Physiological Quality-of-Experience Assessment of Text-to-Speech Systems

Rishabh Gupta and Tiago H. Falk
INRS-EMT, University of Quebec, Montreal, Canada
Email: falk@emt.inrs.ca

***Abstract*—With the emergence of various text-to-speech (TTS) systems, developers have to provide superior user experience in order to remain competitive. To this end, quality-of-experience (QoE) perception modelling and measurement has become a key priority. QoE models rely on three influence factors: technological, contextual and human. Existing solutions have typically relied on using individual physiological modalities, such as electroencephalography (EEG), to model human influence factors (HIFs). In this paper, we show that fusion of physiological modalities, such as EEG, functional near infrared spectroscopy (fNIRS) and heart rate, provide gains of up to 18.4% relative to utilizing only technological factors and 4% relative to using the best performing individual physiological modality.**

I. Introduction

Quality-of-Experience, defined as 'the degree of delight or annoyance of the user of an application, resulting from the fulfillment of his/her expectations in light of the user's personality and current mental state', is driven by three key influence factors: technological, contextual, and human [1]. Technological influence factors (TIFs) refer to system and network parameters that can be readily measured (e.g., delay, bitrate). Contextual influence factors, in turn, can describe the user's environment, as well as economic aspects (e.g., pricing, churn rate). Lastly, human influence factors (HIFs) characterize the user's perception, emotional and mental state with respect to a service [1]. For much of the last decade, experts have advocated for QoE to be used as the standard user-centric quality metric for emerging applications and products. Notwithstanding, the majority of existing work has focused only on technological and contextual aspects [2]. In order to develop true QoE assessment methods, however, HIFs also need to be incorporated. In this paper, we propose the use of the fusion of multiple physiological modalities (such as EEG, fNIRS and heart rate), during speech QoE perception tests to measure HIFs.

The recent advances in wireless and portable neurotechnologies have lead to the development of a plethora of applications, such as to measure implicit information from the users including their mental states (e.g., stress level), fatigue levels, and more recently, their mood and emotional states [3]. In fact, EEG has proved effective in modelling HIFs, such as affective states [4]. In this paper, we have compared the performances of fusion of multiple physiological modalities and individual modalities for monitoring the human influential factors needed in Quality-of-Experience (QoE) perception models. Here, a case scenario based on text-to-speech (TTS) systems has been explored, as TTS systems have gained tremendous popularity, particularly in the domain of personal digital assistants (e.g., Apple's Siri, Google Now, and Microsoft's Cortana), automated call centres and global positioning systems.

II. Methods and Materials

The database used for the characterization of synthesized speech QoE were obtained from the PhySyQX database [5], which is a publicly-available database for physiological evaluation of synthesized speech QoE. The database consists of multimodal neurophysiological data recorded from 21 healthy participants (8 females, average age = 23.8$\pm$4.35 years), while they experienced 44 synthesized speech stimuli, generated from 7 commercially available TTS systems along with 4 natural voices. The database consisted of EEG and fNIRS signal recorded from the scalp. The EEG signals was recorded from 62 electrodes (AF7 and AF8 were removed), placed on the scalp according 10/20 standard system, at 512 Hz. The fNIRS signals were recorded from 60 channels that were collocated with EEG channels, as described in [5], at 4.46 Hz. Furthermore, after listening to each speech stimulus, participants scored their valence and arousal, along with QoE.

The raw EEG and fNIRS signals pre-processed as described in [5], to obtain clean EEG signals and $\Delta[HbO]$ and $\Delta[HbR]$ concentrations. Furthermore, raw fNIRS signals were bandpass filtered between 0.05-2 Hz to create a heart rate and heart rate variability (HRV) time series. Next, the pre-processed EEG, fNIRS and HRV time series signals were then used to extract features that encode users' valence and arousal thus, forming physiological feature set. As such, EEG signals were used to extract graph theoretical features, such as local efficiency (E_l) and global efficiency (E_g), as described in [6]; and asymmetry index and medial beta power (MBP), as described in [4]. Also, using fNIRS signals, certain features, such as average and peak values for $\Delta[HbO]$ and $\Delta[HbR]$ concentrations, as described in [7] were extracted. Moreover, leveraging HRV time series statistical features, such as mean, median, skewness and kurtosis, were extracted. Finally, two quantitative measures were extracted from speech stimuli themselves to model the technological influence factors, as described in [4]. The measures included the slope of the second order derivative of the fundamental frequency ($sF0''$) and the absolute mean of the second order mel frequency cepstrum coefficient ($MFCC_2$). While the $sF0''$ feature models

TABLE I: The goodness-of-fit (r^2) values are reported for each equation developed using different modalities. In the table S, Sub, E, F and H represent Speech, Subjective, EEG, fNIRS and heart rate modalities, respectively.

