Using fNIRS to Characterize Human Influential Factors: Towards Models of Quality of Experience Perception for Text-to-Speech Systems

Rishabh Gupta, Hubert J. Banville, Isabela Albuquerque and Tiago H. Falk

INRS-EMT, University of Quebec, Canada

INTRODUCTION

• With the thriving **speech communications industry**, the global market for speech technology is going through a phase of rapid growth and a recent market analysis report predicts the speech technology market to cross **$31.3 billion** dollars by 2017. Technologies such as, automatic speech recognition (ASR), speaker verification (SV) and **text-to-speech (TTS)** (i.e., synthesized speech) will form the major component of this market.

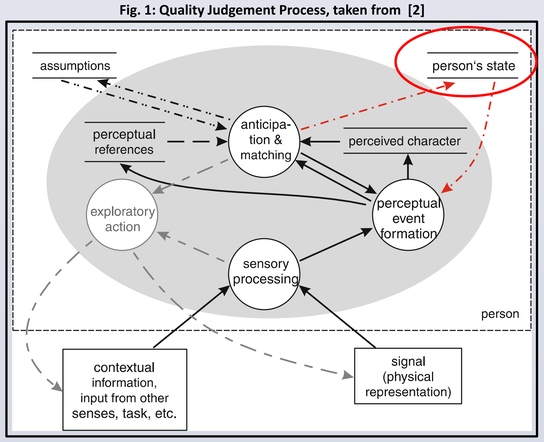
• There is a greater push by the industry and researchers towards evaluating the quality of these technologies through concepts like ***Quality-of-Experience (QoE)***.

• QoE takes a **user centred approach** towards characterising the quality of a product, which ultimately leads to its greater acceptability.

• For most of the last decade, experts have focused on the development of methods for **objective characterisation** of QoE, so as to expedite the process of its quantification.

•A recent expert panel pointed out that existing objective methods lack insights from the so-called **'Human Influence Factors' (HIFs),** which characterize users' emotional or cognitive states, preference, attention etc [1].

Fig. 1: Quality Judgement Process, taken from [2]



• The basic constructs of HIFs are not directly observable and take form inside the users' brain.

• Thus, there is a shift towards probing the brain activity using technologies such as electroencephalography (EEG) or functional near infrared spectroscopy (fNIRS) to:

➢ understand and characterize the HIFs

➢ develop better objective QoE quantification techniques

• In this study, we have used the features derived from oxy /deoxygenated haemoglobin [HbO] or [HbR] concentrations from fNIRS to find neurophysiological correlates of HIFs.

MATERIALS AND METHODS

Participants: Eight fluent English speakers consented to participate in the study.

Stimuli: 16 Natural, 28 Synthesized.

- **16 Natural speech** stimuli = 4 Female speakers X 4 different sentences**. 28 Synthesized speech** stimuli = 7 TTS Engines X 4 different sentences.
- Average duration = 20s.

Experimental Design: Block design

• 15s rest for baseline, 20s stimuli presentation.

• Followed by, scoring the stimuli.

• Rating scales for: user perceived overall quality, **voice pleasantness**, comprehension, emotion, intonation, listening effort, **naturalness** etc.

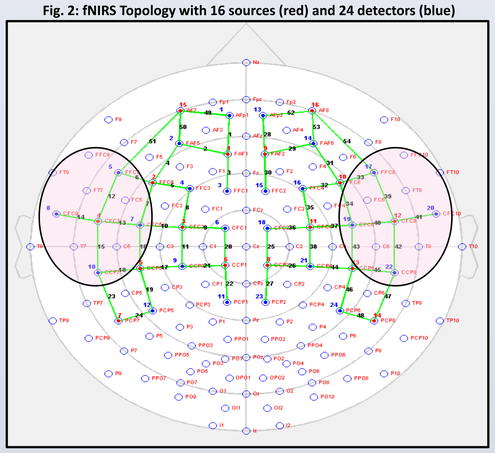
Experimental Setup: NIRx NIRScout System

• 16 Sources and 24 Detectors

• 54 functional channels.

• Wavelengths: 760nm and 850nm

Fig. 2: fNIRS Topology with 16 sources (red) and 24 detectors (blue)

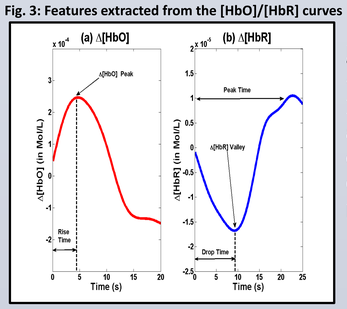


Pre-processing: nilab2 and NIRS-SPM software were used

• Highpass filter (fc=0.001 Hz) to remove slow trends and lowpass (fc=0.2 Hz) to remove cardio-respiratory noise.

