective Performance Assessment of the Quality of Speech Voice Output Devices," *International Telecommunication Union, CH-Genf*, 1994.

[2] C. Mayo, R.A.J. Clark, and S. King, "Listeners weighting of acoustic cues to synthetic speech naturalness: A multidimensional scaling analysis," *Speech Communication*, vol. 53, no. 3, pp. 311–326, 2011.

[3] K. Laghari, R. Gupta, S. Arndt, J.N. Antons, R. Schleicher, S. Moller, and T.H. Falk, "Neurophysiological experimental facility for Quality of Experience (QoE) assessment," in *Proceedings of International Conference on Quality of Experience Centric Management (QCMan)*, 2013.

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

[5] K. Laghari, R. Gupta, S. Arndt, J.N. Antons, S. Moller, R. Schleicher, D. O'Shaughnessy, and T.H. Falk, "Auditory BCIs for Visually Impaired Users: Should Developers Worry About the Quality of Text-to-Speech Readers?," in *Proceedings of International Brain Computer Interface (BCI) Meeting*, 2013.

[6] G. Strangman et al., "A quantitative comparison of simultaneous bold fMRI and NIRS recordings during functional brain activation," *Neuroimage*, vol. 17, no. 2, pp. 719–731, 2002.

[7] M. Okamoto et al., "Multimodal assessment of cortical activation during apple peeling by NIRS and fMRI," *Neuroimage*, vol. 21, no. 4, pp. 1275–1288, 2004.

[8] E. Koechlin, C. Ody, and F. Kouneiher, "The architecture of cognitive control in the human prefrontal cortex," *Science*, vol. 302, no. 5648, pp. 1181–1185, 2003.

[9] A. Bechara, H. Damasio, A.R. Damasio, and G.P. Lee, "Different contributions of the human amygdala and ventromedial prefrontal cortex to decision-making," *The Journal of Neuroscience*, vol. 19, no. 13, pp. 5473–5481, 1999.

[10] E.T. Rolls and F. Grabenhorst, "The orbitofrontal cortex and beyond: from affect to decision-making," *Progress in Neurobiology*, vol. 86, no. 3, pp. 216–244, 2008.

[11] A. Rangel, C. Camerer, and P.R. Montague, "A framework for studying the neurobiology of value-based decision making," *Nature Reviews Neuroscience*, vol. 9, no. 7, pp. 545–556, 2008.

[12] A.J. Blood and R.J. Zatorre, "Intensely pleasurable responses to music correlate with activity in brain regions implicated in reward and emotion," *Proceedings of the National Academy of Sciences*, vol. 98, no. 20, pp. 11818–11823, 2001.

[13] S.M. Tom, C.R. Fox, C. Trepel, and R.A. Poldrack, "The neural basis of loss aversion in decision-making under risk," *Science*, vol. 315, no. 5811, pp. 515–518, 2007.

[14] H. Plassmann, J. O'Doherty, and A. Rangel, "Orbitofrontal cortex encodes willingness to pay in everyday economic transactions," *The Journal of Neuroscience*, vol. 27, no. 37, pp. 9984–9988, 2007.

[15] M.P. Paulus and L.R. Frank, "Ventromedial prefrontal cortex activation is critical for preference judgments," *Neuroreport*, vol. 14, no. 10, pp. 1311–1315, 2003.

[16] K. Simon and K. Vasilis, "The Blizzard Challenge 2009," in *Proceedings of Blizzard Challenge Workshop*, 2009.

[17] J.C. Ye et al., "NIRS-SPM: Statistical parametric mapping for near-infrared spectroscopy," *Neuroimage*, vol. 44, no. 2, pp. 428–447, 2009.

[18] D.A. Boas, T. Gaudette, G. Strangman, X. Cheng, J.J.A. Marota, and J.B. Mandeville, "The accuracy of near infrared spectroscopy and imaging during focal changes in cerebral hemodynamics," *Neuroimage*, vol. 13, no. 1, pp. 76–90, 2001.

[19] N.K. Logothetis, "The underpinnings of the BOLD functional Magnetic Resonance Imaging signal," *The Journal of Neuroscience*, vol. 23, no. 10, pp. 3963–3971, 2003.

[20] B. Hallgrímsson and B.K. Hall, *Variation: A central concept in biology*, Academic Press, 2011.

[21] B.W. Balleine, "Neural bases of food-seeking: affect, arousal and reward in corticostriatolimbic circuits," *Physiology & behavior*, vol. 86, no. 5, pp. 717–730, 2005.