No.	Modalities	QoE Equations	r^2
1	S	$0.36 - 0.56 * MFCC_2 + 0.44 * sF0''$	0.76
2	S, Sub	$0.004 + 0.02 * MFCC_2 + 0.05 * sF0'' + 1.53 * Val - 0.52 * Ar$	0.96
3	S, E	$0.30 - 0.87 * MFCC_2 + 0.67 * sF0'' + 0.51 * E_l + 0.80 * MBP$	0.87
4	S, F	$0.87 - 0.56 * MFCC_2 + 0.34 * sF0'' - 0.33 * HbR - 0.44 * HbO$	0.84
5	S, H	$0.46 - 0.49 * MFCC_2 + 0.39 * sF0'' - 0.15 * HRV$	0.79
6	S, E, H	$-0.18 - 0.81 * MFCC_2 + 0.62 * sF0'' + 0.48 * E_l + 0.74 * MBP - 0.10 * HRV$	0.88
7	S, F, H	$1.04 - 0.47 * MFCC_2 + 0.27 * sF0'' - 0.38 * HbR - 0.49 * HbO - 0.19 * HRV$	0.88
8	S, E, F	$-0.62 - 0.88 * MFCC_2 + 0.81 * sF0'' + 0.57 * E_l + 1.17 * MBP - 0.07 * HbR + 0.16 * HbO$	0.89
9	S, E, F, H	$0.05 - 0.74 * MFCC_2 + 0.59 * sF0'' + 0.40 * E_l + 0.70 * MBP - 0.19 * HbR - 0.10 * HbO - 0.12 * HRV$	0.90

the macro-prosodic or intonation-related properties of speech, $MFCC_2$ models articulation related properties [8].

In order to assess QoE model performance, first, we explored the goodness-of-fit (r^2) achieved by using only the technology-centric speech metric as a correlate of the QoE score reported by the listeners. Second, we investigated the gains obtained by including HIFs into the QoE models. Here, we measured the r^2 obtained from a linear combination of the technology-centric speech metric combined with the subjective valence and arousal ('ground truth') ratings reported by the listeners. Gains in the goodness-of-fit metric should indicate the benefits of including HIFs into QoE perception models. Lastly, we replaced the ground truth HIFs by the physiological features that showed maximum correlation with affective dimensions. It is expected that the r^2 achieved will lie between those achieved without and with HIFs, thus signalling the importance of fusion of physiological modalities in QoE perception modelling.

III. Results

The physiological features that correlated well with the subjective dimensions of affect were used to develop regression equations. It was observed that, for EEG, E_l computed from high beta band (24-30 Hz) showed maximum correlation with valence whereas, MBP showed maximum correlation with arousal. For fNIRS, average $\Delta[HbR]$ computed from right temporal region showed maximum correlation with valence and average $\Delta[HbO]$ at temporo-parietal region showed maximum correlation with arousal. Moreover, for hear rate signal, average HRV correlated with valence however, there was significant correlation between hear rate derived features and arousal. Therefore, the features that showed maximum correlation with affective dimensions were used to develop the regression equations. Table I reports the regression equations using combinations of features from various modalities along with the goodness-of-fit (r^2) values.

IV. Discussion and Conclusion

Recently, HIFs and objective HIF characterization have gained burgeoning attention from QoE researchers. Towards combining HIFs, such as affective states, with technology-centric speech quality metrics, researchers have investigated the use of EEG in [4]. Here, we investigated the use of fusion of physiological modalities to model affective states. As evident from Table I, amongst individual physiological modalities, EEG based model performed best. However, combination of the three physiological modalities (EEG, fNIRS and heart rate), to model HIFs, resulted in best performing objective model, in comparison to using individual modalities to model HIFs. Therefore, these findings indicate that fusion of multiple physiological modalities would allow for better continuous real-time monitoring of listener affective states. Moreover, using fusion feature set lead to an overall gain of 18.4% in QoE measurement whereas, using EEG feature set lead to an overall gain of 14.5% in QoE measurement, relative to using only technological factors.

References

[1] Q. team, "Qualinet white paper on definitions of quality of experience," Qualinet Cost: European Network on Quality of Experience in Multimedia Systems and Services, Tech. Rep., 2012.

[2] S. Moller *et al.*, "Speech Quality Estimation: Models and Trends," *Signal Processing Magazine IEEE*, vol. 28, no. 6, pp. 18–28, 2011.

[3] C. Mühl *et al.*, "Modality-specific Affective Responses and their Implications for Affective BCI," *Proceedings of the 5th International Brain-Computer Interface Conference 2011*, pp. 120–123, September 2011. [Online]. Available: http://doc.utwente.nl/78294/

[4] R. Gupta, K. Laghari, H. Banville, and T. H. Falk, "Using affective brain-computer interfaces to characterize human influential factors for speech quality-of-experience perception modelling," *Human-centric Computing and Information Sciences*, vol. 6, no. 1, pp. 1–19, 2016.

[5] R. Gupta, H. J. Banville, and T. H. Falk, "PhySyQX: A database for physiological evaluation of synthesised speech quality-of-experience," *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 1–5, 2015.

[6] R. Gupta and T. Falk, "Affective state characterization based on electroencephalography graph-theoretic features," *7th International IEEE/EMBS Conference on Neural Engineering*, pp. 577–580, 2015.

[7] R. Gupta *et al.*, "Using fNIRS to Characterize Human Perception of TTS System Quality, Comprehension, and Fluency: Preliminary Findings," *Proceedings of The Fourth Workshop on Perceptual Quality of Systems (PQS)*, pp. 73–78, 2013.

[8] C. Norrenbrock *et al.*, "Quality prediction of synthesized speech based on perceptual quality dimensions," *Speech Communication*, vol. 66, pp. 17–35, 2015.