Université du Québec
Institut national de la recherche scientifique
Centre Énergie Matériaux Télécommunications

Physiology-based Quality-of-Experience Assessment for Next Generation Multimedia Technologies

By

Rishabh Gupta

A thesis submitted in fulfillment of the requirements for the degree of
Doctorate of Sciences, Ph.D.
in Telecommunications

Evaluation Committee

Internal evaluator and committee president:	Prof. Douglas O'Shaughnessy
External evaluator 1:	Prof. Hantao Liu Cardiff University
External evaluator 2:	Prof. Hussein Al Osman University of Ottawa
Research advisor:	Prof. Tiago H. Falk

©Rishabh Gupta, 2016

Acknowledgements

Firstly, I would like to express my sincere gratitude to my advisor Prof. Tiago H. Falk for the continuous support of my Ph.D study and related research, for his patience, motivation, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis. I could not have imagined having a better advisor and mentor for my Ph.D study.

I would also like to thank Dr. Douglas O'Shaughnessy, Dr. Hantao Liu, and Dr. Hussein Al Osman for taking time off their busy schedules to serve on my Thesis Examination Committee and also for their valuable comments.

My sincere thanks also goes to Mr. Mojtaba Khomami and Mr. Jean-Phillipe Poulin, who provided me an opportunity to join their team as intern. Without their precious support it would not be possible to conduct this research.

I thank Dr. Khalil Laghari, Mr. Hubert J. Banville, Mr. Andrea Clerico and Mr. Raymundo Cassani for the stimulating discussions, for the sleepless nights we were working together before deadlines, and for all the fun we have had in the last four years.

I would not have been here today if it were not for the love and care of my family. I would like to greatly acknowledge my parents who taught me to be a sincere learner and a good human being. I am also grateful to my brother who has had to deal with my naive teaching skills for the past few years. To my dearest wife, without your endless support, encouragement and companionship during those sleepless nights, this thesis would not have been possible. You really helped keep me sane during these last four years and to you I dedicate this thesis.

Abstract

As new multimedia technologies emerge, telecommunication service providers 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 the former two and human influence factors (HIFs) have been mostly neglected due to difficulty in measuring them. In this thesis, we show that measuring HIFs is important for QoE measurement and propose the use of hybrid brain-computer interfaces (hBCIs) for objective measurement of perceived QoE for multimedia technologies, such as affective music videos and text-to-speech systems.

For the development of hBCIs, we explore the use of two neuroimaging techniques, namely electroencephalography (EEG) and functional near-infrared spectroscopy (fNIRS), to better understand neuronal and cerebral haemodynamic changes resultant from multimedia signals of varying quality. Neural correlates of several QoE dimensions were derived and validated on the publicly available DEAP and PhySyQX databases. In general, the parameters derived from EEG and fNIRS indicated correlation between neural activation, in various cortical regions, and signal quality. These individual features derived from EEG and fNIRS were then used to develop classifiers to establish their usability as QoE monitoring modalities. The parameters derived from EEG and fNIRS showed to accurately classify different user states and subjective QoE dimensions. Interestingly, features derived from heart rate, extracted from fNIRS signals, also showed to encode information regarding HIFs. Next, fusion of EEG, fNIRS, and fNIRS-derived heart rate parameters showed to accurately represent several QoE dimensions, including those related to listener affective states.

Finally, the subjectively-derived HIFs were incorporated into the QoE model, leading to gains of up to 26.3% relative to utilizing only technological factors. When utilizing HIFs derived from individual modalities, on the other hand, gains of up to 14.5%, 10.6% and 4% were observed for EEG, fNIRS and heart rate, respectively. The hybrid model based on features from all three physiological modalities resulted in gains of up to 18.4%. These findings show the importance of using BCIs and hBCIs in QoE measurement and also highlight that further improvement may be warranted once improved HIFs correlates are found from EEGs and/or other neurophysiological modalities. It is hoped that these findings will help researchers build better instrumental QoE models that incorporate technological, contextual, and human influence factors.

Keywords Quality-of-Experience, Hybrid Brain-Computer Interfaces, Electroencephalography, Functional Near-infrared spectroscopy, Human Factors

Contents

List of Figures

List of Tables

Liste des Figures

Liste des Tableaux

List of Abbreviations

ACR Absolute Category Rating
AI Asymmetry Index
ANIQUE+ Auditory Non Intrusive Quality Estimation Plus
ANOVA Analysis of Variance
BCI Brain-Computer Interface
CIF Contextual Influence Factor
CCR Comparison Category Rating
CFA Confirmatory Factor Analysis
CFI Comparative Fit Index
CNS Central Nervous System
DCR Degradation Category Rating
DEAP Database for Emotion Analysis using EEG, Physiological and Video signals
ECG Electrocardiography
EEG Electroencephalography
EFA Exploratory Factor Analysis
ERD Event Related Desynchronization
ERP Event Related Potential
ERS Event Related Synchronization
fMRI Functional Magnetic Resonance Imaging
fNIRS Functional Near-Infrared Spectroscopy
GSR Galvanic Skin Response
GOF Goodness of Fit
GFI Goodness of Fit Index
hBCI Hybrid Brain-Computer Interface
HAHV High Arousal High Valence
HALV High Arousal Low Valence
HIF Human Influence Factor
HMM Hidden Markov Model
HQ High Quality
HR Heart Rate

HRV Heart Rate Variability
HSD Honestly Significant Difference
IBI Inter Beat Interval
ICA Independent Component Analysis
ICC Intra Class Correlation
IF Influence Factor
IFI Incremental Fit Index
ITU International Telecommunications Union
ITV Inter Trial Variance
KMO Kaiser Meyer Olkin
LAHV Low Arousal High Valence
LALV Low Arousal Low Valence
LQ Low Quality
MBLL Modified Beer-Lambert Law
MBP Medial Beta Power
MDS Multi Dimensional Scaling
MEG Magnetoencephalography
MFCC Mel Frequency Cepstrum Coefficients
MOS Mean Opinion Score
MSC Magnitude Squared Coherence
NFI Normal Fit Index
NNFI Non Normal Fit Index
NN Normal to Normal
OFC Orbito Frontal Cortex
PANS Peripheral Autonomic Nervous Systems
PANS Peripheral Autonomous Nervous System
PESQ Perceptual Evaluation of Speech Quality
PFC Pre-Frontal Cortex
PhySyQX Physiological Evaluation of Synthetic Speech Quality of Experience
POLQA Perceptual Objective Listening Quality Analysis
QoE Quality of Experience
QoS Quality of Service
RBF Radial Basis Function
RMSE Root Mean Squared Error
RNI Relative Non-Central Index
RVM Relevance Vector Machine
SAM Self-Assessment Manikins
SDNN Standard Deviation to Normal

SEM Standard Error of Mean

SNR Signal to Noise Ratio

SRMR Standardised Root Mean-Squared Residual

SVM Support Vector Machine

TIF Technological Influence Factor

TTS Text to Speech

UX User Experience

VoIP Voice over Internet Protocol

Sommaire récapitulatif

0.1 Introduction

Avec une industrie des communications multimédias en plein essor, de nouveaux services ne cessent d'apparaître. Pour réussir, ces services doivent subir une évaluation continue des performances pour assurer une bonne *qualité* du contenu livré. Auparavant, le terme *qualité* était utilisé par les ingénieurs pour décrire le soi-disant paramètre de *qualité du service* (QS), qui est défini par l'Union internationale des télécommunications (UIT) comme l'*'ensemble des caractéristiques d'un service de télécommunications qui portent sur sa capacité à satisfaire les besoins explicites et implicites de l'utilisateur du service'* [3]. Ainsi, on a supposé que, pour un service particulier, une QS plus élevée conduirait à sa plus grande acceptabilité et, par conséquent, à une augmentation du nombre d'utilisateurs. Cependant, le succès de certains services malgré la mauvaise *qualité*, tels que les premiers systèmes de SMS, a suscité le besoin de comprendre la *qualité* du point de vue des utilisateurs. Cela a conduit à la formulation du terme *qualité de l'expérience* (QE), qui prend en compte le processus de perception et de jugement des utilisateurs. La qualité de l'expérience a été formellement définie comme 'le degré de plaisir ou d'agacement de l'utilisateur/utilisatrice d'une application, provenant de la réalisation de ses attentes selon la personnalité de l'utilisateur/utilisatrice et de son état d'esprit réel' [4].

La QE perçue est le résultat d'un processus de formation de la qualité qui implique une phase de sensibilisation à la qualité qui se traduit par l'identification des caractéristiques de qualité émotionnelles, sensorielles et conceptuelles par la réflexion et l'attribution; puis qui compare les caractéristiques de qualités attendues et expérimentées pour former la QE, telle que présentée à la Fig. 1.1. La QE vise spécifiquement les perceptions et les expériences des utilisateurs qui sont considérées comme plus appropriées pour la conception des services avec une acceptabilité plus élevée. Par conséquent, une meilleure QE assure l'avantage concurrentiel et, finalement, le succès de ce service. Traditionnellement, les aspects de la QE peuvent être évalués soit sur la base de la perception (méthodes subjectives) ou des méthodes instrumentales (méthodes objectives). Les méthodes d'évaluation subjectives exigent des évaluateurs humains de recueillir les informations relatives à la QE pour les différents stimuli multimédias. En règle générale, pour l'évaluation subjective, les utilisateurs interagissent avec les stimuli multimédias et ensuite donnent une note quantitative de la QE momentanée et retenue sur un ensemble d'échelles. Ces méthodes subjectives fournissent également les données de réalité de terrain pour le développement de méthodes objectives d'évaluation

de la QE. Les méthodes objectives, d'autre part, évaluent la QE en utilisant un algorithme ou un instrument. Ces algorithmes sont généralement alimentés à partir d'un ensemble de traits caractéristiques dérivés du système technique ou du signal multimédia à l'étude, qui sont utilisés pour développer des modèles afin d'estimer les données de réalité de terrain fournies par les méthodes subjectives. Les modèles objectifs sont peu coûteux et représentent une alternative plus rapide aux méthodes d'évaluation subjectives de QE. Aussi, récemment, l'intérêt principal des chercheurs s'est concentré sur le développement de nouvelles techniques de caractérisation objective pour la QE en vue d'automatiser le processus d'évaluation de la qualité.

Toutefois, pour le développement des modèles d'évaluation objective de la QE, il est important d'identifier les attributs qui influencent la perception de la QE. Ces attributs sont appelés 'facteurs d'influence' (FI). Comme indiqué dans [4], les FI peuvent être généralement classés en FI technologiques, contextuels et humains. Les FI technologiques (FIT) se rapportent aux paramètres du système et du réseau qui peuvent être facilement mesurés (par exemple, le retard et le débit binaire). Les FI contextuels (FIC) englobent les propriétés situationnelles qui décrivent l'environnement des utilisateurs à l'aide de ses caractéristiques physiques, temporelles ou techniques. Enfin, les FI humains (FIH) prennent en compte toutes les caractéristiques variables et invariables des utilisateurs humains, telles que leurs états émotionnels, préférences, attitudes et objectifs, ainsi que de nombreux autres facteurs subjectifs (par exemple, la fatigue). Au cours des 10 dernières années, les chercheurs se sont concentrés sur les FIT et les FIC pour mettre au point de nouveaux modèles objectifs. Cependant, les prédicteurs de ces techniques ne représentent pas la véritable perception de la QE car ils ne tiennent pas compte des FIH. Fondamentalement, l'incorporation de FIH dans les modèles de QE objectifs peut permettre le développement de modèles plus centrés sur l'utilisateur. Par ailleurs, la Fig. 1.1 montre que les facteurs humains, représentés par les états internes des utilisateurs, modulent manifestement les scores de la QE des utilisateurs. Par conséquent, la recherche interdisciplinaire est encore nécessaire pour combler cette lacune par l'incorporation des constructs FIH dans de nouvelles techniques de caractérisation objective.

À cette fin, il est important de noter que les aspects fondamentaux de la QE, spécifiquement les FIH, sont de nature subjective et ne sont pas directement observables, comme il ressort de la discussion précédente. En effet, la majeure partie du processus de jugement et de formation de la qualité a lieu à l'intérieur du cerveau de l'utilisateur. Ainsi, un sondage des activités neuronales (impulsions électriques) ou hémodynamiques (débit sanguin) du cerveau est censé fournir des indications importantes concernant le processus de jugement de la qualité et les FIH. En fin de compte, ces découvertes pourraient servir à développer de meilleures techniques de caractérisation objective de la QE. Les modalités neurophysiologiques, comme l'électroencéphalographie (EEG) et la spectroscopie proche infrarouge fonctionnelle (SPIRf), peuvent sonder les activités hémodynamiques neuronales et cérébrales. En particulier, les modalités neurophysiologiques peuvent agir en tant qu'interface entre les états internes des utilisateurs et les modèles d'évaluation de QE objectifs. Ce système constitue ce que l'on appelle interface cerveau-ordinateur (ICO), qui est défini comme *'système qui mesure l'activité du système nerveux central (CNS) et la convertit en rendement artificiel*

qui remplace, restaure, améliore, supplémente, ou améliore le rendement naturel du CNS et change ainsi les interactions continues entre le système nerveux central et son environnement externe ou interne' [5]. Plus récemment, une nouvelle approche appelée ICO hybride (ICOh) a émergé, qui vise à fusionner les différentes modalités physiologiques, telles que l'électrocardiographie (ECG), les capteurs de réponse électrodermale (RÉ) et le traqueur oculaire, avec au moins une modalité neurophysiologique [6]. Ces ICO peuvent être utilisées pour estimer les FIH et incorporées dans les modèles d'évaluation de QE objectifs de pointe basés sur une ICO. Les sous-sections suivantes décrivent l'état de l'art et les méthodes d'évaluation de la QE basées sur les ICO. Aussi, les principales contributions de cette thèse sont répertoriées et l'organisation de ce document décrite.

0.1.1 Méthodes d'évaluation de la QE de pointe

La QE peut être évaluée subjectivement ou objectivement [7, 8]. Les méthodes d'évaluation subjective quantitative impliquent généralement l'élaboration de questionnaires avec des échelles de notation, des enquêtes et des études d'utilisateurs qui peuvent être effectuées soit dans un laboratoire ou au moyen de paramètres du « monde réel ». Le plus souvent, les tests subjectifs réalisés sont des tests de notations de catégorie absolue (ACR). L'UIT a élaboré des directives d'étude subjective pour les évaluations de la qualité perceptuelle. Plus précisément, pour différentes applications, l'UIT a émis différentes recommandations, telles que l'ITU-T Rec. P.800 pour la qualité de la parole, P.910 pour la qualité vidéo, P.911 pour la qualité audiovisuelle et P.85 pour les systèmes de synthèse vocale ('Text-to-Speech', TTS) [9]. Par ailleurs, pour évaluer la réaction affective des utilisateurs à la qualité des signaux multimédia, les chercheurs utilisent deux dimensions affectives, comprenant la valence, qui mesure le plaisir devant un événement et l'activation physiologique, qui mesure l'intensité devant l'événement, sur des graphiques bidimensionnels [10]. Pour caractériser quantitativement ces deux primitives émotionnelles, le système pictural d'auto-évaluation 'Self Assessment Manikin' (SAM) est couramment utilisé, comme illustré à la Fig. 1.3 [11, 10].

Les méthodes objectives, d'autre part, utilisent des mesures centrées sur la technologie pour estimer la QE. Les modèles centrés sur la technologie remplacent l'évaluateur humain par un algorithme informatique qui a été développé pour extraire les caractéristiques pertinentes du signal analysé (parole, audio, image ou vidéo) et cartographier un sous-ensemble / une combinaison de ces caractéristiques dans une valeur QE *estimée*. L'UIT, par exemple, a normalisé plusieurs modèles objectifs au cours de la dernière décennie, comme PESQ (Recommandation P.862 [12]) et POLQA (Recommandation P.863 [13]) et la Recommandation ITU P.563 [14].

Comme mentionné ci-dessus, des méthodes objectives existantes ont été « centrées sur la technologie », en se basant principalement sur les facteurs technologiques [7]. Afin de développer des méthodes d'évaluation de QE, cependant, les facteurs d'influence contextuels et humains doivent également être incorporés. Il y a eu quelques tentatives pour intégrer les facteurs contextuels dans les modèles de QE objectifs, tels que le modèle décrit dans [15], qui utilise certains traits caractéristiques du signal pour estimer les effets de la réverbération de la pièce sur le signal de parole. Toutefois, l'incorporation des facteurs humains n'a pas été l'objet de tentatives similaires principa-

lement parce que les FIH, formulées à l'intérieur du cerveau des utilisateurs, ne sont pas directement observables. À cette fin, sonder l'activité du cerveau en utilisant des outils basés sur une ICO (ou de neuro-imagerie) devrait fournir une estimation objective intéressante pour les FIH. La section suivante décrit la méthodologie et les outils nécessaires à l'évaluation de la QE basée sur une ICO.

0.1.2 Évaluation de la QE basée sur une ICO

La modélisation objective de la perception humaine de la QE basée sur une ICO a fait beaucoup de progrès dans les dernières années. Ces techniques pourraient fournir une alternative viable aux modèles prédictifs de QE objectifs existants ou aider les modèles existants à fournir une meilleure prédiction de la qualité. Cette idée provient du fait que ces techniques mesurent directement l'activité neurophysiologique; comme la plus grande partie de processus de jugement de la qualité a lieu à l'intérieur du cerveau de l'utilisateur, celles-ci pourraient fournir une meilleure approximation de la QE perçue par l'humain. Ainsi, le système ICO proposé pour l'évaluation de la QE relève de la prémisse des soi-disant ICO passives. Les ICO passives se basent sur l'analyse de l'activité cérébrale arbitraire qui survient sans l'objectif d'un contrôle volontaire, dans le but d'enrichir les interactions homme-machine en utilisant les informations implicites concernant l'état réel des utilisateurs [16].

Une ICO caractéristique se compose de plusieurs modules qui forment une boucle entre l'utilisateur et l'ordinateur ou la machine [6], tel que représenté par la Fig. 1.4. Par conséquent, une ICO comprend une étape de collecte de données, suivie par des étapes d'extraction de traits caractéristiques et de nettoyage des signaux. Enfin, les traits caractéristiques extraits sont traduits en décisions. Plusieurs classifieurs, comme les arbres décisionnels [17, 18], les machines à vecteurs de support (SVM) [19, 20, 21] ou les machines à vecteurs de pertinence (RVM) [22, 23], peuvent être utilisés pour traduire les traits caractéristiques en décisions. Enfin, la décision concernant l'état mental actuel des utilisateurs est envoyée à un dispositif de commande, tel qu'un robot ou un téléphone intelligent, qui fournit une réponse aux utilisateurs.

Électroencéphalographie (EEG)

En général, la collecte de données d'EEG consiste à placer plusieurs électrodes sur la tête avec un gel conducteur suivant le système international 'Standard 10-20' [24], tel qu'illustré par la Fig. 1.5. Suivant la collecte de l'EEG, les données sont prétraitées pour supprimer les artefacts (bruit) et atteindre un rapport de signal-bruit (SNR) élevé, à l'aide de techniques de traitement de signaux avancées. Ensuite, un autre traitement des données est effectué pour extraire certains traits caractéristiques utiles pour caractériser un événement particulier. Le trait caractéristique le plus couramment extrait des données d'EEG est ce que l'on appelle le « potentiel lié à l'événement » (ERP). Un ERP est une variation de l'amplitude du signal temporel d'EEG en réponse à un événement sensoriel, cognitif ou moteur [25]. Toutefois, l'application de ces techniques est surtout limitée à des signaux multimédias courts qui incorporent des bruits perceptibles à leur tout début [26, 27]. Par conséquent, il y a une nécessité d'évaluer les traits caractéristiques d'EEG qui peuvent être plus pratiques pour l'évaluation des états des utilisateurs à long terme.

À cette fin, dans cette thèse, nous avons proposé deux séries de traits caractéristiques d'EEG, qui sont soit basés sur le spectre de puissance ou basés sur le spectre croisé. Les traits caractéristiques basées sur le spectre de puissance mesurent les modifications relatives dans les bandes spectrales des signaux d'EEG [28], en réponse à un événement ou un stimulus et sont appelées traits caractéristiques de desynchronisation liées à un événement (DRÉ). Les signaux d'EEG se composent de cinq sous-bandes majeurs d'intérêt, à savoir : delta (0 à 4 Hz), thêta (4 à 8 Hz), alpha (8 à 12 Hz), bêta (12 à 30 Hz) et gamma (30 Hz). Les traits caractéristiques basés sur le spectre de puissance se sont avérés utiles pour caractériser les GIH et la QE [28, 29, 30, 31]. Cependant, pendant la découverte des contenus multimédias différentes régions du cerveau sont censées communiquer entre elles, comme le montre la Figure 1.6 [32]. Par conséquent, pour collecter des informations concernant la dynamique du flux d'informations, les traits caractéristiques dérivés du spectre croisé et inspirés de la théorie des graphes [33] peuvent être utiles.