• Modified Beer Lambert's Law to convert to [HbO], [HbR].

Fig. 3: Features extracted from the [HbO]/[HbR] curves



Feature Extraction:

• Various features such as HbO peak, HbR valleys and area under the HbO curve (AUC_{HbO}) were extracted from the pre-processed signals.

RESULTS

Subjective Evaluation of Speech

• ANOVA followed by Tuckey's HSD test showed significant differences between perceived pleasantness and naturalness of synthesized and natural speech, and different TTS Engines.

Table 1: ANOVA showing difference between natural and synthesized speech

Subjective Dimension	Mean Sq.	F-statistic	P-value
Voice Pleasantness	284.12	474.01	<0.01
Naturalness	382.49	687.74	<0.01

Fig. 3: Post-hoc Tuckey HSD test: natural and synthesized speech

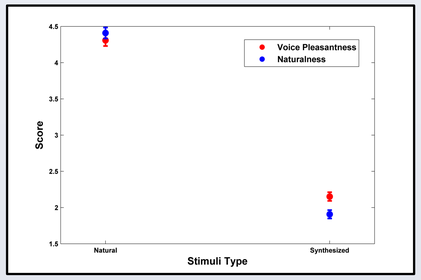
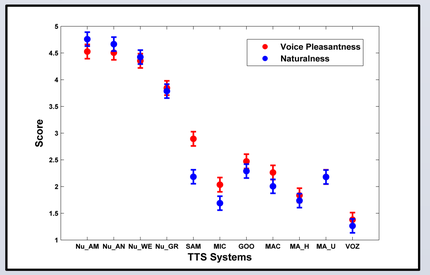


Table 2: ANOVA showing difference between different TTS Engines

Subjective Dimension	Mean Sq.	F-statistic	P-value
Voice Pleasantness	32.44	75.65	<0.01
Naturalness	41.54	101.35	<0.01

Fig. 4: Post-hoc Tuckey HSD test: different TTS Engines



Neurophysiological Correlates

Table 3: Pearson Correlation between the subjective ratings and fNIRS features

Factors	Voice Pleasantness	Naturalness	AUC_{HbO} (Right)	AUC_{HbO} (Left)
Voice Pleasantness	1	0.92	0.32	0.47
Naturalness	-	1	0.40	0.36
AUC_{HbO} (Right)	-	-	1	0.74
AUC_{HbO} (Left)	-	-	-	1

• Four channels located over the temporal areas on both left and right hemispheres were identified (see Fig. 2).

• Features derived from [HbO] and [HbR] curves, averaged over the identified channels, showed significant correlations with subjective ratings.

• Specifically, AUC_{HbO} features computed over the right and left hemispheres were found to be moderately correlated with the subjective ratings (see Table 3).

DISCUSSION

• Synthesized speech has a lower quality as compared to natural speech. Moreover, there is also a significant difference in quality of speech files from different TTS Engines.

• The positive correlation between the subjective ratings and AUC_{HbO} suggests that there is an increase in HbO concentrations in the temporal regions of the brain with better speech quality, higher voice pleasantness, and naturalness. The temporal regions are located close to the auditory cortex.

• This observation concurs with the findings from [3], where the authors found higher activation of the temporal regions in response to natural speech stimuli.

•Also in [4], authors have found higher activations in temporal regions in response to natural speech stimuli, as compared to synthesized speech lacking affective prosody.

CONCLUSION

• The current study demonstrates the ability of fNIRS in encoding information related to users' QoE perception.

• Features extracted from fNIRS recorded over the temporal lobes showed to be useful for QoE perception modeling.

• However, these are some of the preliminary results which have helped us identify the useful features which should be extracted and the regions of the brain which should be probed in order to model user perception of quality.

REFERENCES

1. U. Reiter et al. "Factors Influencing Quality of Experience." *Quality of Experience*. Springer International Publishing, 2014. 55-72.

2. A. Raake and S. Egger, "Quality and Quality of Experience." In *Quality of Experience.* Springer International Publishing, 2014, 11-33.

3. P. Belin et al. "Voice-selective areas in human auditory cortex." *Nature* 403.6767 (2000): 309-312.

4. V. Beaucousin et al. "FMRI study of emotional speech comprehension."*Cerebral cortex* 17.2 (2007): 339-352.