Spectroscopie proche infrarouge fonctionnelle (SPIRf)

À l'aide de SPIRf, les variations des niveaux d'oxyhémoglobine ($\Delta[HbO]$) et de désoxy-hémoglobine ($\Delta[HbR]$) peuvent être détectées. Ceci est possible en plaçant une source et un détecteur à 3 cm environ l'une de l'autre sur la surface du crâne. La surface du cerveau est ensuite éclairée par un rayonnement infrarouge de faible énergie (avec généralement deux différentes longueurs d'onde, par exemple 760 nm et 850 nm) qui se déplace à travers la peau et le crâne jusqu'au cortex. Ce rayonnement est par la suite réfléchi par la surface corticale du cerveau, puis capturé à l'aide de détecteurs à la surface de la peau. Le chemin emprunté par le rayonnement d'une source jusqu'au détecteur forme un *canal* de SPIRf. Les variations de $[HbO]$ et $[HbR]$ se manifestent habituellement dans l'intensité du rayonnement réfléchi. Les mesures d'intensité sont convertis en $\Delta[HbO]$ et $\Delta[HbR]$ en utilisant ce que l'on appelle la loi de Beer-Lambert modifiée (LBLM) [34], qui est utilisée pour interpréter l'activation corticale.

Les signaux de SPIRf bruts (intensités enregistrées par les détecteurs) sont souvent corrompus par divers artefacts physiologiques, tels que les pulsations cardiaques ou le cœur (fréquence cardiaque), la respiration et les ondes de Mayer (ou ondes de tension artérielle) [35], qui peuvent être facilement éliminées en utilisant les filtres passe-bande. Ensuite, afin de rendre possible la description de l'état des utilisateurs à long terme à l'aide de la SPIRf, divers traits caractéristiques liés à la dynamique temporelle de l'amplitude $\Delta[HbO]/\Delta[HbR]$, tel que représenté par la Fig. 1.7, sont extraits et utilisés pour caractériser les diverses FIH, tels que la sympathie [36] et les états émotionnels [37]. Cependant, la SPIRf comme technique en est encore à ses débuts dans le domaine de l'évaluation de la QE et pourrait s'avérer très efficace aux côtés de l'EEG, dans l'élaboration de meilleurs modèles objectifs. À cette fin, dans cette thèse de doctorat, nous avons exploré un ensemble de traits caractéristiques basés sur la SPIRf, comme la moyenne, la variance, le coefficient d'aplatissement et le coefficient de dissymétrie de $\Delta[HbO]$ et $\Delta[HbR]$ comme corrélations de la QE.

Système nerveux autonome périphérique (SNAP)

En plus d'enregistrer l'activité du cerveau pour la caractérisation de l'état des utilisateurs, il est également utile d'enregistrer certains signaux physiologiques périphériques en utilisant des techniques telles que l'ECG, la RÉ, la température cutanée et la mesure de l'activité respiratoire. Ces mesures sont les manifestations du SNAP qui aide le système nerveux central à communiquer avec le reste du corps. Récemment, les signaux basés sur le SNAP ont prouvé qu'ils pouvaient caractériser les FIH avec précision [28, 38, 39, 40, 41, 42]. Ainsi, on pourrait émettre l'hypothèse qu'un cadre conceptuel intégrant toutes les techniques de mesure de la réponse neurophysiologique pourrait permettre de développer des modèles d'évaluation de la QE objectifs plus précis. Des études antérieures, cependant, se sont appuyées sur du matériel et des périphériques dédiés pour collecter des données physiologiques, tels que la pléthysmographie pour la surveillance de la fréquence cardiaque et des ceintures de respiration à base de jauge de déformations. Ici, nous empruntons un autre chemin et nous proposons d'extraire des informations de fréquence cardiaque en traitant le signal SPIRf brut. Cela permettra d'obtenir un ensemble plus riche de données multimodales pour caractériser les FIH qui affectent la QE perçue.

Contributions de la thèse

L'objectif de cette thèse est de développer des mesures pertinentes aux ICO, basées sur la fusion des différentes modalités neurophysiologiques, pour caractériser les facteurs humains qui influencent la perception de la QE, de même que d'incorporer ces ICO hybrides dans les modèles de la QE objectifs de pointe. Pour l'évaluation des mesures réalisées, un scénario basé sur les systèmes (TTS) de synthèse texte-parole a été étudié, comme au cours des dernières années, les systèmes TTS sont devenus énormément populaires, en particulier dans le domaine des assistants numériques personnels (par exemple, Siri d'Apple, Google Now, et Cortana de Microsoft), les centres d'appels automatisés, les assistants de lecture aux aveugles et les systèmes de positionnement global. Ici, nous présentons une liste des principales contributions de cette thèse:

1. Une base de données neurophysiologiques multimodales en source libre, la base de données appelée PhySyQX, a été développée pour la caractérisation des FIH. Les données subjectives et neurophysiologiques de la base de données ont été utilisées pour élaborer différents modèles basés sur la ICO pour caractériser la QE. Les publications qui ont résulté de cette contribution comprennent [43, 44].
2. Deux classes de traits caractéristiques basés sur l'analyse du spectre de puissance et du spectre croisé des signaux d'EEG ont été proposées pour la caractérisation à long terme des FIH. Ces approches ont permis de comprendre l'interaction des informations entre les régions du cerveau responsables du traitement des informations relatives à la QE. Pour la validation, les traits caractéristiques proposées ont été testées sur deux bases de données distinctes, à savoir, PhySyQX et la base de données DEAP. Les publications qui ont résulté de cette contribution comprennent [45, 32, 46].

3. Les traits caractéristiques basés sur la SPIRf, sur la base de la dynamique temporelle de $\Delta[HbO]$ et $\Delta[HbR]$ ont été proposées pour la caractérisation à long terme des FIH. Par ailleurs, la pulsation cardiaque dans les données SPIRf brutes a été extraite et utilisée pour développer des modèles pour le monitorage des FIH. En outre, les traits caractéristiques d'EEG, de la SPIRf et de fréquence cardiaque basée sur la SPIRf ont été utilisés pour développer une ICO hybride dans le même but. Malheureusement, la base de données DEAP ne contient pas de données SPIRf et par conséquent, n'a pas été utilisée pour tester les traits caractéristiques basés sur la SPIRf. Les publications qui ont résulté de cette contribution comprennent [47, 46].
4. Les premiers pas vers l'intégration des outils basés sur une ICO hybride dans les modèles de QE objectifs de pointe ont été posés. Les mesures basées sur une ICO hybride qui caractérisent les états affectifs humains ont été incorporées dans un modèle de QE objectif pour la synthèse texte-parole. Les publications qui ont résulté de cette contribution comprennent [48, 49].

0.1.3 Organisation de la thèse

La structure de cette thèse suit le processus d'intégration des ICO hybrides, qui caractérisent les FIH, dans des modèles d'évaluation de QE objectifs de pointe. Par conséquent, dans le chapitre 1, nous avons présenté le thème de l'évaluation de QE basée sur les ICO. Le chapitre 2 décrit la méthodologie de collecte de données neurophysiologiques multimodales et subjectives simultanées pour le développement des ICO hybrides qui caractérisent les FIH. Les tests d'évaluation de QE subjective ont réuni les informations dites 'vérité de terrain' concernant les FIH, sur la base desquelles les ICO hybrides ont été développées. Par ailleurs, ce chapitre fournit une analyse en profondeur des données subjectives qui explorent les divers facteurs humains comportementaux et affectifs qui peuvent potentiellement influencer la QE perçue. Les Chapitres 3 et 4 abordent le développement des caractéristiques basées sur l'EEG et la SPIRf, respectivement, qui sont utiles pour le suivi à long terme des FIH en utilisant les ICO. Ensuite, les résultats de chacune des ICO développées en utilisant l'EEG et la SPIRf, ont été fusionnés pour développer une ICO hybride au chapitre 5. Dans le chapitre 6, les mesures basées sur les ICO développées dans les chapitres précédents ont été intégrées dans le modèle d'évaluation de QE objectif de pointe. Enfin, le chapitre 7 présente une discussion générale et les conclusions.

0.2 Développement de la base des données physiologiques

L'incorporation des FIH dans les modèles de QE objectifs en utilisant les ICO exige des procédures de collecte de données rigoureuses. En outre, les données neurophysiologiques sont sujettes à diverses sources de bruits physiologiques et instrumentaux, ce qui nécessite la mise en œuvre de techniques d'analyse et de nettoyage de données complexes. Par conséquent, du fait de leur nature chronophage et coûteuse, et qu'ils nécessitent la supervision d'experts avisés, l'adoption de méthodes d'évaluation de QE basées sur les ICO, a été lente. La disponibilité de bases de données en libre accès, sur la base des données collectées par le biais des méthodologies d'évaluation de QE basées sur

les ICO, peut aider à atténuer les craintes de la communauté des chercheurs liées à l'utilisation de ces nouvelles techniques. Ces bases de données permettent un accès facile aux données d'évaluation de QE basée sur les ICO à une plus grande partie de la communauté de chercheurs, ce qui n'est pas le cas avec la collecte de données neurophysiologiques. De plus, la disponibilité en libre d'accès permet la réplication de l'étude, à partir de laquelle les données ont été collectées, fournissant ainsi plus de validité aux résultats.

À cette fin, la base de données PhySyQX présentée ici, a été élaborée pour explorer les effets des FIH sur les processus de formation de la qualité de la parole synthétisée en utilisant l'enregistrement simultané de l'électroencéphalographie (EEG) et de la spectroscopie proche infrarouge fonctionnelle (SPIRf). On espère que la base de données complétera le corpus traditionnellement en libre accès de synthèse texte-parole (TTS), tels que ceux de Blizzard Challenges [50] (qui ne fournissent que des fichiers audio avec des notes subjectives pour les facteurs comportementaux humains). Ainsi, cela permet aux chercheurs de corréler leurs perceptions neuronales avec des notes subjectives obtenues et de développer des méthodes d'évaluations de la QE objectives basées sur des ICO hybrides. Une autre différence essentielle entre la base de données PhySyQX et d'autres bases de données de paroles TTS traditionnelles en accès ouvert est qu'elle comprend une plus grande variété de dimensions subjectives, qui comprennent à la fois les facteurs humains comportementaux et affectifs. Par conséquent, cela nous permet d'établir les effets des états affectifs des utilisateurs sur la perception de la QE.

0.2.1 Documents et méthodes

Les données ont été collectées sur deux sessions. Les données de la première session (pilote) ont été utilisées pour la phase exploratoire de l'analyse des facteurs et les données de la deuxième (principale) session, qui ont formé la base de données PhySyQX, ont été utilisées pour la phase de confirmation de l'analyse factorielle. Pour collecter les données, vingt et un participants en bonne santé (8 femmes, âge moyen = 23,8±4,35 ans) ont été recrutés et ont consenti à participer à l'étude. Les données d'EEG de soixante-deux canaux (AF7 et AF8 ont été retirés) ont été enregistrées en utilisant un système Biosemi ActiveTwo (Amsterdam, Pays-Bas). Les données d'EEG ont été enregistrées à une fréquence d'échantillonnage de 512 Hz et des électrodes ont été placées sur le cuir chevelu selon le système international 10-20 (Fig.2.2)[24]. En même temps, les données de la SPIRf ont été enregistrées en utilisant le système NIRx NIRScout avec 16 sources bi-longueur d'onde (longueurs d'onde sondées de 760 nm et 850 nm) et 24 détecteurs et une fréquence d'échantillonnage de 4,46 Hz. Les optodes SPIRf (sources et détecteurs) ont été placés à côté des électrodes d'EEG, comme présenté à la Fig. 2.2. Chaque paire source-détecteur avec une distance d'environ 3 cm formant un canal dit «fonctionnel», un total de 60 canaux fonctionnels était disponible. Le Tableau 2.1 énumère les stimuli vocaux utilisés pour cette étude avec certains aspects importants. Les stimuli se composaient de quatre voix naturelles et sept voix de synthèse obteneus à partir de systèmes commerciaux à savoir, Microsoft, Apple, Mary TTS Unit Selection & HMM, vozMe, Google et Samsung. Les sujets ont évalué les stimuli vocaux sur 12 échelles subjectives différentes,

telles que la compréhension, la douceur de la voix et l'acceptation, comme présenté au Tableau 2.2. La plupart des dimensions subjectives étaient conformes aux recommandations P.85. Cependant, les dimensions supplémentaires de valence, d'activation physiologique et de dominance ont également été introduites, qui ont été notées en utilisant des mannequins d'auto-évaluation (SAM) [51].

Après la collecte de données, une analyse préliminaire des données subjectives a été effectuée pour étudier la qualité des données en utilisant la corrélation intra-classe et l'analyse de sous-échantillons [52]. De plus, les différences entre les divers systèmes TTS ont été étudiées en utilisant une analyse de la variance (ANOVA) et une analyse de corrélation. En outre, pour valider l'introduction des dimensions affectives (par exemple, la valence), ainsi que les dimensions subjectives des recommandations P.85 [9], pour mesurer la mesure subjective de QE, nous avons effectué une analyse factorielle exploratoire et confirmatoire.

0.2.2 Résultats

La Figure 2.3 montre le diagramme en boîte pour toutes les évaluations subjectives acquises au cours de l'étude. Ces diagrammes en boîte montrent la répartition de chacune des évaluations subjectives sur les différents systèmes TTS. Au Tableau 2.3, sont présentées les valeurs F de l'analyse unidirectionnelle de variance (ANOVA). L'ANOVA a systématiquement montré des différences significatives entre les voix naturelles et les systèmes TTS pour toutes les dimensions subjectives, sauf la dominance et la vitesse d'élocution. De plus, pour quantifier la fiabilité inter-évaluateurs des dimensions subjectives, le Tableau 2.3 présente également les coefficients de corrélation interclasse (CIC). Par ailleurs, le tracé résultant de l'analyse de sous-échantillonnage, pour chaque dimension subjective, est présenté à la Figure 2.4. Toujours au Tableau 2.3, nous avons présenté la matrice du coefficient de corrélation croisée de Pearson pour les 12 dimensions subjectives.

Ensuite, une analyse factorielle exploratoire (AFE) a été réalisée et a abouti à deux facteurs, où les notes *douceur de la voix*, *acceptation*, *effort d'écoute*, *problèmes de compréhension* et *valence* étaient enregistrées de façon plus significative sur le facteur 1. Les notes *intonation*, *émotion*, *naturel* et *activation physiologique*, à leur tour, étaient chargées sur le facteur 2, comme indiqué au Tableau 2.4. Comme les charges pour la dominance et le débit d'élocution ne sont pas significatives, celles-ci ont été ignorées des analyses ultérieures. De plus, l'analyse des facteurs de sous-échantillonnage a validé la fiabilité de la structure factorielle et de la suffisance des données. Pour la vérification de la structure factorielle obtenue, une analyse factorielle confirmatoire (AFC) a été réalisée. Les paramètres d'ajustement de modèles obtenus à partir de la AFC, ainsi que des tests d'invariance de mesure, comme présenté dans le Tableau 2.5, ont validé le modèle.

0.2.3 Discussion

L'analyse des données d'exploration a établi la qualité des données obtenues. D'autre part, partant de l'analyse exploratoire présentée au Tableau 2.4, il est évident que l'AFE a donné lieu à l'extraction de deux facteurs qui encodent le *plaisir d'écoute* et la *prosodie*. Cela établit le 'plaisir d'écoute 'et la 'prosodie' comme les deux dimensions perceptuelles de la QE de la synthèse vocale.

De plus, les résultats corroborent l'AFE réalisée pour les livres audio, comme présenté dans [53]. Fait intéressant, les échelles de valence et d'activation physiologique ont été chargées sur deux facteurs différents, les facteurs 1 et 2, respectivement. Les échelles de valence et d'activation physiologique forment les deux dimensions orthogonales de l'expérience émotionnelle/affective correspondant aux caractères positif/plaisant et d'alerte [54], respectivement. Ainsi, la charge de l'élément de valence sur le facteur 1 établit en outre la relation du facteur 1 à la dimension perceptive du 'plaisir d'écoute'. Aussi, le chargement de l'échelle d'activation physiologique sur le facteur 2 associe le caractère d'alerte du stimulus suscité à la prosodie de la parole, ce qui est également corroboré par les résultats antérieurs présentés dans [55]. Ces résultats indiquent que les modifications dans les constructions perceptuelles sous-jacentes de QE, dues aux modifications dans la qualité du système, modifient les états affectifs des utilisateurs. Par conséquent, il est évident que les échelles affectives correspondant aux dimensions de valence et d'activation physiologique sont importantes pour estimer les dimensions perceptuelles sous-jacentes de l'expérience des utilisateurs avec des assistants numériques personnels.

Les prétraitements et les étapes d'extraction de traits caractéristiques pour les signaux physiologiques sont décrits dans les sections suivantes. Par ailleurs, les sections suivantes décrivent le développement des techniques de caractérisation objective des FIH en utilisant des outils neurophysiologiques.

0.3 Caractérisation des FIH en utilisant des ICO basées sur l'EEG

Dans cette thèse, deux traits caractéristiques différents basés sur l'EEG ont été étudiés pour l'évaluation de la QE de signaux multimédia de longue durée. La première série de traits caractéristiques contient les traits caractéristiques indice d'asymétrie et (dé)synchronisation reliées à un événement (DRÉ/SRÉ), qui mesurent les *modifications* relatives dans le spectre de puissance dans les sous-bandes d'EEG [56], alors que la deuxième série de traits caractéristiques a été dérivée de l'analyse graphique théorique de l'activité cérébrale [33].

0.3.1 Méthodes

Lors de l'évaluation des traits caractéristiques proposés basés sur l'EEG, nous avons étudié deux bases de données, à savoir les bases de données PhySyQX et DEAP [28]. À ce titre, la base de données PhySyQX utilise des signaux sonores produits par les différents systèmes de synthèse texte-parole, alors que la base de données DEAP utilise des stimuli audiovisuels plus complexes issus de ses vidéos musicales affectives. Cependant, la base de données PhySyQX se compose de multiples dimensions comportementales et affectives subjectives, alors que la base de données DEAP n'explore que les facteurs affectifs. Comme la base de données PhySyQX, la base de données DEAP comprend des données d'EEG acquises alors que les participants découvraient les signaux multimédia. Cependant, la base de données DEAP se composait d'enregistrements d'EEG de seulement 32 canaux et ne contient pas d'enregistrements de SPIRf.

Après la collecte de données, les signaux enregistrés ont été nettoyés et utilisés pour l'extraction des traits caractéristiques. Tout d'abord, les traits caractéristiques basés sur la DRÉ, elle-même basée sur le spectre de puissance, ont été extraits car il a été constaté qu'elles encodaient l'activation corticale concernée à diverses activités du cerveau, comme la perception et le jugement [57, 56] et le traitement d'informations sensorielles, cognitives ou motrices [58]. Les traits caractéristiques de DRÉ ont été calculés en utilisant la méthode de la variance inter-essai [59] et la méthode de Welch [60]. Ensuite, afin de coder les interactions asymétriques entre les hémisphères corticaux, les traits caractéristiques d'indice d'asymétrie (IA) ont été calculés pour chaque sous-bande d'EEG [28]. Enfin, les traits caractéristiques basés sur le spectre croisé dérivés en utilisant une analyse graphique théorique ont été calculés. À cette fin, ses matrices d'adjacence ou graphiques ont été créées sur la base de la fonction de cohérence entre les différents canaux d'EEG. Les graphiques résultants ont été utilisés pour calculer les divers paramètres qui codent le traitement des informations intégrées, telles que la longueur du chemin caractéristique et l'efficacité globale, le traitement des informations séparées, telles que le coefficient de clustering et de l'efficacité locale, la propriété de petit réseau [33].

Ensuite, pour la compréhension des bases neurales des FIH et pour quantifier la relation entre les FIH et traits caractéristiques basés sur l'EEG, nous avons étudié les corrélats neuraux des FIH en utilisant la statistique de corrélation de Pearson entre les notes subjectives et les traits caractéristiques d'EEG.

Suite à l'analyse de corrélation, les traits caractéristiques basés sur l'EEG ont ensuite été utilisés pour résoudre le problème de classification binaire pour les dimensions subjectives (par exemple, activation physiologique haute/basse) [28, 29, 45] où les notes subjectives de chaque participant ont été utilisées comme vérité de terrain. Pour classer les dimensions subjectives de la base de données PhySyQX, un arbre de décision raccourci basé sur un indice de Gini a été utilisé. Cependant, pour la base de données DEAP, les classifieurs basés sur des SVM et RVM (notés RVM_1 dans le texte suivant) ont été étudiés. Plus précisément, nous avons utilisé la nature probabiliste du classifieur RVM pour trouver le meilleur seuil de probabilité de telle sorte qu'il optimise les performances du classifieur, noté comme classifieur RVM_2 dans le texte suivant. Les classifieurs ont été validés par la technique de validation croisée leave-one-out [28]. Les performances du classifieur ont été quantifiées en utilisant l'exactitude de classification et la valeur F1 pondérée [61]. Enfin, la classification de fusion basée sur la décision a aussi été mise en œuvre pour les bases de données PhySyQX et DEAP.

0.3.2 Résultats expérimentaux

D'abord, pour la base de données PhySyQX, les corrélats neuraux de percepts de la QE ont été analysés à l'aide des cartes topographiques pour les corrélations moyennes entre les caractéristiques neurales et toutes les dimensions subjectives de la QE. En tant que telles, des cartes de corrélation topographiques ne sont présentées que pour la dimension de QE d'impression générale et pour l'EEG à la Fig. 3.2. Il est évident que l'augmentation de la qualité vocale a induit des

augmentations dans les sous-bandes d'EEG en-dessous de la bande bêta inférieure dans les régions gauche et droite fronto-temporales du cerveau. De plus, $DR\acute{E}_{h-\beta}$ et $DR\acute{E}_{\gamma}$ a toujours montré des corrélations significatives pour les régions temporales et occipitales du cerveau, avec toutes les dimensions subjectives. Une analyse de corrélation similaire avec des caractéristiques IA et graphiques théoriques a révélé des corrélations significatives avec certaines dimensions subjectives. En général, il a été constaté que les traits caractéristiques IA dérivés de la sous-bande α sur la région frontale du cuir chevelu étaient positivement corrélées avec certaines dimensions subjectives, telles que la valence et l'impression générale. Pour les traits caractéristiques graphiques théoriques, d'autre part, la majorité ($> 50\%$) des traits caractéristiques représentés par des mesures de traitement séparées ont conduit à des coefficients de corrélation significatifs au seuil entre 0,2 et 0,6. De même, pour la base de données DEAP, on a observé que l'importance n'a été atteinte que pour les caractéristiques calculées à des seuils compris entre 0,2 et 0,5. Parmi les traits caractéristiques significativement corrélés, les mesures de ségrégation, d'intégration et de petit réseau étaient également réparties. Cependant, les coefficients de corrélation positifs ont été observés pour E_L, E_g, C_{mean} et S et des corrélations négatives ont été obtenues pour L dans la plupart des sous-bandes d'EEG. Par ailleurs, pour visualiser les différences significatives entre les catégories 'haute' et 'basse', ainsi que l'interaction entre les modules d'intégration et de ségrégation, des tests T ont été calculés entre des échantillons de catégorie 'haute' et 'basse' pour E_L et E_g, tels que présentés aux Fig. 3.3 et Fig. 3.4.

Pour la base de données PhySyQX, le Tableau 3.1 présente les exactitudes moyenne et les valeurs F1 sur tous les participants pour chaque trait caractéristique défini pour l'EEG, pour chaque dimension subjective. Le tableau indique les classifieurs les plus performants de chaque sous-bande d'EEG ainsi que la fusion de décision pour chaque ensemble de traits caractéristiques. Les performances pour chaque classifieur ont été testées à l'aide de tests T d'échantillons répétés bilatéraux sur les résultats concaténés de chaque échelle de notation et participant, tel que suggéré par [28]. Au vu du Tableau 3.1, il est évident que l'utilisation d'une combinaison de séries de traits caractéristiques individuelles, il est possible de classifier *toutes* les dimensions subjectives de la QE (sauf la dimension émotionnelle de la voix en utilisant les traits caractéristiques graphiques) avec des performances supérieures au hasard. Enfin, il a été constaté que la fusion des classifieurs basés sur les traits caractéristiques de la théorie des graphes, la DRÉ et l'IA résultaient en une augmentation de $1-2\%$ des performances pour chaque dimension; cependant, cette observation n'a pas été significative.

Pour la base de données DEAP, le Tableau 3.2 présente la valeur F1 et l'exactitude de classification pour le classifieur correspondant à la sous-bande de fréquences avec une valeur F1 maximum choisi parmi les 10 classifieurs individuels pour les catégories d'activation physiologique et de valence. On peut observer que les caractéristiques graphiques, de puissance spectrale et d'IA donnent de bien meilleures performances que le hasard dans la classification des états émotionnels des utilisateurs. De plus, il y avait une amélioration importante ($p < 0.05$) de l'exactitude de classification des dimensions affectives à l'aide des traits caractéristiques graphiques théoriques proposés par rapport aux traits caractéristiques traditionnels, comme le montre le Tableau 3.2. Par ailleurs, le Tableau 3.3

montre les performances réalisées avec une fusion au niveau de la décision de chacun des classifieurs RVM_2 et le vote majoritaire. Au vu du Tableau 3.3, on peut remarquer que tous les classifieurs ont donné de bien meilleurs résultats que le hasard, et que les classifieurs correspondant à la fusion de décision des traits caractéristiques graphiques ont donné de bien meilleurs résultats que les classifieurs correspondants utilisant des traits caractéristiques traditionnels. De plus, en comparant les classifieurs individuels et les classifieurs basés sur la fusion de décisions, les classifieurs de fusion de décisions ont donné de bien meilleurs résultats que les classifieurs individuels.

0.3.3 Discussion

Des recherches antérieures ont montré qu'une augmentation de la $DR\acute{E}_{\alpha}$ ou $DR\acute{E}_{l-\beta}$, ou une diminution de la $DR\acute{E}_{\gamma}$ ou $DR\acute{E}_{h-\beta}$ (suggérant une SRÉ en γ et h-β) indiquaient une activation corticale [56]. Par conséquent, les graphiques de corrélation topographiques pour la base de données PhySyQX présentés par la Fig. 3.2 indiquent l'activation fronto-temporale supérieure gauche et droite avec l'augmentation de la qualité vocale. Les régions temporales gauche et droite du cerveau sont considérées comme importantes pour la perception de la parole [62], la compréhension [63] et la perception de la tonie [64]. Comme la prosodie de la parole synthétique (ou l'intonation) est connue pour affecter la perception du naturel [65] et du module de l'état émotionnel associé à l'extrait de la parole produite [66], l'activation de la région temporale gauche et droite est attendue pour les systèmes TTS de qualité variable. Ceci est corroboré par les graphiques de corrélation topographiques à la Fig. 3.5 entre $DR\acute{E}_{h-\alpha}$ et le CP et les dimensions Int, car ils montrent des corrélations significatives dans la région gauche fronto-temporale et droite fronto-temporale, respectivement. L'analyse de corrélation utilisant les traits caractéristiques graphiques a révélé des corrélations significatives entre les seuils de 0,2 et 0,6. Cela peut être dû à des effets de conduction de volume pour les seuils inférieurs à 0,2 et de graphiques déconnectés pour les seuils inférieurs à 0,6 [33]. Toutefois, un pourcentage plus élevé de traits caractéristiques d'efficacité locale corrélée de façon significative est indicatif de l'augmentation de traitement des informations locales (dans le contexte des réseaux cérébraux) avec l'augmentation de la qualité des stimuli de synthèse texte-parole.

Pour la base de données DEAP, l'analyse de corrélation entre les dimensions affectives et les mesures graphiques a révélé que L est inversement lié aux notes subjectives, alors que E_g, E_l, C et S sont tous corrélés positivement. Une diminution de L et une augmentation de E_g indiquent une augmentation du flux d'informations globales séquentiel et parallèle, ce qui conduit à une meilleure intégration des informations dans les connectomes du cerveau [33]. D'autre part, une augmentation de C et E_l suggère une augmentation de l'efficacité du flux d'informations locales ou de la ségrégation dans les connectomes du cerveau. En fait, les stimuli saillants sont connus pour induire des niveaux d'activation physiologique élevés [67], conduisant ainsi à un traitement plus intégré des informations par le biais de ce qu'on appelle 'neurones d'espace de travail', comme proposé par la théorie globale d'espace de travail [68]. Ensemble, ces deux résultats conduisent à une augmentation globale des propriétés de petit réseau avec l'augmentation de l'activation physiologique. Ainsi, nous pouvons affirmer que les traits caractéristiques basés sur l'EEG encodent des informations reliées aux

FIH significatives car les corrélations observées concordent partiellement avec certaines des études précédentes, et peuvent être utilisées comme traits caractéristiques valables pour la classification des FIH.

Ensuite, la possibilité de caractériser plusieurs percepts de la QE en utilisant des signaux d'EEG est évidente à partir des résultats de classification présentés aux Tableaux 5.1 et 3.2, où des résultats nettement meilleurs que le hasard ont été obtenus. La fusion de décisions des ensembles de traits caractéristiques d'EEG s'est avérée conduire à des améliorations significatives, suggérant ainsi que la fusion de décisions de différentes sources d'informations sur l'EEG améliore la caractérisation des percepts de la QE.

En plus, les traits caractéristiques graphiques théoriques se sont avérés être plus performants que les traits caractéristiques basés sur le spectre de puissance (par exemple, la DRÉ et l'IA), dans la classification des dimensions affectives, pour des vidéos affectives (base de données DEAP), alors que, pour les stimuli TTS, les traits caractéristiques graphiques théoriques et les traits caractéristiques basés sur le spectre de puissance ont montré des performances équivalentes. Cette observation peut être attribuée à la différence de la nature des stimuli qui a induit des modifications dans les états affectifs. La base de données DEAP a utilisé des vidéos musicales affectives comme stimuli, ce qui implique un traitement d'informations visuelles et audio, alors que la base de données PhySyQX a utilisé la TTS comme stimuli, ce qui implique uniquement un traitement auditif. Ceci est également corroboré par le fait que la plupart du temps les mesures de traitement des informations séparées ont été corrélées avec les dimensions de la QE (et affective) pour les stimuli TTS, alors que pour les stimuli vidéo, les mesures de traitement des informations séparées et intégrées ont été corrélées avec les dimensions affectives. De plus, comme indiqué dans [28], les stimuli vidéo couvraient les quatre quadrants d'échelle de valence-activation physiologique (soit, HAHV, HALV, LALV et LAHV), cependant, les stimuli TTS couvraient seulement les quadrants HAHV et LALV. Cela peut conduire à une dynamique différente de traitement des informations corticales.

0.4 Caractériser les FIH à l'aide de SPIRf

Récemment, la SPIRf a émergé comme modalité alternative de neuro-imagerie fournissant des informations complémentaires à l'EEG pour étudier les états mentaux à long terme [36]. La SPIRf offre une bonne résolution spatiale, surmontant ainsi une limitation majeure de l'EEG. De telles qualités de la SPIRf indiquent que l'utilisation d'une ICO basée sur la SPIRf pourrait être efficace pour l'évaluation de la QE objective. Cependant, la SPIRf comme technique en est encore à ses débuts dans le domaine de l'évaluation de la QE et pourrait s'avérer très efficace aux côtés de l'EEG, dans l'élaboration de meilleurs modèles objectifs pour caractériser la QE. À cette fin, nous avons étudié des traits caractéristiques dérivés de la SPIRf pour la caractérisation des FIH. Ceci a été possible en sondant le cortex préfrontal et la tête entière dans l'étude préliminaire et dans la base de données PhySyQX, respectivement.

0.4.1 Sondage du cortex préfrontal - Étude préliminaire

Pour la validation de la SPIRf comme outil menant au développement d'ICO pour l'évaluation de la QE, le cortex préfrontal (CPF) du cerveau a été sondé. Le CPF est responsable de la cognition et la prise de décision peut fournir des indications utiles sur la perception de la qualité de la parole. Plus distinctement, il a été constaté qu'une région du CPF appelée cortex orbito-frontal (COF) s'activait lors des tâches de la prise de décision [69]. Ces résultats ont motivé le sondage de régions telles que les CPF/COF pour obtenir des informations sur les processus de perception de la qualité de la parole humaine en utilisant la SPIRf.

Méthodologie

Quatorze locuteurs parlant l'anglais couramment ont été recrutés pour participer au test d'écoute subjective. Les participants ont testé des stimuli de parole synthétisée représentatifs des systèmes existants. Les données du Blizzard TTS Challenge 2009 ont été utilisées [50]. Les stimuli comprenaient quatre phrases en anglais (neutres dans le contenu) d'une durée de 8 à 10 secondes, ce qui correspond à des réponses d'un système de recommandation de restaurant. Ici, nous avons utilisé les données d'un discours naturel, des systèmes de note de qualité haute (MOS = 3,7) et basse (MOS = 1,9). Les données SPIRf ont été collectées à partir du CPF en utilisant un bandeau SPIRf personnalisé, tandis que les participants testaient les stimuli vocaux. Après avoir écouté les fichiers vocaux, les participants ont noté leur compréhension perçue, la fluidité et la qualité globale, sur une échelle de '1' à '5'.

Après la collecte de données, les données SPIRf ont été nettoyées et certains traits caractéristiques descriptifs de signaux ont été extraits, comme l'amplitude de crête de $\Delta[HbO]$ et son temps de montée correspondant, l'amplitude de la valeur minimale de $\Delta[HbR]$ et son temps de descente correspondant, ainsi que le temps de crête et le coefficient de variation.

Résultats

La Fig. 4.2 présente les traits caractéristiques statistiques exploratoires des données subjectives. Elle montre que les notes MOS et de fluidité ont diminué linéairement suivant la qualité de la parole, alors que la compréhension diminuait de façon non linéaire. Les résultats ANOVA, présentés au Tableau 4.1, montrent qu'il y a un effet principal significatif dans les variables de réponses subjectives sur trois conditions de qualité différentes.

De plus, comme la littérature le suggère, l'augmentation de l'activation du COF était attendue avec l'augmentation de la qualité perçue des stimuli [70]. Cet effet est clairement visible aux Fig. 4.3 et Fig. 4.4, où les amplitudes des crêtes $\Delta[HbO]$ augmentent et les creux $\Delta[HBR]$ diminuent avec l'augmentation la qualité perçue dans la région COF (canaux 14, 16, 17 et 18), ce qui suggère que l'activation de la région a augmenté. Un effet significatif des différents niveaux de qualité des stimuli sur les amplitudes des traits caractéristiques SPIRf a été observé, comme indiqué au Tableau 4.2. Par ailleurs, une statistique F importante ($p < 0,05$) avec η^2 >0.50 a été constatée pour tous les

traits caractéristiques SPIRf à partir de l'analyse de tendance linéaire comme indiqué au Tableau 4.3. Cela confirmait la relation linéaire entre les traits caractéristiques SPIRf et les évaluations de qualité subjectives des stimuli.

Pour évaluer la relation entre les scores subjectifs et les traits caractéristiques de SPIRf, les coefficients de corrélation de Pearson et de Spearman, dénotés par ρ et ρ_{spear}, respectivement, ont été utilisés. Des corrélations modérément élevées et significatives ($p < 0.05$) ont été constatées avec toutes les traits caractéristiques SPIRf et au moins une des trois notes subjectives (MOS, compréhension et fluidité). Le creux $\Delta[HbR]$ du canal 17 a montré la plus forte corrélation de - 0,54 avec MOS. La crête $\Delta[HbO]$ du canal 14 et le creux $\Delta[HbR]$ du canal 17 ont montré la plus forte corrélation de 0,58 et - 0,59 avec la compréhension, respectivement. Ces deux traits caractéristiques ont également été bien corrélées avec la fluidité. Les traits caractéristiques qui étaient significativement corrélés avec les trois notes subjectives sont présentés au Tableau 4.4.

Discussion

À la lumière des résultats obtenus à partir des données subjectives et physiologiques, notre étude a montré une augmentation linéaire de l'activation du COF avec une augmentation linéaire de la qualité TTS. Cette activation différentielle du COF pourrait être attribuée à l'évaluation basée sur la qualité perçue des stimuli de parole dans le cerveau, corroborant ainsi les résultats d'études précédentes [71]. Ces résultats établissent le fait que la SPIRf peut être un outil utile pour caractériser la QE.

0.4.2 Sondage de la tête entière - Base de données PhySyQX

Dans la section précédente, nous avons montré l'importance de la SPIRf et des phénomènes hémodynamiques corticaux pour la mesure de la QE de TTS, en particulier dans la région du cortex préfrontal (CPF) du cerveau qui s'est avérée être associée à la cognition et à la prise de décision [72]. Ici, nous élargissons le travail en sondant l'ensemble de la tête pour tenter de décoder les interactions corticales plus complexes impliquées dans le processus de perception de la QE. De plus, le sondage SPIRf-EEG simultané le long des mêmes régions corticales peut fournir une validation aux observations de l'EEG. Par ailleurs, des mesures physiologiques (par exemple, la fréquence cardiaque) se sont également avérées utiles dans la surveillance des états affectifs, en particulier dans une configuration multimodale [28, 29]. Ici, nous essayons d'extraire des traits caractéristiques basés sur la fréquence cardiaque, à partir de signaux SPIRf, pour caractériser les FIH en utilisant la base de données PhySyQX décrite précédemment.

Méthodologie

Les signaux de SPIRf bruts ont été nettoyés en utilisant des techniques de traitement de signaux simples tels que des filtres passe-bande. Par la suite, nous avons calculé les crêtes, les creux, l'amplitude de montée, l'amplitude de descente et le taux de passage à zéro calculé à partir des

courbes $\Delta[HbO]$ et $\Delta[HbR]$, comme le montre la Fig. 1.7. Cela a abouti à 300 traits caractéristiques (60 canaux $\times$ 5 caractéristiques) pour chacune des courbes $\Delta[HbO]$ et $\Delta[HbR]$. Désormais, ces deux ensembles de traits caractéristiques seront appelés HbO_{temp} et HbR_{temp}, respectivement. De plus, nous avons également calculé les mesures statistiques, telles que la moyenne, la médiane, l'écart type, l'asymétrie et l'aplatissement, sur quatre fenêtres de 5 secondes (d'une durée de 20 secondes post stimuli) ne se chevauchant pas, pour $\Delta[HbO]$ et $\Delta[HbR]$ pour chacun des 60 canaux. Cela a abouti à 1200 traits caractéristiques (60 canaux $\times$ 5 traits caractéristiques $\times$ 4 fenêtres) pour les courbes $\Delta[HbO]$ et $\Delta[HBR]$. Désormais, ces traits caractéristiques sont appelés HbO_{stats} and HbR_{stats}, respectivement. Ensuite, nous avons extrait la fréquence cardiaque et le signal de la variabilité de la fréquence cardiaque du signal SPIRf [73]. À partir de ces deux séries temporelles, deux séries de traits caractéristiques ont été extraites. La première série, désormais appelée 'HRF1', a été calculée en fonction de [28] et se composait de 20 traits caractéristiques. La deuxième série de traits caractéristiques, désormais appelée 'HRF2 ', a été calculée en fonction de [73] et consistait en un total de 9 traits caractéristiques.

Résultats

De la Fig. 4.7 il a été observé que les crêtes $\Delta[Hbo]$ montraient des corrélations positives significatives dans les régions fronto-temporale droite et temporo-centrale gauche du cerveau, avec l'augmentation de la qualité du signal. De plus, la concentration moyenne $\Delta[HbO]$, au cours de la présentation du stimulus, a montré des modèles de corrélation similaires à ceux observés pour les crêtes $\Delta[HbO]$. La moyenne $\Delta[HbR]$ et la concentration des creux $\Delta[HbR]$ ont montré des corrélations négatives significatives, dans les régions fronto-temporale gauche et droite avec l'augmentation de la qualité du signal, de manière cohérente pour toutes les dimensions subjectives.

Le Tableau 4.5 présente les exactitudes moyennes et les valeurs F1 sur tous les participants pour chaque série de traits caractéristiques pour SPIRf, pour chaque dimension subjective. Le tableau présente les meilleurs classifieurs de chaque trait caractéristique de SPIRf ainsi que la fusion de décisions de chaque classifieur élaborée en utilisant différents traits caractéristiques de SPIRf. Les performances pour chaque classifieur ont été testées à l'aide de test T d'échantillons répétés bilatéraux sur les résultats concaténés de chaque échelle de notation et de chaque participant, tel que suggéré par [28]. Il a été observé que la plupart des classifieurs étaient bien plus performants que le hasard. Par ailleurs, il a été observé que les classifieurs basés sur la fusion de décisions donnaient de bien meilleurs résultats que les classifieurs basés sur les SPIRf les plus performants et les traits caractéristiques de HR. De plus, en comparant les performances de chaque SPIRf et les traits caractéristiques de HR, il a été observé que les traits caractéristiques basés sur la SPIRf donnaient de meilleurs résultats que le HR, mais cette découverte n'a pas révélé de différences significatives. Par ailleurs, en comparant les classifieurs de fusion élaborés en utilisant chaque SPIRf et les traits caractéristiques de HR, il a été observé que les classifieurs de fusion basés sur la SPIRf donnaient de meilleurs résultats que les HR et cette constatation a été importante pour la dimension de CP.

Discussion

Les graphiques de corrélation topographiques représentés par la Fig. 4.7 indiquent une activation fronto-temporale droite et gauche plus forte avec l'augmentation de la qualité de la parole; ces résultats étaient cohérents avec les résultats de l'EEG. Par conséquent, la présente étude de corrélat neural impliquant des systèmes TTS était en accord avec les études antérieures sur la perception de la parole, validant ainsi les traits caractéristiques extraits comme corrélats de la perception de QE de la parole synthétisée. La faisabilité de caractérisation de plusieurs percepts de la QE en utilisant des signaux physiologiques est évidente à partir des résultats de classification présentés au Tableau 5.1, où des résultats nettement meilleurs que le hasard ont été obtenus. Fait intéressant, les traits caractéristiques de la fréquence cardiaque dérivés de la SPIRf ont également donné lieu à de meilleures performances que la chance, indiquant ainsi que le bruit dit physiologique de signaux de SPIRf peut être utilisé pour la modélisation de la perception de la QE.

Par ailleurs, la fusion des séries de traits caractéristiques d'une modalité spécifique a démontré qu'elle conduisait à des améliorations significatives pour SPIRf et HR, suggérant ainsi que la fusion de décision de différentes sources d'informations à partir d'une modalité améliore la caractérisation des percepts de QE. De plus, il a été observé que les classifieurs basés sur la SPIRf donnaient de meilleurs résultats que les classifieurs basés sur la HR, ce qui suggère que les signaux SPIRf fournissent des informations plus riches concernant les dimensions de la QE. Aussi, en comparant les performances des traits caractéristiques basées sur l'EEG et la SPIRf, il n'y avait pas de différence significative entre les deux modalités. Cela indique que les deux modalités fournissent la même quantité d'informations pour la caractérisation des FIH. Dans la section suivante, la fusion des deux ICO a été étudiée, dans l'espoir que l'approche des ICO hybrides puisse améliorer les performances du modèle de caractérisation de la QE.

0.5 Caractérisation des FIH en utilisant la fusion multimodale

Récemment, dans le domaine de la recherche sur les facteurs humains, en particulier dans le domaine de l'informatique affective, les chercheurs ont étudié la fusion de mesures de différentes modalités de capteurs grâce à diverses méthodes de reconnaissance de modèles [74]. Il est noté dans [75], qu'une approche multimodale vers la reconnaissance affective conduit à des résultats plus précis. Par conséquent, pour caractériser les FIH avec plus de précision, nous avons étudié la fusion de différentes modalités neurophysiologiques pour développer des ICO hybrides. Nous avons tenté d'élaborer des mesures précises et objectives de FIH en étudiant la fusion de décisions de multiples modalités de signaux de la base de données PhySyQX. Plus précisément, l'EEG, la SPIRf et les mesures de fréquence cardiaque dérivées de la SPIRf sont combinées avec des arbres décisionnels pour mesurer plusieurs dimensions subjectives de la QE. On suppose qu'une exactitude de classification importante peut être obtenue avec les traits caractéristique extraits pour chaque modalité. Ces résultats corroborent l'importance des FIH dans l'évaluation globale de la QE objective des systèmes TTS.

0.5.1 Méthodologie

La fusion de décisions de classifieurs obtenue à partir des différentes modalités (à savoir, la fréquence cardiaque dérivée de la SPIRf, l'EEG et la SPIRf) a été mise en œuvre en utilisant la fusion de décisions pondérée proposée dans [76], en utilisant la fusion de décisions également pondérée. La fusion de décisions également pondérée attribue un poids égal à chaque modalité. La fusion des différentes modalités a été effectuée en deux phases. Dans la première phase, toutes les séries de traits caractéristiques de chaque modalité ont été utilisées pour la fusion, tout en pondérant également chaque trait caractéristique. Cependant, pour la deuxième phase de fusion de la décision, seules les séries de traits caractéristiques les plus performantes (sur l'ensemble d'entraînement) ont été utilisées à partir de chaque modalité.

0.5.2 Résultats

Le Tableau 5.1 présente des niveaux de performance de classification pour la fusion de décisions de *toutes* les séries de traits caractéristiques disponibles pour une modalité donnée et différentes stratégies de combinaison de modalités. Il a été observé que la combinaison de différentes modalités physiologiques augmentait la performance de chacune des modalités pour toutes les dimensions subjectives, sauf Ar, par rapport aux classifieurs développés en utilisant les traits caractéristiques les plus performantes de chacune des modalités. Cependant, l'ANOVA n'a révélé un effet principal significatif que pour les combinaisons 3 (SPIRf et HR) et 4 (EEG, SPIRf et HR) avec Nat ($F(2, 60) = 4.23, p < 0.05$) et MOS ($F(2, 60) = 2.76, p < 0.05$), respectivement. Les tests post-hoc ont révélé des différences significatives entre les classifieurs basés sur les traits caractéristiques de fréquence cardiaque et la fusion de décisions. Enfin, la comparaison des classifieurs mis au point en utilisant des modalités de fusion (présentés dans le Table 5.1) et la fusion des traits caractéristiques de chaque modalité n'a entraîné aucune différence significative.

Enfin, le Tableau 5.2 présente la performance de classification des combinaisons de fusion des séries de traits caractéristiques les plus performantes (sur l'ensemble d'entraînement) de chaque modalité physiologique. Comme on peut le voir, cette stratégie de fusion a amélioré la performance de classification par rapport à l'utilisation de toutes les séries de traits caractéristiques de chaque modalité pour la fusion de décisions (par exemple, comme présenté au Tableau 5.1). En comparant les performances des classifieurs présentés au Tableau 5.2 avec ceux du Tableau 5.1 une analyse de la variance a été réalisée, comme indiqué dans la dernière ligne du Table 5.2. À l'évidence, la comparaison a été significative pour les dimensions subjectives Ac et Ar. Par ailleurs, les tests post hoc ont révélé une différence significative entre la combinaison 1, du Tableau 5.2; la combinaison 1-4, du Tableau 5.1 pour Ac; les combinaisons 1-2 du Tableau 5.2, et la combinaison 1-4 du Tableau 5.1 pour Ar.

0.5.3 Discussion

L'approche ICO hybride conduit à des améliorations significatives par rapport à l'utilisation d'ICO basées sur une seule modalité, suggérant ainsi que la fusion de décisions de différentes sources d'informations améliore la caractérisation des percepts de la QE. Cela met également en évidence la nécessité d'une surveillance physiologique multimodale pour la modélisation de la perception de la QE. Dans l'ensemble, la combinaison de l'EEG avec la SPIRf ou la HR a abouti à des classifieurs plus performants par rapport aux classifieurs développés en combinant la SPIRf et la HR, mais cette observation n'a pas été significative. En général, la combinaison de deux des modalités a donné de meilleurs résultats que la combinaison des trois modalités. Cela pourrait être dû à la pondération inappropriée des trois modalités lorsqu'elles sont combinées et donc de meilleures stratégies de fusion (par exemple, la fusion de poids optimal décrite dans [76]) peut conduire à une amélioration des résultats. Compte tenu de la taille limitée de l'ensemble des données PhySyQX, cela ne pouvait pas être examiné ici. Enfin, la fusion des séries de traits caractéristiques les plus performantes, à partir de chaque modalité, a conduit à des améliorations significatives des performances de classifieur pour les dimensions Ac et Ar, soulignant ainsi la nécessité d'intégrer des séries de traits caractéristiques optimales, de chaque modalité, pendant le développement des classifieurs de fusion de décision pour la caractérisation des FIH.

0.6 Caractérisation de la QE en utilisant une ICO hybride passive

La majorité des modèles de QE objectifs de pointe comptent sur les aspects technologiques et contextuels du service [7, 77]. Toutefois, afin de développer des modèles de QE vraiment 'centrés sur les utilisateurs', des FIH, comme les émotions et les attitudes des utilisateurs, doivent également être incorporés. Pour le développement de ces modèles de QE objectifs, des ICO peuvent être utilisées. Ainsi, les sections précédentes ont établi l'utilisation des ICO pour caractériser les informations relatives aux FIH des utilisateurs, en utilisant des classifieurs binaires. Cependant, l'objectif de modèles de QE objectifs de pointe est de quantifier la QE. Par conséquent, pour le développement de modèles objectifs basés sur des ICO hybrides qui quantifient la QE, tout en intégrant des FIH, une analyse de régression doit être entreprise pour estimer la QE sur une échelle continue.

Dans une première étape pour l'intégration des FIH dans les modèles de QE objectifs, nous avons étudié l'utilisation des ICO dites hybrides passives. Les ICO hybrides passives caractérisent les états affectifs des utilisateurs en utilisant la fusion de multiples modalités neurophysiologiques et physiologiques [78]. Plus précisément, les facteurs humains affectifs ont été choisis car ils sont des indicateurs significatifs des deux constructions perceptives, 'plaisir d'écoute' et 'prosodie', de la QE de TTS, tels qu'établis dans **Chapitre 2**. Fondamentalement, les dimensions affectives de valence et d'activation physiologique indiquent les changements dans le 'plaisir d'écoute' et la 'prosodie', respectivement.

Ici, nous sommes intéressés par l'étude de l'utilisation des ICO hybrides, basées sur la fusion de la fréquence cardiaque dérivée de la SPIRf, l'EEG et la SPIRf, pour mesurer les FIH, et les

incorporer dans des modèles objectifs de perception de QE de la parole. Un scénario basé sur des systèmes de synthèse texte-parole (TTS), basé sur la base de données PhySyQX, est étudié. Le schéma général du système ICO hybride proposé pour la surveillance de la perception de la QE des utilisateurs est présenté à la Fig. 6.1. Dans ce cadre, le signal audio, généré par un système TTS, est utilisé pour extraire des facteurs d'influence technologiques, alors qu'une ICO hybride collecte des facteurs influents humains implicites. Ainsi, le modèle de QE perçue par l'utilisateur final est obtenu sous forme d'une combinaison de FIT mesurée et de FIH. La QE mesurée offre aux fournisseurs de services des renseignements précieux qui leur permettent d'affiner les paramètres de service, conduisant ainsi à l'amélioration de l'expérience de l'utilisateur. Par conséquent, ici, nous avons proposé une approche nouvelle de la ICO qui intègre des mesures objectives de FIH basées sur la neurophysiologie dans les modèles objectifs de pointe pour la mesure de la QE, entraînant ainsi des gains importants en performance. Par ailleurs, ce chapitre compare les performances des modèles ICO objectifs, basés sur des modalités neurophysiologiques individuelles et le modèle de ICO hybride développé en utilisant la fusion des modalités individuelles.

0.6.1 Méthodologie

Des recherches antérieures ont montré l'importance des mesures basées sur les signaux [15], telles que la prosodie et l'articulation [79] pour la modélisation objective de la QE de TTS. Récemment, deux paramètres quantitatifs se sont montrés utiles [79], et sont donc utilisés dans notre étude TTS : la pente de la dérivée de second ordre de la fréquence fondamentale ($SF0''$) et la moyenne absolue du coefficient de spectre de fréquence de second ordre ($MFCC_2$). Alors que le trait caractéristique $SF0''$ modélise les propriétés macro-prosodiques ou liées à l'intonation de la parole, $MFCC_2$ modélise les propriétés liées à l'articulation [79]. Dans nos expériences, la boîte à outils openSMILE [80] a été utilisée pour extraire ces traits caractéristiques en utilisant la longueur de fenêtre par défaut de 25 ms et le décalage du cadre de 12,5 ms.

Pour le développement d'ICO hybrides, nous avons tiré parti des données acquises de l'EEG et de la SPIRf. Par conséquent, sur la base des résultats de l'analyse de corrélation, les traits caractéristiques qui ont montré une corrélation maximale avec la valence et l'activation physiologique subjectives ont été utilisées comme corrélats neurophysiologiques des états affectifs. Ainsi, pour les signaux EEG, les mesures d'efficacité locale (E_L) dérivées de la sous-bande bêta supérieure, et la puissance bêta médiale (PBM) [48], ont été utilisées pour modéliser la valence et l'activation physiologique, respectivement. Pour les signaux de SPIRf, le $\Delta[HbR]$ moyen calculé à partir de la région temporale droite a montré une corrélation maximale avec la valence et le $\Delta[HbO]$ moyen à la région temporo-pariétale a montré une corrélation maximale avec l'activation physiologique. Enfin, pour le signal de fréquence cardiaque, la HRV moyenne s'est avérée corrélée avec la valence, alors qu'aucune des caractéristiques dérivées de la fréquence cardiaque n'a montré de corrélation avec l'activation physiologique. Par conséquent, les traits caractéristiques neurophysiologiques mentionnés ci-dessus ont été utilisés pour développer des modèles d'ICO basés sur chaque modalité individuelle, ainsi que d'un modèle d'ICO hybride.

Afin d'évaluer la performance du modèle de la QE, six tests ont été effectués. Tout d'abord, nous avons étudié la qualité de l'ajustement (r^2) obtenue en utilisant uniquement la mesure du discours centrée sur la technologie comme corrélat du score de QE indiqué par les auditeurs (désignés comme QE_{Tech}). Deuxièmement, nous avons étudié les gains obtenus en incluant les FIH dans les modèles de la QE. Ici, nous avons mesuré les r^2 obtenus à partir d'une combinaison linéaire de mesure de discours centrée sur la technologie combinée avec les notes de valence et d'activation physiologique subjectives ('vérité du terrain') rapportées par les auditeurs (désignées comme QE_{FIH}). Ensuite, nous avons remplacé les FIH de vérité de terrain par les traits caractéristiques de l'ICO, sur la base de chaque modalité individuelle, qui sont utilisés comme corrélats des états émotionnels des auditeurs (désignés comme QE_{EEG}, QE_{SPIRf} et QE_{RH}). Enfin, nous avons remplacé les FIH de vérité de terrain par les traits caractéristiques d'ICO hybrides (combinaison des traits caractéristiques d'ICO individuelles). Il est prévu qu'une augmentation de r^2 peut être obtenue en utilisant des modèles d'ICO hybrides, établissant ainsi l'utilisation de modèles d'ICO hybrides.

0.6.2 Résultats expérimentaux

Comme mentionné ci-dessus, trois modèles de la QE ont été mis en œuvre afin d'évaluer les avantages d'inclure des FIH, ainsi que des traits caractéristiques d'ICO dans l'équation. Pour cette étude, les modèles de la QE présentés au Tableau 6.1 ont été trouvés. La qualité d'ajustement obtenue de la valeur (r^2) pour le modèle 1 était de 0.76 avec une racine carrée de l'erreur quadratique moyenne (REQM) de 0.136. Pour le modèle 2, à son tour, la valeur r^2 obtenue était de 0.96 avec une REQM de 0.05, ce qui souligne encore une fois l'importance des FIH dans la modélisation de la perception de la QE. Pour les modèles basés sur les traits caractéristiques d'ICO dérivés de modalités physiologiques individuelles, les valeurs r^2 obtenues étaient de 0.87, 0.84 et 0.79 pour l'EEG, la SPIRf et la fréquence cardiaque, respectivement, obtenues à partir de modèles 3, 4 et 5. Enfin, pour le modèle d'ICO hybride 6, la valeur r^2 obtenue était de 0.90 avec un REQM de 0,10. Lorsque l'on compare le résultat du modèle de QE objectif en 6 et le modèle en 2, un coefficient de corrélation de Pearson de 0.96 est obtenu.

0.6.3 Discussion

Récemment, la caractérisation des FIH et des FIH objectifs a attiré l'attention croissante des chercheurs en QE [81, 82, 83]. Dans la même veine, nous avons évalué les effets des états affectifs des utilisateurs sur la perception globale de la QE. A partir des tests d'évaluation subjective, nous avons trouvé des preuves, qu'en effet, les états affectifs perçus des utilisateurs changeaient avec la qualité variable de la parole. Plus précisément, lorsque des FIH étaient combinés avec des mesures existantes de qualité de parole centrée sur la technologie, comme dans le modèle 2 du Tableau 6.1, des améliorations des performances de mesure de la QE ont été observées et un gain relatif de 26,3% a été constaté pour les systèmes TTS. Ces résultats suggèrent que les états affectifs peuvent en effet influer directement sur l'expérience perçue des auditeurs (ou QE) avec des synthétiseurs vocaux. Néanmoins, en dépit des améliorations constatées lors de l'ajout des FIH à des modèles de qualité

objectifs (par exemple, le modèle 2 du Tableau 6.1), il y avait encore un écart jusqu'à la qualité d'ajustement parfaite, suggérant ainsi que l'inclusion de FIH supplémentaires alternatives peut être importante. À cette fin, les études futures devraient étudier les effets, par exemple, de l'attention, de la charge cognitive, de la fatigue et / ou de l'engagement des utilisateurs.

Pour l'amélioration des performances des modèles basés sur les modalités neurophysiologiques individuelles, une approche basée sur des ICO 'hybrides' a été utilisée dans le modèle 6, présenté au Tableau 6.1. Le modèle 6 a donné de meilleurs résultats que les modèles basés sur les ICO développés en utilisant des modalités individuelles, cependant, le modèle 2 a donné de meilleurs résultats que le modèle 6. L'écart observé entre les modèles de la QE trouvés avec des traits caractéristiques subjectifs et d'ICO hybrides peut être attribué à des corrélations moyennes faibles obtenues entre les traits caractéristiques neurophysiologiques et la valence et l'activation physiologique subjectives. Dans l'ensemble, il est prévu que des modèles plus puissants pourront être obtenus une fois que des traits caractéristiques améliorés seront développés. Alternativement, d'autres modalités de signaux neurophysiologiques peuvent être incorporés pour la surveillance de l'état affectif humain, tels que la réponse électrodermale et l'oculométrie. Le développement de ces ICO 'hybrides' est le but de notre recherche actuelle.

0.7 Conclusion

Dans cette thèse, nous avons démontré l'utilisation d'ICO basées sur l'EEG et la SPIRf pour la caractérisation de la QE. Plus précisément, les traits caractéristiques d'EEG basés sur le spectre de puissance et le spectre croisé ont été étudiés pour caractériser la QE. Les traits caractéristiques spectraux croisés se sont révélés être mieux adaptées à des stimuli complexes. D'autre part, les traits caractéristiques basés sur la SPIRf ont montré des performances équivalentes par rapport aux traits caractéristiques basés sur l'EEG. De plus, nous avons établi la supériorité de l'approche des ICO hybrides pour caractériser la QE. Enfin, les traits caractéristiques basés sur l'EEG, la SPIRf et une ICOh ont été incorporés dans les modèles d'évaluation de la QE objectifs de pointe. Fait intéressant, avec des caractéristiques basées sur des ICOh une amélioration d'environ 18% a été obtenue. Cependant, le modèle développé a montré des performances moindres par rapport au modèle développé en utilisant des scores subjectifs, indiquant ainsi une possibilité d'amélioration du modèle. À cette fin, les études futures peuvent se concentrer sur le développement de meilleures traits caractéristiques basées sur l'EEG et la SPIRf, des techniques de fusion EEG et SPIRf et des systèmes adaptatifs de qualité du signal.

Chapter 1

Introduction

With the burgeoning multimedia communications industry, new services are emerging increasingly. To be successful, these services have to undergo continuous performance evaluation to ensure good *quality* of the delivered content. Previously, the term *quality* had been used by engineers to describe the so-called *'Quality-of-Service'* (QoS) parameter, which is defined by the International Telecommunications Union (ITU) as the *'totality of characteristics of a telecommunications service that bear on its ability to satisfy stated and implied needs of the user of the service'* [3]. Thus, it was assumed that, for a particular service, higher QoS would lead to its higher acceptability and hence, have more users. However, the success of some services despite having low *quality*, such as the early SMS systems, sparked the need to understand *quality* from the perceiving users' standpoint. This led to formulation of the term *'Quality-of-Experience'* (QoE), which takes the user's perceptual and judgment process into consideration. As recently defined by the Qualinet group, however, *'Quality of Experience (QoE) is the degree of delight or annoyance of the user of an application or service. It results from the fulfillment of his or her expectations with respect to the utility and / or enjoyment of the application or service in the light of the user's personality and current state'* [4]. In this definition of QoE, the 'personality' is used in terms of 'those characteristics of a person that account for consistent patterns of feeling, thinking and behaving' and 'current state' in terms of 'situational or temporal changes in the feeling, thinking or behavior of a person'. Also, in this definition, the current state is both an influencing factor of QoE and a consequence of the experience [4].

The role of the users' current state as an influencing factor and a consequence of the experience is further evident in the quality formation process, shown in Fig. 1.1. In this figure, the perceived

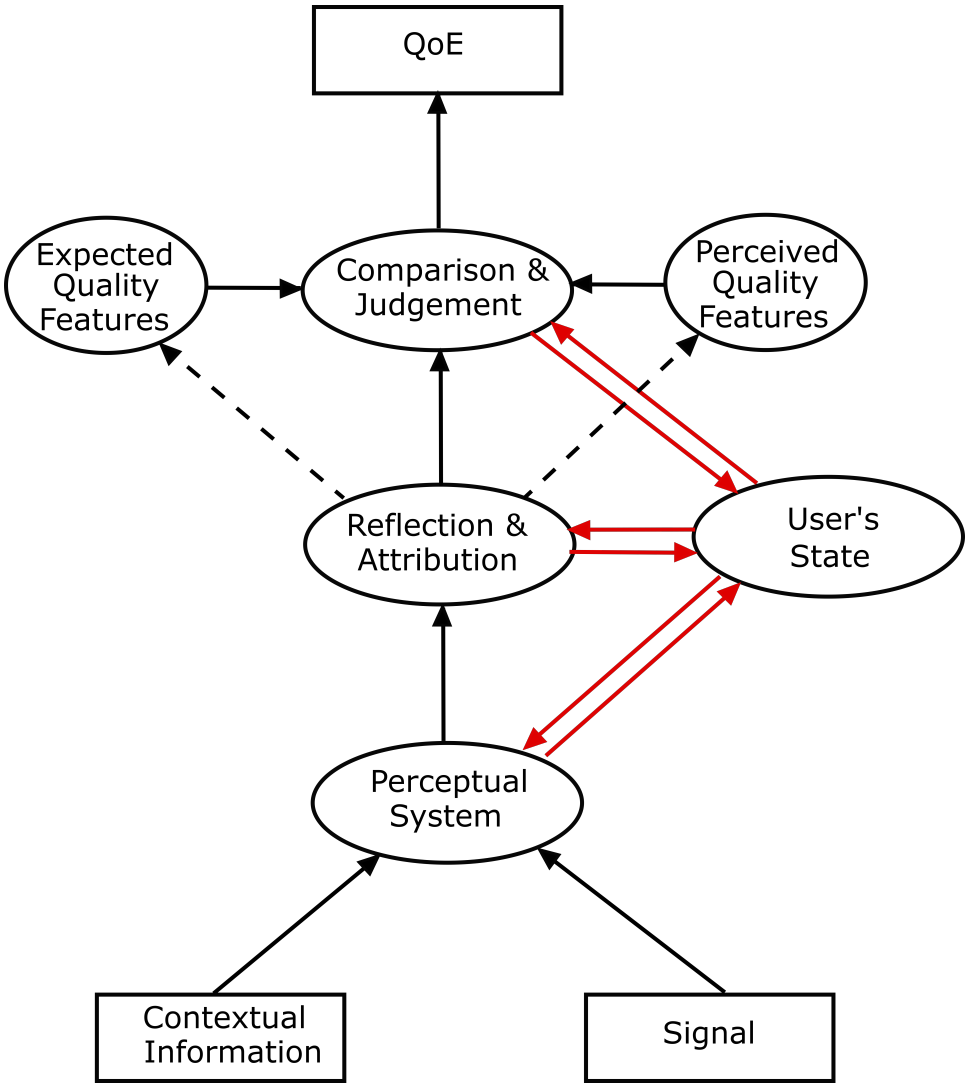


Figure 1.1 – The figure depicts a simplified version of quality formation process, modified from [1]. The comparison of users' expectations and perceived event affects their QoE perception. The (cognitive or affective) states of the user can influence all the stages of QoE formation and can be influenced by the same.

QoE is the output of a quality formation process that involves a quality awareness phase that results in identification of emotional, sensory and conceptual quality features, through reflection and attribution and then compares the expected and experienced quality features to form the QoE. The QoE specifically addresses the perceptions and experiences of the users that are deemed more appropriate for the design of services with higher acceptability. Therefore, a better QoE ensures the competitive edge and ultimately, the success of that service.

Traditionally, the aspects of QoE can be assessed using either perception-based (subjective) or instrumental (objective) methods. The subjective assessment methodologies are the most valid way

to assess the QoE. These methods require human evaluators to gather QoE related information for different multimedia stimuli. Typically, for subjective assessment the users interact with the multimedia stimuli and subsequently provide quantitative ratings of momentary and remembered QoE on a set of scales. Such subjective methods also provide the ground-truth data for the development of objective methods for QoE assessment. Objective methods, on the other hand, estimate the QoE using an algorithm or instrument. These algorithms are generally fed by a set of features, derived from the technical system or the multimedia signal under consideration, that are used to develop models to estimate the ground-truth data provided by the subjective methods. The objective models are low-cost and a quicker alternative to subjective QoE assessment methods. Therefore, recently, the main focus of the researchers has been on the development of new objective characterization techniques for QoE to automate/expedite the quality assessment process.

However, for the development of the objective QoE assessment models, it is important to identify the attributes that influence the perception of QoE. Such attributes are called 'Influence Factors' (IFs). These are highly complex and inter-related aspects of the system, environment or the user himself. As reported in [4], the IFs can be broadly categorized into technological, context and human IFs. The technological IFs (TIFs) refer to the system and network parameters that can be readily measured (e.g., delay, bitrate). The contextual IFs (CIFs) encompass the situational properties that describe user's environment using its physical, temporal, social, economic, task or technical characteristics. Lastly, the human IFs (HIFs) enlist any variant or invariant characteristics of the human user, such as their emotional state, preferences, attitudes and goals, along with many other subjective factors (e.g., fatigue). Over the last decade, researchers have focussed on the TIFs and CIFs to develop new objective models. However, the predictors from these techniques do not represent the true perception of QoE as they lack inputs from the HIFs. Fundamentally, the incorporation of HIFs into the objective QoE models can make the models more user-centric in nature. Moreover, it is evident from Fig. 1.1 that human factors, represented by users' internal states, modulate users' QoE scores. Hence, interdisciplinary research is still needed to bridge this gap through incorporation of HIFs constructs into new objective characterization techniques.

Towards this end, it is important to notice that the basic facets of QoE, specifically HIFs, are subjective in nature and are not directly observable, as is evident from the above discussion. This is because most of the quality formation and judgment process takes place inside the user's brain. Thus, probing the neuronal (electrical impulses) or haemodynamic (blood flow) activities

of the brain are expected to provide significant information regarding quality judgment processes and HIFs. Ultimately, these insights could be used to develop better objective characterization techniques for QoE. The neurophysiological modalities, such as electroencephalography (EEG) and functional near infrared spectroscopy (fNIRS), can probe neuronal and cerebral haemodynamic activities. Specifically, the neurophysiological modalities can act as an interface between users' internal states and the objective QoE assessment models. This system forms the so-called brain computer interface (BCI), which is defined as *'a system that measures central nervous system (CNS) activity and converts it into artificial output that replaces, restores, enhances, supplements, or improves natural CNS output and thereby changes the ongoing interactions between the CNS and its external or internal environment'* [5]. More recently, a new approach called hybrid BCI (hBCI) has emerged, which seeks to fuse different physiological modalities, such as heart rate, skin conductance or skin temperature, along with at least one neurophysiological modality [6]. Such BCIs can be leveraged to estimate HIFs, and incorporated into state-of-the-art objective QoE assessment models.

Recent QoE studies have explored the use of EEG as a BCI tool for neurophysiological evaluation of QoE [27, 84, 85]. These studies leveraged the so-called 'event-related potential' (ERP), which is a time and phase locked change in the voltage amplitude of the EEG time series due to changes in neuronal populations, in response to a sensory, cognitive or motor event [25, 56]. This assessment methodology assumes that the distortions that affect the perceived QoE are time-locked to the beginning of the multimedia signal. While this may be a valid assumption when testing, e.g., the effects of ambient noise [84], it is not the case with most of the multimedia signals, such as TTS speech, which could be comprised of time-varying distortions. In such cases, longer duration test signals are needed. Therefore, in this doctoral thesis we have proposed new EEG-based features that can be used to accurately characterise QoE of multimedia signals of longer duration. Moreover, we have also explored fNIRS-based features for characterising the QoE of multimedia signals, as fNIRS has recently emerged as an alternate neuroimaging modality providing complementary information to EEG for studying long duration multimedia signals [36]. Finally, we have also proposed a hBCI system that fuses information from multiple neurophysiological sources to estimate HIFs, and then incorporates them into a state-of-the-art objective QoE model.

The remainder of this chapter is organised as follows: section 1.1 describes the state-of-the-art QoE assessment methods, section 1.2 describes the methodology for BCI-based QoE assessment.

In section 1.3, we have listed the major contributions of this thesis. Finally, in section 1.4 we have described the organisation of this thesis.

1.1 State-of-the-art QoE Assessment Methods

QoE can be assessed either subjectively or objectively [7, 8]. Subjective testing typically involves user interviews, ratings and surveys to obtain insights about the end-user's perception, opinion and emotions about multimedia quality and their overall experience, thus forming the 'ground truth'. Objective assessment, on the other hand, replaces the listener with a computational algorithm that has learned complex mappings between several key factors and previously-recorded subjective ratings. The following sections describe the state-of-the-art QoE assessment methods.

1.1.1 Subjective Assessment Methods

Quantitative subjective assessment methods typically involve the construction of questionnaires with rating scales, surveys, and user studies that can be conducted either in laboratory or "real-world" settings. Subjective assessment methods can be broadly classified into two categories: with reference and without reference tests. The 'with reference' subjective tests involve asking the participants to rate the signal quality in comparison to a reference sample, for example comparison category rating (CCR) and degradation category rating (DCR) [86]. For signals with high quality, a subjective test with a reference stimulus is more suitable. However, the most commonly implemented class of subjective tests are the 'without reference' tests, which are conducted without a reference sample, for example absolute category ratings (ACR) tests. The International Telecommunications Union (ITU) has developed subjective study guidelines for perceptual quality evaluations. Specifically, for different applications, ITU has different recommendations, such as ITU-T Rec. P.800 for speech quality, P.910 for video quality, P.911 for audiovisual quality and P.85 for text-to-speech (TTS) systems [9]. In accordance with these recommendations, test participants score the QoE percepts on ACR scales that result in a mean opinion score (MOS), which is expressed on a scale between 1 (bad) and 5 (excellent). For text-to-speech systems, for example, ITU-T Rec. P.85 recommends using 9 different 5-point rating scales, such as overall impression and voice pleasantness, that range from bad to excellent, and very pleasant to very unpleasant, respectively. Together, the

different subjective scales can be used to extract (latent) perceptual constructs of QoE using various techniques, such as multidimensional scaling (MDS) [65] and factor analysis [87, 88, 53, 89]. Fundamentally, the latent perceptual constructs are the internal attributes that influence the surface attributes in a systematic manner, thus, measurements obtained from subjective indicators are, at least in part, the result of the linear influence of the underlying latent factors [90]. Therefore, such analyses provide an in-depth understanding of the perceptual processes related to QoE.

Towards this end the ITU recommendations have mainly focussed on utilizing attitudinal HIFs, such as voice pleasantness and acceptance scales for P.85 recommendations [91]. However, human affective states are also important HIFs that influence the perception of QoE [92, 93, 94], and must be subjectively evaluated along with attitudinal HIFs. As such, affect is defined as the experience of feeling or emotion [95], and in the affective computing domain, human affect is considered to manifest itself through multifaceted verbal and non-verbal expressions. Therefore, one common approach is to categorize affective factors using two broad dimensions comprising valence (V) and arousal (A) on two-dimensional plots [10]. Valence refers to the (un)pleasantness of an event, whereas arousal refers to the intensity of the event, ranging from very calming to highly exciting [96, 11]. Using the valence-arousal (VA) model, various emotional constructs have been developed, as depicted by Fig. 1.2 [11, 10, 97]. As depicted by the VA model, in Fig. 1.2, various emotions lie in four different quadrants namely, high arousal high valence (HAHV), high arousal low valence (HALV), low arousal high valence (LAHV) and low arousal low valence (LALV). In order to quantitatively characterize these two emotional primitives (i.e., valence and arousal), the Self Assessment Manikin (SAM) pictorial system is commonly used, as shown in Fig. 1.3 [11, 10]. As can be seen, the SAM for valence ranges from a smiling, happy manikin to a frowning, unhappy one. For arousal, in turn, SAM ranges from very excited, eyes-open manikin to a sleepy, eyes closed one [51]. It is important to emphasize that a third dimension, dominance, has also been proposed and refers to the controlling/dominant nature of the felt emotion [96]. While dominance has shown to be useful in characterizing emotions felt by subjects viewing pictures [98] and watching movies [99], it has shown limited use with speech stimuli.

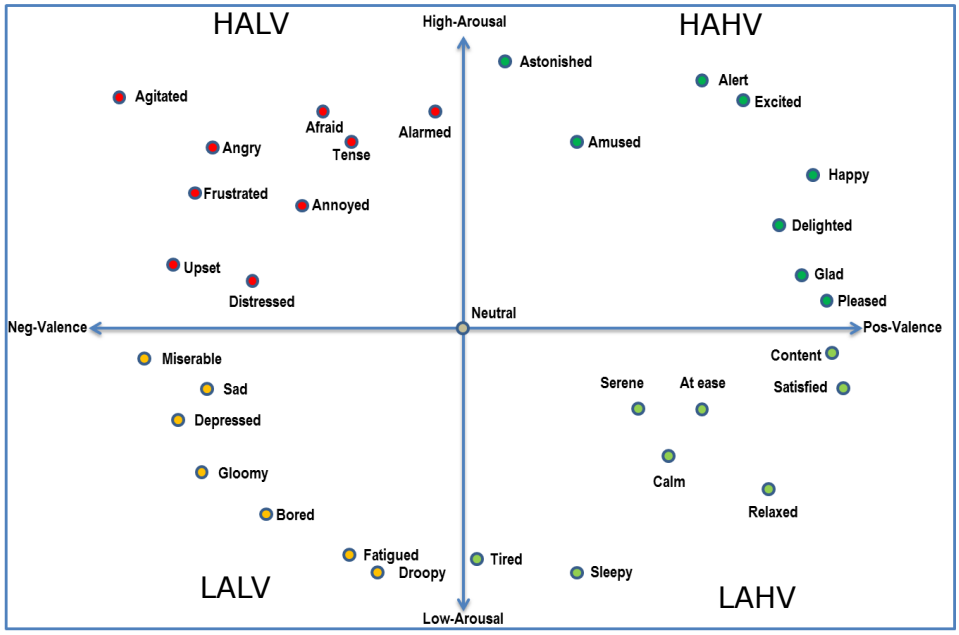


Figure 1.2 – This figure shows a two-dimensional Valence-Arousal (VA) emotion map with representative emotions.

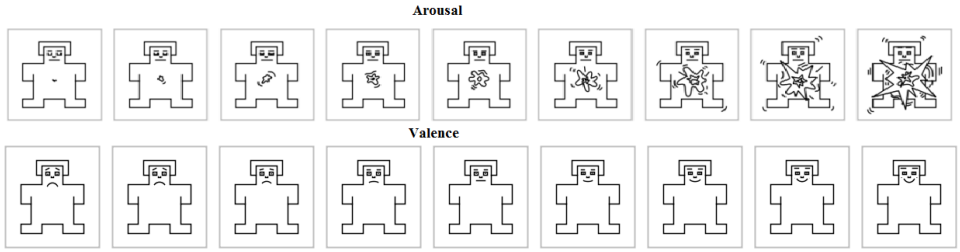


Figure 1.3 – This figure shows two self assessment manikin scales for emotion assessment; top: Arousal; bottom: Valence.

1.1.2 Objective Assessment Methods

Objective assessment methods are also often referred to as instrumental measures. Objective methods usually leverage the technology-centric metrics to estimate QoE. Technology-centric models replace the human rater by a computer algorithm which has been developed to extract relevant features from the analyzed signal (speech, audio, image, or video) and map a subset/combination of such features into an *estimated* QoE value. For speech technologies, models can be further categorized as full-reference (also known as double-ended, intrusive) or no-reference (single-ended, non-intrusive), depending on the need, or not, of a reference signal, respectively. The ITU, for example, has standardized several objective models over the last decade, such as PESQ (Recom-

mendation P.862 [12]) and POLQA (Recommendation P.863 [13]) as full-reference models and ITU Recommendation P.563 [14] as no-reference.

As mentioned above, existing objective methods have been "technology-centric", thus relying mostly on technological factors [7]. In order to develop QoE assessment methods, however, contextual and human influential factors also need to be incorporated. There have been some attempts to incorporate contextual factors into the objective QoE models, such as the model described in [15], which uses certain characteristics of the signal to estimate the effects of room reverberation on the speech signal. However, the incorporation of human factors has not seen similar attempts mainly because the HIFs are formulated inside the users' brain, thus are not directly observable. Towards this end, probing brain activity using BCI-based (or neuroimaging) tools is expected to provide insightful objective estimation for HIFs. The next section describes the methodology and tools needed for BCI-based assessment of QoE.

1.2 BCI-based QoE Assessment

Recently, BCI-based objective modelling of human QoE perception has gained much ground. These techniques could provide a viable alternative to existing objective QoE predictive models or aid the existing models in better prediction of quality. This idea stems from the fact that such techniques directly measure neuro-physiological activity. As most of the quality judgment process takes place inside the user's brain, these could provide better approximation of the human perceived QoE.

BCIs can be broadly classified into three categories: active, reactive and passive [16]. Active BCIs derive their outputs from brain activity that is directly and consciously controlled by the user for controlling an application. Some of the prominent examples of active BCIs are the basket paradigm [100] and the Hex-O-Spell [101]. Reactive BCIs derive their outputs from brain activity arising in reaction to external stimulation that is indirectly modulated by the user for controlling an application, for example the P300 speller [102] and steady-state visual potentials-based systems [103]. Passive BCIs, in turn, derive their outputs from arbitrary brain activity that arises without the purpose of voluntary control, with the aim of enriching human-computer interaction using the implicit information regarding the user's actual state. For example, passive BCIs have been used

for monitoring mental fatigue and working memory load estimation [104]. As such, the proposed BCI system for QoE assessment falls under the premise of passive BCIs.

A typical BCI system consists of multiple modules that form a loop between the user and the computer or machine [6], as depicted by Fig. 1.4. First, a brain imaging modality, such as EEG or fNIRS, is used to collect information regarding neuronal or humoral activity of the brain. For hBCI systems, the supplementary physiological modalities, such as electrocardiography (ECG), photoplethysmography, galvanic skin response (GSR) sensors and eye tracker, are used additionally. In the next step, the acquired brain activity signals are processed following three general steps of signal preprocessing, feature extraction and feature translation. The signal preprocessing step involves correction or rejection of signal components that do not convey the desired information and arise from other physiological processes, such as blood circulation, respiration, eye blinks, etc. Next, the clean signals are used to generate features that are relevant to the application of interest, using feature extraction techniques. The so-called features are expected to reliably indicate the presence or absence of a phenomenon in the user's brain activity. The last step in the processing stage is the so-called feature translation, which consists of making sense of the extracted features using different classifiers or a set of predefined rules. Several classifiers, such as decision trees [17, 18] and support vector machines (SVM) [19, 20, 21], can be used to translate features into decisions. Furthermore, the SVM classifier can be converted into a relevance vector machine (RVM) classifier [22, 23] to provide probabilistic inference regarding the classifier output. Finally, the decision regarding the user's current mental state is fed into a control device, such as display, robot or a smart-phone, that provides a feedback to the user. Typically, for passive BCIs, the so-called feedback step can be very subtle as the main goal of a passive BCI is to enrich the human-computer interaction and not voluntarily control an application.

Specifically, in the context of passive hBCIs for QoE assessment, the brain activity monitoring can be achieved using techniques such as: electroencephalography (EEG), magnetoencephalography (MEG), functional magnetic resonance imaging (fMRI), functional near infrared spectroscopy (fNIRS) etc. Usually, the electrical activity of the brain is recorded using the EEG and MEG, whereas the cerebral hemodynamics is captured using fMRI and fNIRS. Techniques such as fMRI and MEG provide better spatial resolution, as well as probe deeper into the human brain as compared to EEG and fNIRS. But EEG and fNIRS provide better temporal resolution and are more portable and inexpensive. Given their lower cost, portability and ease of use, EEG and fNIRS have

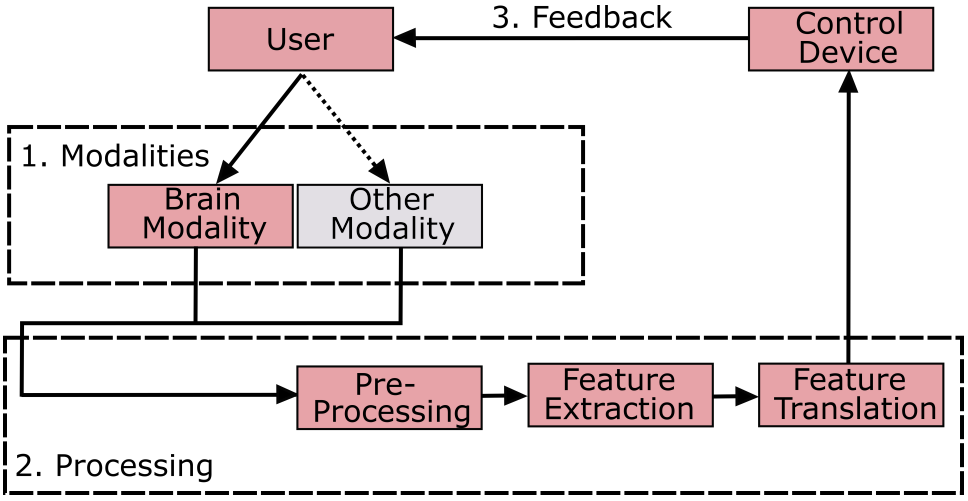


Figure 1.4 – Structure of a standard hybrid BCI.

found more use in the QoE studies lately [84, 105, 106]. Moreover, recently, signals characterizing the effects of peripheral autonomous nervous system (PANS), such as electrocardiography (ECG), photo-plethysmography and galvanic skin response (GSR), have also shown promise towards users' states assessment. The description for each technique is provided below.

1.2.1 Electroencephalography (EEG)

The activities of the brain are maintained via billions of neurons (nerve cells). These cells manage this feat by gathering and transmitting various electro-chemical signals. Thus in effect, acting as dynamically oscillating batteries which produce electrical currents throughout the brain [107]. Each neuron receives a vast amount of information from other neurons or sensory cells [108]. The transfer of information between neurons take place via. specific chemical molecules, known as neurotransmitters. These neurotransmitters are released from one neuron and are received by downstream neurons. As such, the neurotransmitters can either inhibit, excite or modulate the downstream neuronal activity. At any moment, a particular neuron receives and integrates information from many different neurons and fire only if the inputs exceed the neuron's threshold potential, known as action potential. This action potential results in propagation of an electrical current along the main body of the neuron, the axon. Once the current reaches the neuronal terminal, it results in the release of neurotransmitters onto the downstream neurons, and so on. Inside the cortex, the outer layer of the brain, neurons exist in a highly layered structured. Due to this layered architecture, in the cortical regions where the orientation of a population of the neuronal

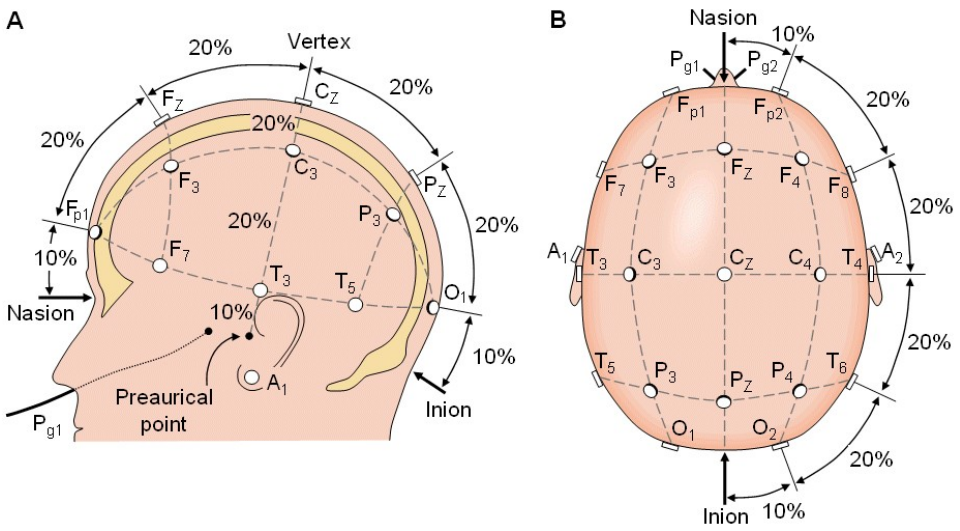


Figure 1.5 – The International 10-20 system. This standard positioning system for EEG electrodes is widely used to ensure good reproducibility between experiments and across subjects. A View of the left side and B of the top of the head. Letters refer to the different brain regions (F: frontal, C: central, T: temporal, P: parietal and O: occipital). [2]

axons get aligned perpendicular to the scalp, the electrical potentials of the neurons add up instead of cancelling out. Once this neuronal firing synchronizes, an electrical potential of approximately 100μV can be measured on the scalp.

Data Acquisition

The electrical field potentials are recorded using EEG by placing many electrodes on the head surface. For EEG data acquisition using electrodes, two contact points are necessary to pick up the signal, one of which is used as a reference. Therefore, for the sake of reproducibility of recordings across the same subject, and across subjects with different head sizes and shapes, different conventions for electrode placement have been adopted. The most common electrode placement convention is the so-called international '10-20 standard' system [24]. As depicted by Fig. 1.5, the '10-20' system consists of electrodes positioned in 10% and 20% marks along the lines that link the front (the nasion, Nz) and the back of the head (the inion, Iz), and the one linking the left and right sides of the head (A1 on the left and A2 on the right). Furthermore, the extension of the 10-20 system, the so-called 10-10 system, allows increments of 10% only, thus providing a higher number of recording sites.

The EEG electrodes are placed over the scalp with a conductive gel, paste or saline solution to reduce impedance between sensors and the skin. Such electrodes are known as wet electrodes. However, a new class of electrodes, the so-called dry electrodes, are slowly gaining recognition, as they have the advantage that the small metal pins directly contact the scalp without any abrasion fluid, such as conductive gel or paste, thus improving the wearability of such electrodes and reduced experimental setup time. However, dry electrodes are sensitive to motion and thus can result in lower quality EEG signals [109]. Moreover, another important issue to be considered while recording the EEG data is the sampling rate. According to the Nyquist theorem, the sampling rate should be twice that of the highest frequency of interest, which can change depending on the application. Commonly, EEG signals are recorded at a sampling frequency of 512 Hz, as the majority of the useful EEG frequency content lies below 200 Hz [110].

Preprocessing

Following the acquisition of EEG data, the data is preprocessed to remove the artefacts (noise) and attain a high signal-to-noise ratio (SNR). EEG signals require amplification of very low amplitude electrical signals, thus the EEG signals can be strongly corrupted by electromagnetic noise, such as powerline interference or other sources of electrical noise such as electrical appliances and electronics. EEG signals are also sensitive to various unwanted artefacts that arise from different physiological processes, such as eye blinks, eye movements, beating of the heart, muscle movement, teeth clenching or even tongue movement. Such artefacts and noises, which corrupt the EEG signals, must be eliminated in order to extract useful information from the signals.

Various signal preprocessing techniques can aid in reducing the effects of such physiological artefacts and noises. As such, powerline interference can be eliminated using a notch filter with a cut-off frequency of 60 Hz or 50 Hz, depending on the country where the signals were recorded. Furthermore, physiological artefacts are reduced from the recorded signals using the independent component analysis (ICA) technique, where the EEG signals are decomposed into separate source signals, and only the signals of interest are retained while the signals from other sources are discarded [111]. This technique effectively removes eye blink, eye movement and muscle related artefacts. Traditionally, ICA is carried out in the supervision of EEG experts; however, recently automated or semi-automated methods for ICA have gained prominence, such as ADJUST [112], AWICA [113].

Ultimately, the signal preprocessing techniques aid in achieving higher signal-to-noise ratio (SNR) and once the achieved SNR reaches a satisfactory level, further processing of data is carried out to extract certain signal descriptive features to characterize an event.

Feature Extraction

The most common EEG feature extracted from the processed data is the so-called 'event-related potential' (ERP). An ERP is a change in the voltage amplitude of the EEG time series (temporal) data in response to a sensory, cognitive or motor event [25]. The so-called 'P300' is among the most well understood ERP components, which is observed as a positive peak which appears after 300 ms of the presentation of an event. The appearance of P300 can be conceptually explained on the basis of the resource allocation model [114], which entails an indirect relationship between the amplitude of P300 and a direct relationship of its latency with the attentional (arousal) demand of the task. This property of P300 has been leveraged in many recent QoE studies [26, 27]. These studies have found a direct relationship between the latency and the quality of the multimedia signals along with an indirect relationship with the amplitude of P300. This suggests an increase in attentional demand for better quality signals. However, application of such techniques is mostly limited to short multimedia signals that incorporate perceivable noises at the very beginning.

Most of the real world multimedia signals consist of noises that are inconsistently located temporally. Therefore, there is a need to evaluate the EEG features that can be more practical for long-term users' states assessment. Towards this end, in this thesis, we have proposed two sets of EEG features, that are either power spectrum-based or cross-spectrum based. The power spectrum-based features measure the relative changes in the power spectral bands of EEG signals [28], in response to an event or stimulus. Such power spectrum-based features are called event-related synchronisation (ERS) or event-related desynchronisation (ERD), depending the increase or decrease in power of a frequency band, respectively. The EEG signals consist of five major subbands of interest, namely: delta (0-4 Hz), theta (4-8 Hz), alpha (8-12 Hz), beta (12-30 Hz) and gamma (30 Hz). Power spectrum based features have been found to be useful in characterising HIFs and QoE [28, 29, 30, 31].

However, the power spectrum-based features only record the brain activity in localized regions, while experiencing multimedia content different regions of the brain are expected to communicate

with each other. For example, the flow of information while watching an emotional music video clip can be represented by the model shown in Figure 1.6 [32]. The model consists of a sensory processing block and an affect-cognition processing block. The specialised sensory information is processed in a segregated manner in densely interconnected brain regions of auditory and visual cortices. The processed information is then evaluated through a highly interactive and integrated interplay of information between affective and cognitive neuronal networks [68]. Through top-down effects, these affective-cognitive networks can further influence the analysis, processing, and integration of the multiple sensory streams, thus closing the information flow loop. Within this information flow model, the affect-cognition block modulates integrated processing of information, and also influences its segregated processing through top-down effects. These properties of neuronal networks can be quantified using cross-spectrum based features, through graph-theoretical analysis [33]. For this, the cross-spectrum analysis, based on magnitude squared coherence (MSC) between various electrodes, is used to populate the so-called graph, which is then used to compute various graph metrics that quantify the dynamics of information flow.

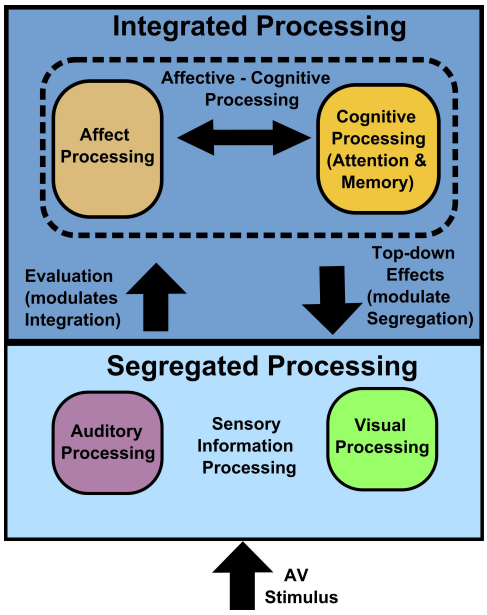


Figure 1.6 – Information flow model for affect-laden audio-video stimuli

1.2.2 Functional Near Infrared Spectroscopy (fNIRS)

The neural activity of the brain is also regulated via changes in the blood oxygenation in the immediate vicinity of an active neuronal network. Specifically with increase in activity, the

metabolic demands of a neuronal network increase, which in turn leads to an increase in oxygen consumption [115]. Therefore to meet this increased oxygen demand, vasodilation of the local arterioles bring about increase in cerebral blood flow and blood volume [116]. This leads to changes in concentrations of oxy-haemoglobin ($\Delta[HbO]$; 'Δ' represents the change in concentration w.r.t. rest) and deoxy-haemoglobin ($\Delta[HbR]$). Using fNIRS, these changes in $[HbO]$ and $[HbR]$ can be detected. This is achieved by placing a source and a detector approximately 3 cm apart from each other on the scalp surface. Then the brain surface is illuminated with a low energy infrared radiation (with generally two different wavelengths, e.g., 760 nm and 850 nm) which travels through the skin and skull into the cortex. This radiation subsequently gets reflected from the cortical surface of the brain, which is then captured using the detectors. The path taken by the radiation from a source till the detector forms a *channel* of the fNIRS instrument. The changes in $[HbO]$ and $[HbR]$ are usually manifested in the intensities of the reflected radiation. The intensity measurements are converted into $\Delta[HbO]$ and $\Delta[HbR]$ using the so-called 'Modified Beer-Lambert law' (MBLL) [34], which is used to interpret the cortical activation.

Typical $\Delta[HbO]$ and $\Delta[HbR]$ curves are depicted by Fig. 1.7. After cortical activation, it is known that oxygenation peaks between 3-6 seconds post-activation [117]. Moreover, many studies have established that $\Delta[HbO]$ and $\Delta[HbR]$ temporal dynamics are negatively correlated [118, 119]. The temporal dynamics of $\Delta[HbO]$ and $\Delta[HbR]$ have been found to be correlated with the blood oxygenation level dependent (BOLD) signal measured using magnetic resonance imaging (MRI) that is directly proportional to cortical activation. In fact, the $\Delta[HbO]$ concentration is directly proportional, whereas $\Delta[HbR]$ concentration is inversely proportional to BOLD [120]. Therefore, cortical activation is reflected by an increase in $\Delta[HbO]$ and decrease in $\Delta[HbR]$.

Data Acquisition

The fNIRS data acquisition system consists of two types of probes called optodes: sources and detectors. The fNIRS sources are typically light emitting diodes (LEDs) or lasers, that shine near infrared radiation of two different wavelengths (e.g., 760nm and 880nm) through the skin. The fNIRS detectors are built using photodiodes that capture the reflected near infrared radiation and convert it into an electrical signal. The sources and detectors are placed approximately 3 cms apart, so that some of the photons that are reflected from the cortical surface are captured by the detector

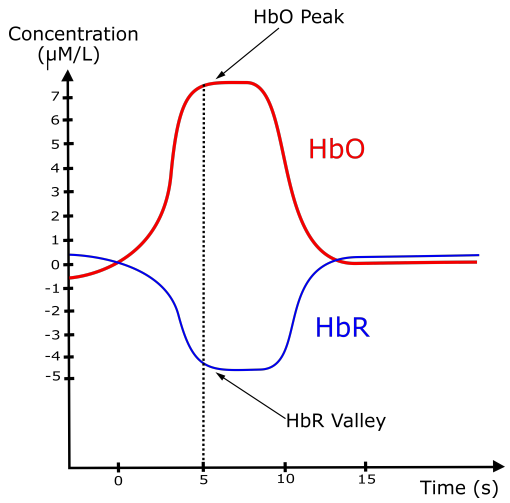


Figure 1.7 – Visual representation of typical $\Delta[HbO]$ and $\Delta[HbR]$ waveforms.

[121]. There is no fNIRS specific standard for placing the optodes in order to probe haemodynamic activity of the brain. However, most of the fNIRS systems follow the international 10-20 system for the placement of fNIRS optodes. This enables co-localized and simultaneous recordings of EEG and fNIRS. Furthermore, fNIRS signals are generally recorded at low sampling frequencies around 1 Hz, as the BOLD response manifests itself at approximately 0.1 Hz [122].

Preprocessing

The raw fNIRS signals (intensities recorded by the detectors) are often corrupted by various physiological artifacts, such as heart or cardiac pulsation (heart rate), respiration, and Mayer waves (or blood pressure waves) [35]. The noise generated from heart pulsations, respiration and Mayer waves are typically noticeable around 1 Hz, 0.5-0.08 Hz and 0.1 Hz, respectively, in the power spectrum of the raw fNIRS signal [35]. These artifacts can lead to inaccurate quantification of $\Delta[HbO]$ and $\Delta[HbR]$ temporal dynamics, thus must be removed [35]. The physiological artefacts can be easily removed using the band-pass filters. The raw fNIRS signals quantify the amount of reflected intensity of the near infrared radiation injected into the cortex. The raw intensity of the radiation is converted into measures of $\Delta[HbO]$ and $\Delta[HbR]$ using the so-called 'Modified Beer-Lambert law' (MBLL) [34].

Feature Extraction

Towards describing long-term users' states using fNIRS, various features have been investigated recently. One of the most common features is the temporal dynamics of the $\Delta[HbO]/\Delta[HbR]$ amplitude itself. For example, [36] used it to characterize the users' preference of a movie based on its likability. In [37], the authors have described the laterality (such as lateral absolute mean difference between two brain hemispheres, lateral slope ratio) and single channel based features (e.g., stimuli period mean, stimuli period slope) to characterize users' emotional states. Various other features along with their applications in the user experience (UX) domain (which is closely related to QoE [123]) have been studied and are listed in [124]. However, fNIRS as a technique is in its infancy in the QoE assessment domain and could prove to be very effective alongside EEG, in developing better objective models to characterize QoE. Towards this end, in this doctoral thesis, we have explored a set of fNIRS-based features, such as mean, variance, kurtosis and skewness of $\Delta[HbO]$ and $\Delta[HbR]$ as QoE correlates.

1.2.3 Peripheral Autonomous Nervous System (PANS)

In addition to recording brain activity for users' state characterization, it is also useful to record certain peripheral physiological signals using techniques such as plethysmography, galvanic skin response (GSR), skin temperature, respiration activity measurement and eye-tracking. These measurements are the manifestations of PANS, which helps the central nervous system communicate with the rest of the body. The GSR is the direct measurement of the electrical conductance of the skin induced by the changes in the activity of the sweat glands due to an event. The features, such as amplitude changes, frequency of electrodermal activation and skin conductance variability derived from GSR, have been used to describe users' states [28, 38, 39, 40]. Moreover, plethysmogram derived features such as heart rate, heart rate variability (HRV) etc., have been used as indicators of various emotional responses [41, 42]. In a similar vein, quantification of respiratory features, such as inhalation/exhalation amplitude and duration, have been used to characterize emotion states [125]. In addition, eye-tracking has proven to be a very useful tool in the UX domain [126]. Thus, it could be hypothesized that a multimodal framework of all the neuro-physiological response measurement techniques could help develop more accurate objective QoE assessment models. Previous studies, however, have relied on dedicated hardware to collect physiological data, such as plethysmography

for heart rate monitoring and strain gauge-based respiration belts. Here, an alternate route is attempted and we propose to extract heart rate information by processing the raw fNIRS signal. This will allow for a richer pool of multimodal data to characterize HIFs that affect perceived QoE.

1.3 Thesis Contributions

The aim of this thesis is to develop BCI measures, based on the fusion of different neurophysiological modalities, to characterize human factors that influence QoE perception, and to incorporate such hybrid BCIs into state-of-the-art objective QoE models. For the evaluation of the developed measures, a scenario based on text-to-speech (TTS) systems was explored, as over the last few years, 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, reading assistants to the blind, and global positioning systems. Here, we have listed the key contributions of this thesis:

1. The development of an open-source multimodal neurophysiological database, the so-called PhySyQX database, for the characterisation of HIFs. The database uses synthesised and natural speech, used for the development of personal digital assistants, as stimuli. The subjective data from the PhySyQX database was used to establish the importance of affective human factors as indicators of underlying perceptual constructs of QoE. Publications that have resulted from this contribution include [43, 44].
2. The proposal of two classes of features based on EEG for the long-term characterisation of HIFs. The first is the power spectrum-based features, such as event related (de)synchronization (ERD) and asymmetry index (AI). The ERD-based approach captures the regional cortical activities, whereas the AI-based approach captures the interhemispheric differences in cortical activity while experiencing multimedia signals. These approaches helped localise the brain regions responsible for processing QoE-related information, which can be useful for future studies. The second approach uses cross-spectrum analysis, to decode the interplay of information flow between different regions of the brain, using graph-theory principles. For validation, the proposed features were tested on two separate databases, namely, the PhySyQX and the DEAP database. Publications that have resulted from this contribution include [45, 32, 46].

3. The proposal of fNIRS-based features for the long-term characterisation of HIFs. The fNIRS based features leverage the temporal dynamics of $\Delta[HbO]$ and $\Delta[HbR]$ for the characterization of HIFs. Moreover, the so-called physiological noise of heart pulsation in the raw fNIRS data was extracted and used to develop models for monitoring HIFs. Furthermore, the features from EEG, fNIRS and fNIRS-based heart rate were used to develop hybrid BCI for the same purpose. The features were formulated using the database collected during a preliminary study, which probed only the prefrontal cortex (forehead). The formulated features were then validated using the PhySyQX database. Unfortunately, the DEAP database does not consist of fNIRS data and therefore, it was not used to test the fNIRS-based features. Publications that have resulted from this contribution include [47, 46].
4. The first steps towards incorporating hybrid BCI-based tools into state state-of-the-art objective QoE models are taken. The hybrid BCI-based measures that characterize human affective states were incorporated into an objective QoE model for synthesised speech. Publications that have resulted from this contribution include [48, 49].

1.4 Thesis Organization

This doctoral thesis summarises the work of the author on hybrid BCI-based characterisation of human factors that influence the perception of quality-of-experience of complex multimedia signals. Some of the results of this thesis have been published or are under review in several conference proceedings and international journals [48, 49, 47, 46, 45, 32, 43, 44, 127]. Along with summarising the previously published aspects of its results, this work presents an overview and a restructured discourse concerning how different neuroimaging tools, which form a hybrid BCI, complement each other while characterising the subjective human factors, such as users' affective states that influence QoE perception. Furthermore, the structure of this thesis outlines the process of integration of the hybrid BCIs, which characterise HIFs, into state-of-the-art objective QoE assessment methods.

Chapter 2 outlines the methodology for simultaneous subjective and multimodal neurophysiological data collection for developing hybrid BCIs that characterise the HIFs. The subjective QoE assessment tests gathered the so-called 'ground-truth' information regarding the HIFs, based on which the hybrid BCIs were developed. Moreover, this chapter provides an in-depth analysis of the subjective data that explores various attitudinal and affective human factors that can potentially

influence the perceived QoE. Chapter 3 and 4 focus on developing EEG and fNIRS-based features, respectively, that are useful for long-term monitoring of HIFs using BCIs. Next, the outputs of individual BCIs, developed using EEG and fNIRS, were fused to develop a hybrid BCI in Chapter 5. In Chapter 6, the BCI-based measures developed in previous chapters were incorporated into state-of-the-art objective QoE assessment model. Lastly, Chapter 7 provides a general discussion and conclusions.

Chapter 2

Physiological Database Development

2.1 Preamble

This chapter is compiled from material extracted from a manuscript published in the Proceedings of the 2015 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics [43] and a manuscript that is under review in the journal for Quality and User Experience [44].

2.2 Introduction

Subjective tests are considered to be the 'go-to' techniques for assessing multimedia QoE [109]. The International Telecommunication Union (ITU) recommends multidimensional subjective listening tests for the evaluation of QoE for different multimedia signals, such as P.800 [86], for speech quality, P.910 [128], for video quality, P.911 [129], for audiovisual quality and P.85 [9], for text-to-speech (TTS) systems. However, these subjective listening tests are time consuming and expensive. Therefore, researchers have been making a concerted effort in developing objective or instrumental models for estimating the QoE, such as for synthesised speech; representative examples of objective models include: P.563 [14], ANIQUE+ [130] and HMMs [131]. The objective models are computer algorithms that try to replicate users' subjective scores; thus subjective assessment provides the so-called 'ground-truth' information for developing the objective models. Towards developing objective QoE models, generally, the technological and contextual factors are considered. However,

human factors, such as users' affect and perceived listening effort, are considered to be very important factors in driving users' perception of QoE [4]. Thus, HIFs need to be accounted for while developing any objective model for QoE evaluation.

The HIFs, however, are subjective in nature and are not directly observable. This is because most of the quality formation and judgment processes take place inside the user's brain [132]. As such, any change in subjective quality is expected to be manifested inside user's brain both neuronally (via, electrical impulses) and haemodynamically (via blood flow). Therefore, probing the brain activity, using BCI-based tools, is expected to provide better understanding of the human quality judgment processes [133]. However, gathering neurophysiological insights, using BCI-based tools, demands rigorous data collection procedures. Moreover, neurophysiological data is prone to various physiological and instrumental noise sources, thus requiring the implementation of complex data cleaning and analysis techniques. Therefore, owing to their time consuming and expensive nature, and the fact that they need careful expert supervision, the adoption of BCI-based QoE assessment methodologies has been slow. Availability of open-access databases, based on data collected from BCI-based QoE assessment methodologies, can help mitigate the apprehensions of the research community related to the use of such new techniques. Such databases allow easy access of BCI-based QoE assessment data to a larger section of the research community that is not experienced with neurophysiological data collection. Furthermore, open-access nature of the database allows replication of the study, from which the data was collected, thus, providing further validity to the results.

Towards this end, so called database for emotion analysis using physiological signals (or DEAP) [28] was developed at EPFL, Switzerland; it consisted of using EEG to characterize HIFs, such as valence and arousal, while watching music videos. In the same vein, the PhySyQX database reported herein, takes a BCI-based approach to explore the effects of HIFs on synthesized speech quality formation processes using EEG and fNIRS. Previous research using EEG [134] and fNIRS [47] to characterize TTS systems' QoE has proven very useful. However, the PhySyQX database takes one step further and provides access to multimodal (fNIRS-EEG) signals collected simultaneously and covering the entire scalp region, along with the related audio files and subjective HIF ratings. It is hoped that the database will complement traditionally open-access TTS speech corpora, such as those from the Blizzard Challenges [50] (which only provide audio files along with subjective scores for attitudinal human factors). This allows researchers to correlate neural insights with the obtained

subjective ratings and develop hybrid BCI-based objective QoE assessment methods. Another key difference between the PhySyQX database and other traditional open-access TTS speech databases is that it encompasses a wider variety of subjective dimensions, which include both attitudinal and affective human factors, therefore allowing us to establish the effects of users' affective states on QoE perception. To the best of the author's knowledge, this is the first such open-access multimodal database of its kind for TTS stimuli.

The remainder of this chapter is organised as follows: Section 2.2 describes the methods and materials used to develop the database. It should be noted that Sections 2.3 and 2.4 only provide the results and discussion for subjective data, respectively, whereas the remaining chapters of this thesis extensively utilize the neurophysiological data collected for PhySyQX database, for the development of hybrid BCI-based objective QoE assessment methods. Lastly, conclusions are drawn in section 2.5.

2.3 Materials and Methods

The data was collected over two sessions. Data from the first session (pilot) was used for the exploratory phase of factor analysis, and the data from the second (main) session was used for the confirmatory phase of factor analysis. The data from the second session has been made publicly available [43] in the form of the so-called PhySyQX database.

2.3.1 Participants

A total of twenty-eight participants were recruited for the study, six of whom participated in session one (pilot) and 21 in session two (main). All participants were fluent in English. For session one, two were female and the participant average age was 31.16 ($\pm$8.18). For session two, (8 females), the average age was 23.8 ($\pm$4.35). None of the participants reported having any hearing or neuro-physiological disorders. The study protocol was approved by the INRS Research Ethics Office and participants consented to participate in the studies and make their de-identified data available freely online. The participants were also compensated monetarily for their time.

Table 2.1 – Description of the stimuli used for the listening test.

Type	System	Sentence Group	Male Sets	Female Sets	Length	Event Markers
Natural	1	A	0	4	17-19s	5-8
	2	A	0	4	18-23s	9-12
	3	A	0	4	17-19s	13-16
	4	B	0	4	13-14s	17-20
Synthesised	5	A	0	4	19-24s	21-24
	6	A	0	4	17-22s	25-28
	7	A	2	2	17-20s	29-32
	8	A	2	2	18-25s	33-36
	9	A	2	2	17-22s	37-40
	10	A	2	2	17-21s	41-44
	11	A	2	2	13-17s	45-48

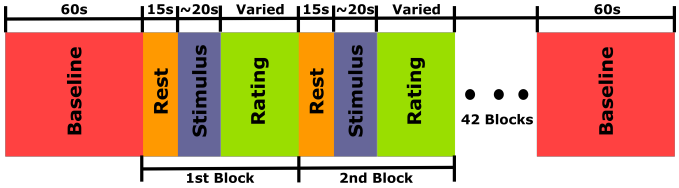


Figure 2.1 – Visual representation of the protocol used in the experimental phase.

2.3.2 Speech Stimuli

Table 2.1 lists the speech stimuli used for this study along with certain important aspects. The stimuli consisted of four natural voices and seven synthesised voices obtained from commercially available systems, namely Microsoft, Apple, Mary TTS Unit selection & HMM, vozMe, Google and Samsung. Tested systems cover a range of different concatenative and hidden markov model-based systems. A non-identifying code is provided for each the 7 TTS systems in Table 2.1. Speech samples were generated from two sentence groups (A & B), listed in Appendix-A, each comprising of 4 sentences. Thus, the total number of stimuli used in this study were forty-four (Natural voices: 4 + Synthesised voices: 7 = 11 voices, 4 sets of sentences = 44 stimuli). The speech stimuli also consisted of both male and female voiced sets of sentences for five of the seven synthesised voices. The speech stimuli were presented to listeners at a sampling rate of 16 kHz and a bitrate of 256 kbps. Table 2.1 also details the length of speech stimuli for each system along with the event markers that were used to synchronize data acquisition over the different neurophysiological modalities.

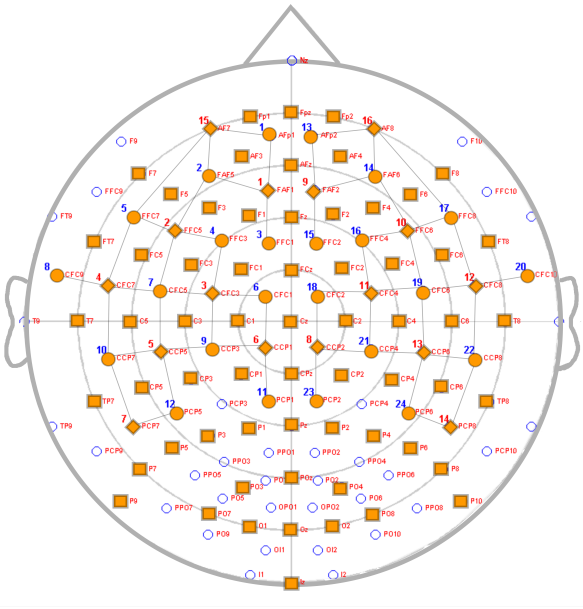


Figure 2.2 – The figure shows the topology for fNIRS optodes and EEG electrodes along the cap. The EEG electrodes, fNIRS detectors and sources are represented by rectangles, circles and diamonds, respectively. The fNIRS channels are shown using straight lines connecting the sources and detectors.

2.3.3 Experimental Protocol

The experimental procedure was carried out in accordance with ITU-T P.85 recommendations [9]. Participants were first comfortably seated in front of the computer screen inside a sound-proof room. Participants were then fitted with EEG electrodes and fNIRS sensors using a compatible fNIRS/EEG cap. Insert earphones were placed comfortably inside the participants' ears to deliver the speech stimuli. The experiment was then carried out in two phases: a familiarity phase and an experimental phase. In the familiarity phase, participants were presented with a sample speech file followed by the series of rating questions, thus illustrating the experimental procedure and giving them the opportunity to report any problem and/or concerns. Next, the experimental phase consisted of several steps as shown in Fig. 2.1. First a baseline (or resting period data) was collected for 1 minute (trigger = 4) for which the participants were advised to focus only on the crossbar in the middle of the screen and not think about anything else. They were then presented with the randomised speech stimuli, one sentence set (approximately 20 s long) at a time. Before each speech stimulus a 15-second rest period was provided to allow neural activity (particularly blood flow) to return to baseline levels. Moreover, following each stimulus participants were presented with a randomised series of rating questions on the screen wherein the participants scored the stimuli on continuous valued scales for quality and various other human influence factors affecting

the perceived quality of user experience. Following the presentation of the 44 speech stimuli, a second baseline or resting period was presented for 1 minute (trigger = 4).

Table 2.2 shows the 12 subjective rating scales used. Most of the subjective dimensions were in accordance with P.85 recommendations. However, additional dimensions of valence, arousal and dominance were also introduced, which were scored using self-assessment manikins (SAM) [51]. The description for each subjective dimension is provided below:

1. **Overall Impression:** This scale evaluated the overall quality of the system considering all the aspects.
2. **Voice Pleasantness:** This measured the degree of voice pleasantness.
3. **Speaking Rate:** This measure reflected the listener's reaction to the speed of delivery in a real situation.
4. **Acceptance:** This scale measured whether the voice could be accepted as a Personal Digital Assistant or not.
5. **Intonation:** This scale gauged whether the produced pitch curve fits to the sentence type.
6. **Naturalness:** This scale measured the level of naturalness/unnaturalness of the voice.
7. **Listening Effort:** This captured the effort required to listen to a particular voice while listening to it for a longer duration of time.
8. **Comprehension Problems:** This scale measured the comprehension problems that might have arisen due to badly synthesized speech.
9. **Emotion:** This item captured the variations of voice that reflected the emotion of the scene being described.
10. **Valence:** This item captured the attractiveness (positiveness) or averseness (negativeness), of the voice, as experienced by the listener.
11. **Arousal:** This item measured the level of mental alertness/excitation of the listener after listening to the voice.
12. **Dominance:** This item measured the feeling of control over the situation after listening to the voice.

Table 2.2 – Subjective dimensions used in the listening test along with their description and abbreviations used herein.

Dimensions	Abbreviation	Description
Overall Impression	MOS	1-Bad,... 5-Excellent
Voice Pleasantness	VP	1-Very Unpleasant,... 5-Very Pleasant
Speaking Rate	SR	1-Slow,... 5-Fast
Acceptance	Ac	1-Strongly don't accept,... 5-Strongly accept
Intonation	Int	1-Melody did not fit sentence type,... 5-Melody fitted the sentence type
Naturalness	Nat	1-Unnatural,... 5-Natural
Listening Effort	LE	1-Very Exhausting,... 5-Very Easy
Comprehension Problems	CP	1-Never,... 5-All the time
Emotions	Emo	1-No expression of emotions,... 5-Authentic expression of emotions
Valence	Val	1-Negative,... 9-Positive
Arousal	Ar	1-Unexcited,... 9-Excited
Dominance	Dom	1-Not in control,... 9-In control

2.3.4 Multimodal Data Acquisition

The subjective data collection, stimulus presentation and data synchronization over all devices was carried out using Presentation software (Neural Behavioral Systems, USA). EEG data was acquired using the Biosemi ActiveTwo system at a sampling rate of 512 Hz, with no online filtering. The cap used for the study consisted of holders placed according to the modified 10/20 system of electrode placement, as shown in Fig. 2.2. The 62 EEG electrodes were then placed over the scalp accordingly (AF7 and AF8 were not used). Furthermore, reference electrodes were placed over the two mastoids, and nose and oculography electrodes were placed to record the horizontal and vertical eye movements.

The fNIRS data, in turn, were recorded using the NIRx NIRScout system with 16 bi-wavelength sources (probed wavelengths were 760 nm and 850 nm) and 24 detectors. The fNIRS optodes were placed alongside the EEG electrodes as shown in Fig. 2.2. Each source-detector pair with a distance below 3 cm formed an fNIRS channel. This resulted in approximately 60 fNIRS channels. The sampling rate used to record the data was 4.46 Hz. The recordings were made using the NIRStar version 13.0 software provided by NIRx. For data collection, optode amplification gains were set for each channel directly proportional to their source-detector distance. In general, the

gains provide insights regarding the signal-to-noise ratio (SNR) for a particular channel, thus higher gains indicated lower SNRs. The NIRStar system automatically adjusted channel gains to result in SNRs above an empirically determined acceptable level.

2.3.5 Subjective Data Analysis

Exploratory Data Analysis

In the exploratory phase of subjective data analysis, the first step is to establish inter-rater agreement and reliability of the subjective dimensions using intra-class correlation. Furthermore, a sub-sampling analysis can be implemented where a large number of random samples are taken from the complete dataset for each subjective dimension. For each sub-sample, the correlation between the average of the subjective dimensions with the average of those subjective ratings for the complete dataset is computed. In particular, for the PhySyQX dataset 100 random samples from the full dataset, at sub-sample sizes ranging from 1-20 in increments of 1, were computed following [52].

The next step involves exploring systemwise boxplots for each subjective dimension. The boxplots help in determining the median values of the subjective ratings for each system along with the spread of the data around the median value, thus visually representing the variability in the data. Additionally, towards establishing significant differences between different TTS systems, along the subjective dimensions, is to compute ANOVA with subjective scores as dependent variable and TTS system as independent variable. In general, the ANOVA is followed by post-hoc tests to determine significant differences, in subjective scores, between between TTS systems.

Factor Analysis

In order to validate the introduction of affective dimensions (e.g., valence), along with subjective dimensions from P.85 recommendations [9], for measuring subjective measurement of QoE, we conducted factor analysis. The factor analysis extracts the unobservable and internal perceptual constructs that account for a participant's scores on the subjective dimensions. The basic principle of factor analysis is that the perceptual constructs influence the subjective dimensions in a systematic manner, thus measurements obtained from subjective rating scales are, at least in part, the result

of the linear influence of the underlying latent factors [90]. The influence of the internal perceptual constructs on the subjective dimensions is quantified using 'factor loadings' [135]. There are two discrete categories of factor analysis techniques: exploratory factor analysis (EFA) and confirmatory factor analysis (CFA). The EFA estimates unrestricted measurement models, whereas CFA analyzes restricted measurement models [136]. Thus, for CFA the correspondence between rating scales and perceptual constructs needs to be specified, whereas, for EFA there are no specific expectations regarding number or nature of underlying perceptual constructs.

Previous research based on factor analysis has leveraged attitudinal HIFs, such as voice pleasantness and acceptance scales for evaluating QoE of TTS systems [91, 88, 53, 89], to extract the perceptual constructs of QoE. Thus, validating the importance of attitudinal HIFs in evaluating QoE. However, in this thesis we hypothesize that the affective human factors are equally important HIFs that must be measured to evaluate QoE. To validate this hypothesis, we conducted a similar factor analysis, based on EFA and CFA, using both attitudinal and affective HIFs. Therefore, using the resulting structure of the perceptual constructs we can determine the validity of the hypothesis. Fundamentally, significant loadings of affective human factors on any of the latent perceptual constructs will validate the importance of measuring users' affective states, for in-depth understanding of QoE related perceptual processes.

For conducting EFA, first, sampling adequacy of the data is measured using the so-called Kaiser-Meyer-Olkin (KMO) measure [137] and Bartlett's test of sphericity [138]. Another recommendation towards establishing sample size adequacy is based on the sample-to-variable ratio, denoted as *N:p* where N refers to the sample size and p refers to number of indicators. The rules of thumb for N:p values have ranged from 3:1 to 20:1 in the literature (e.g., see [139]). In the current EFA study, that used the subjective data from the pilot study, the sample size (N) was 264, as 6 subjects scored 44 speech stimuli, and the number of indicators (p) used were 11, thus leading to a *N:p* ratio of 24:1. The KMO measure and Bartlett's test of sphericity are also used herein to establish sample adequacy. Next, the number of latent constructs (or factor) to be retained during factor analysis is determined using Kaiser's criterion, which recommends to retain all the latent factors that have eigenvalues greater than one, as this is the average size of eigenvalues in the full decomposition [140]. Finally, the rotation method is chosen, as it helps to produce simplified and interpretable results by maximizing high factor loadings and minimizing low factor loadings. Here, we chose oblique rotation using the promax method, thus leading to the production of correlated construct

structures [139, 141]. Moreover, towards establishing a more reliable factor structure, EFA was also performed on random subsamples of data extending from $N = 165$ to $N = 264$ with increments of 2. This exploratory analysis allowed us to vary the sample-to-variable ratios from 15:1 to 24:1, thus further validating the data sufficiency hypothesis.

Finally, after obtaining the factor structure using EFA, CFA can be conducted using a variety of statistical packages available for implementing CFA, such as MPlus [142], AMOS[143], and lavaan [144]. For the current study, we have implemented CFA using the lavaan (Latent Variable Analysis) package for R. The lavaan package allows the specification of the CFA model (as implemented in the path diagram) through the model syntax. The model syntax is a description of the model that needs to be estimated. The lavaan package allows estimates of various goodness-of-fit (GOF) measures, that reflect the acceptability for the developed model. Several GOF indices have been proposed previously, such as the comparative fit index (CFI), normed fit index (NFI), non-normed fit index (NNFI), incremental fit index (IFI), relative non-centrality index (RNI), goodness-of-fit index (GFI), and standardized root mean square residual (SRMR) [145, 146]. The CFI, NFI, NNFI, IFI and RNI indices compare the performance of the model with a baseline (or null) model that assumes zero correlation between all the indicators. The GFI, on the other hand, does not compare the model to a baseline model and is computed based on the amount of variance explained by the model. Finally, the SRMR index is estimated by computing the mean absolute value of the covariance of residuals. Typically, values ≥ 0.90 are considered adequate for the CFI, NFI, NNFI, IFI, RNI and GFI indices [147, 148], whereas a value of SRMR ≤ 0.08 [149] reflects the adequate fit of a model. Here, a combination of these indices is used for model validation.

2.4 Results

2.4.1 Exploratory Subjective Data Analysis

Figure 2.3 shows the box-plots for all the subjective ratings acquired during the study. These box-plots show the distribution of each of the subjective ratings across the different TTS systems. In Table 2.3, the F-statistic values from one-way analysis of variance (ANOVA) are reported. Note that the p-values and the degrees of freedom for all the subjective dimensions were below 0.01 and 10, respectively. Also, the post-hoc tests showed significant differences between the natural voices

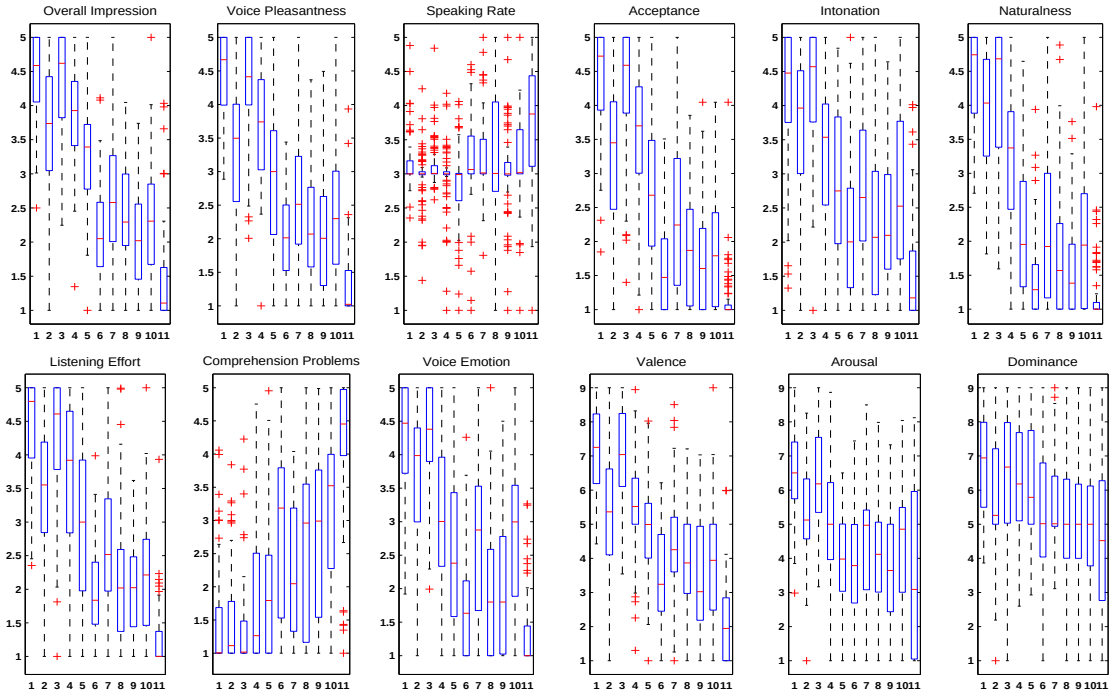


Figure 2.3 – The figure shows the box-plots for the subjective ratings along the TTS systems (labelled 1-11) on the x-axis. The median for each dimension is shown in red line in the box whereas, the outliers are shown using red '+' symbol.

and TTS systems consistently for all the subjective dimensions, except dominance and speaking rate. Moreover, to quantify the inter-rater reliability of the subjective dimensions, Table 2.3 also lists the intra-class correlation coefficients (ICC). Furthermore, the resulting plot for sub-sampling analysis, for each subjective dimension, is shown in Figure 2.4. The figure, only displays the sub-sampling plots for the subjective dimensions that achieved the correlation values of 0.95, which was empirically set. Also, the number of subjects (NOS) required to reach an agreement for a particular subjective dimension has been reported in Table 2.3, by keeping a threshold of 0.95 on the sub-sampling analysis. This indicates the minimum number of subjects required to characterize a subjective dimension with confidence. Also in Table 2.3, we have reported the Pearson cross-correlation coefficient matrix for all 12 subjective dimensions.

Table 2.3 – List of Intra-Class Correlation (ICC), number of subjects (NOS) required to achieve 0.95 confidence in sub-sampling analysis and the ANOVA F-statistic along with the Pearson correlation coefficient matrix for the 12 subjective dimensions.

Dimension	ICC	NOS	F-stat	Pearson Correlation Coefficient											
				MOS	VP	SR	Ac	Int	Nat	LE	CP	Emo	Val	Ar	Dom
MOS	0.63	5	143.3	1	0.85	-0.24	0.84	0.72	0.79	0.83	-0.64	0.71	0.81	0.51	0.27
VP	0.60	5	124.7	0.85	1	-0.23	0.89	0.73	0.80	0.88	-0.58	0.72	0.87	0.57	0.31
SR	0.28	16	16.1	-0.24	-0.23	1	-0.23	-0.18	-0.20	-0.24	0.31	-0.20	-0.21	0.05	-0.12
Ac	0.61	5	132.5	0.84	0.89	-0.23	1	0.72	0.78	0.87	-0.56	0.72	0.84	0.55	0.29
Int	0.44	9	64.3	0.72	0.73	-0.18	0.72	1	0.74	0.71	-0.45	0.80	0.70	0.53	0.23
Nat	0.63	4	144.7	0.79	0.80	-0.20	0.78	0.74	1	0.79	-0.53	0.79	0.73	0.54	0.23
LE	0.57	5	109.3	0.83	0.88	-0.24	0.87	0.71	0.79	1	-0.61	0.69	0.83	0.54	0.27
CP	0.40	8	60.9	-0.64	-0.58	0.31	-0.56	-0.45	-0.53	-0.61	1	-0.41	-0.55	-0.24	-0.33
Emo	0.51	6	87.1	0.71	0.72	-0.20	0.72	0.80	0.79	0.69	-0.41	1	0.70	0.52	0.22
Val	0.52	6	96.3	0.81	0.87	-0.21	0.84	0.70	0.73	0.83	-0.55	0.70	1	0.61	0.33
Ar	0.25	18	31.5	0.51	0.57	0.05	0.55	0.53	0.54	0.54	-0.24	0.52	0.61	1	0.14
Dom	0.09	>20	10.7	0.27	0.31	-0.12	0.29	0.23	0.23	0.27	-0.33	0.22	0.33	0.14	1

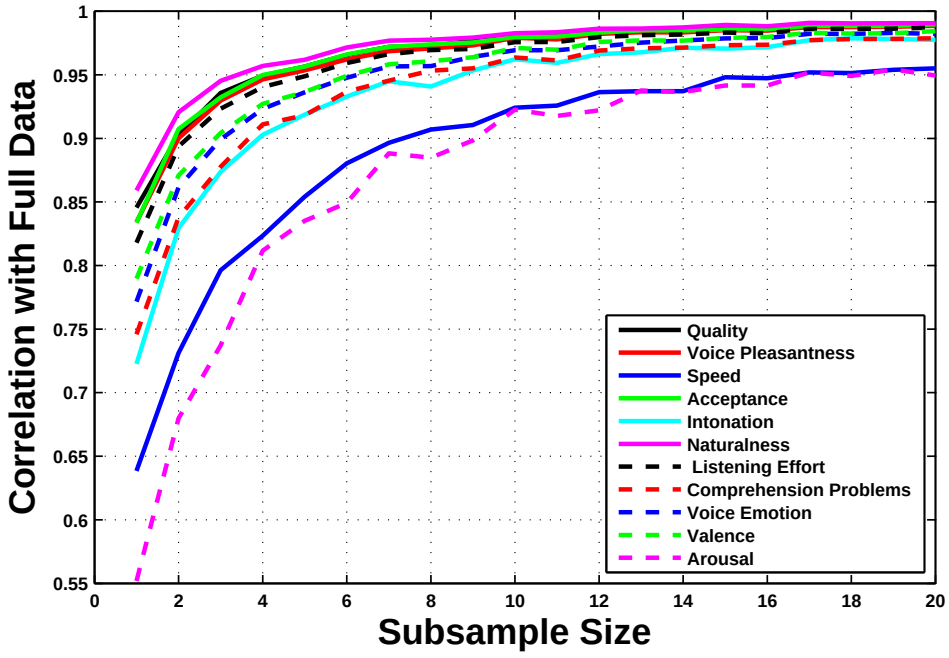


Figure 2.4 – Subsampling analysis over the rating scales.

2.4.2 Factor analysis

Exploratory Factor Analysis

The Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy along with Bartlett's test of sphericity were computed to assess the adequacy correlation matrix for factor analysis. The measured KMO for the data was 0.94, thus supporting factor analysis, as KMO values above 0.5 are typically considered to be adequate for EFA [150]. The Bartlett's test of sphericity resulted in significance levels below 0.05, thus confirming significant relationships between ratings [151]. These measures established the adequacy of the data for exploratory factor analysis. As a next step, the number of factors necessary for EFA was obtained using Kaiser's criterion, which recommended that two factors be retained (Factor 1: eigenvalue = 7.251; Factor 2: eigenvalue = 1.05).

Next, EFA was performed using promax rotation to reduce cross loadings of items. This resulted in two factors, where the ratings *voice pleasantness*, *acceptance*, *listening effort*, *comprehension problems* and *valence* loaded on most significantly to Factor 1. Ratings *intonation*, *emotion*, *naturalness* and *arousal*, in turn, loaded on to Factor 2, as shown in Table 2.4. To obtain meaningful factors, factor loadings below 0.5 were not considered. Therefore, dominance and speaking rate did not load significantly on any of the factors. Also, as visible from Table 2.4, the sub-sampling factor analysis, employed to validate the reliability of the factor structure and data sufficiency, produced similar factor structure and mean loadings along with very low standard deviations over the obtained factor loadings. Moreover, the cumulative variance explained by the two factors was 57%.

Confirmatory Factor Analysis

Towards verifying the factor structure obtained from EFA, a confirmatory factor analysis or CFA was performed. The model fit parameters obtained from CFA, as reported in Table 2.5, validate the model as the fit parameters GFI, NFI, NNFI, CFI, RNI and IFI were observed to be greater than 0.90 and SRMR was found to be less than 0.08. Following the CFA, measurement invariance (MI) and structural invariance (SI) for the model were examined for different groups in the data. First, invariance tests were performed between samples from groups of *female and male* participants, followed by samples from groups of *natural and synthesized* speech stimuli.

Table 2.4 – Factor loadings obtained for each item using EFA.

Rating	General EFA Factor Loadings		Subsampling EFA Mean		Subsampling EFA Std. Dev.	
	1	2	1	2	1	2
VP	**0.85**	0.14	**0.847**	0.140	**0.008**	0.009
Ac	**0.80**	0.15	**0.798**	0.153	**0.007**	0.008
LE	**0.91**	0.03	**0.899**	0.043	**0.010**	0.012
CP	**-0.73**	0.13	**-0.723**	0.116	**0.012**	0.013
Val	**0.84**	0.09	**0.840**	0.092	**0.008**	0.009
Int	0.17	**0.74**	0.170	**0.742**	0.012	**0.012**
Nat	0.45	**0.52**	0.449	**0.520**	0.006	**0.006**
Emo	-0.12	**1.02**	-0.109	**1.013**	0.011	**0.012**
Ar	0.27	**0.52**	0.274	**0.513**	0.012	**0.011**
SR	-0.20	-0.10	-0.198	-0.102	0.012	0.013
Dom	0.34	0.33	0.335	0.335	0.012	0.013

Table 2.5 – Goodness-of-fit metrics obtained using CFA.

GFI	NFI	NNFI	SRMR	CFI	RNI	IFI
0.941	0.968	0.961	0.031	0.971	0.971	0.972

2.5 Discussion

The ANOVA results reported in Table 2.3 and the box-plots shown in Figure 2.3 suggest that there is significant difference between the TTS systems across all subjective dimensions. However, certain dimensions, such as Speaking rate, Dominance and Arousal showed low inter-rater reliability, hence required more subjects to produce reliable inferences. This could be due to the high variability in speaking rate preferences and in the experienced affect between subjects; moreover, it could be due to misinterpretation of these subjective dimensions. Also, the correlation matrix reported in Table 2.3 shows significant correlations between perceived QoE and several HIFs.

EFA was performed using all the subjective dimensions except overall impression, as it comprises information from other dimensions [53]. The EFA resulted in extraction of two factors, with *factor 1* with loadings from voice pleasantness, acceptance, listening effort, comprehension problems and valence, and *factor 2* with loadings from intonation, emotion, naturalness and arousal. Thus, it is evident that the items that load on *factor 1* cover the *listening pleasure* and intelligibility of the systems, whereas items that load on *factor 2* reflect the signal *prosody* and rhythm. Moreover, the

sub-sampling EFA validated the obtained factor structure as it resulted in similar factor loadings with low variations (given by standard deviation) for each indicator. This establishes 'listening pleasure' and 'prosody' as the two perceptual dimensions of synthetic speech QoE. Furthermore, the findings are in corroboration with exploratory factor analysis performed for audiobooks, as reported in [53].

Interestingly, the valence and arousal scales loaded on two different factors, factor 1 and 2, respectively. The valence and arousal scales form the two orthogonal dimensions of the emotional/affective experience corresponding to positiveness/pleasantness and alertness [54], respectively. Thus, the loading of the valence item on factor 1 further establishes the relationship of factor 1 to the perceptual dimension of 'listening pleasure'. Also, the loading of arousal scale on factor 2 relates stimulus-evoked alertness to prosody in speech, which is also corroborated by previous findings reported in [55]. These findings indicate that changes in underlying perceptual constructs of QoE, due to changes in system quality, alter users' affective states. Therefore, it is evident that affective scales corresponding to valence and arousal dimensions are important for estimating underlying perceptual dimensions of users' experience with personal digital assistants.

However, another model for users' affect utilizes an additional dimension of the so-called dominance for describing users' control over a situation [11]. The dominance scale did not load significantly on any of the two extracted factors. Similarly, the speaking rate scale did not show any significant loadings on either of the two factors. The insignificant factor loadings for dominance and speaking rate can be attributed to their low F-statistic values obtained using ANOVA, as reported in Table 2.3, thus suggesting low inter-class variation compared to intra-class variation for these scales. Therefore, both dominance and speaking rate scales were rejected from further analysis.

The EFA established the factor model for confirmatory factor analysis, as shown in Fig. 2.5. The factor model consists of two correlated factors as these were estimated using a promax rotation that results in oblique (non-orthogonal) factors. The first and second factors can be measured using five and four continuous factor indicators, respectively. The loadings from factor indicators and previous research [53] suggest the first and second factors represent listening pleasure and prosody, respectively. Finally, the factor model for evaluating the perceptual dimensions of the synthesized speech QoE was confirmed using the confirmatory factor analysis as all the model fit parameters satisfied the goodness-of-fit criteria.

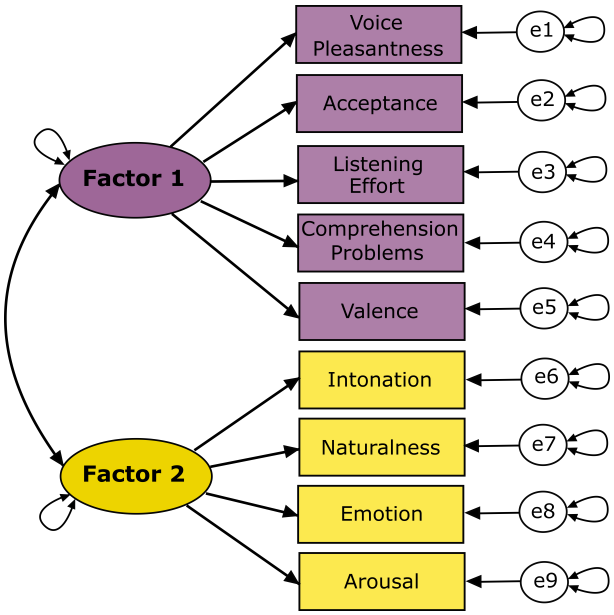


Figure 2.5 – The factor model for confirmatory factor analysis.

The pre-processing and feature extraction steps for physiological signals are described in the following chapters. Furthermore, the following chapters describe the development of techniques for objective characterization of HIFs using neurophysiological tools.

2.6 Conclusion

The described PhySyQX database has been made available online and can be downloaded from: http://musaelab.ca/resources/. The subjective data from PhySyQX database has been extensively analysed and evaluated in order to establish the validity of the ground-truth information that will be utilized in the following chapters to develop BCI-based models. Moreover, factor analysis results established the importance of quantifying affect related human factors in understanding QoE perception. Furthermore, towards developing BCI-based objective assessment models, the neurophysiological data, described herein, will be extensively leveraged in the following chapters.

Chapter 3

Characterization of HIFs using EEG-based BCIs

3.1 Preamble

This chapter is compiled from material extracted from manuscripts published in the Proceedings of 2015 7th International IEEE/EMBS Conference on Neural Engineering [32] and Neurocomputing [45] ; and a manuscript that is under review in the IEEE Journal for Selected Topics in Signal Processing: Special Issue on Measuring Quality of Experience for Advanced Media Technologies and Services [46].

3.2 Introduction

EEG-based QoE assessment is the most common BCI-based approach for QoE assessment [84, 27, 85], as EEG is a well established neuroimaging technique that provides neurophysiological insights at a very low cost. The most common EEG-based feature used for neurophysiological evaluation of HIFs is the so-called 'event-related potential' (ERP). An ERP is a time- and phase-locked change in the voltage amplitude of the EEG time series due to changes in neuronal populations, in response to a sensory, cognitive or motor event [25, 56]. The so-called 'P300' is among the most well understood ERP components [114] and has been leveraged in recent QoE studies [84, 27, 85].

This methodology can be used to effectively understand and characterize the unconscious processes involved in the QoE perception of *short* multimedia signals [85]. Moreover, ERP-based neurophysiological assessment assumes that the signal characteristics that affect the perceived QoE are time-locked to the beginning of the multimedia signal. This may be a valid assumption when testing, e.g., the effects of ambient noise [84], as ERPs can capture momentary changes in HIFs, such as users' attention, that are induced due to changes in stimulus quality. However, it is not the case with long duration multimedia signals, such as synthesised speech or affective music videos, which comprise time-varying changes in HIFs. Also, certain late cognitive neural activities related to downstream processing of a stimulus [152], however, have been shown to be time-locked to the event, but not phase-locked [56, 153, 154]. Such non-phase locked neural activities tend to get suppressed with ERP-based analysis of the EEG signal [153]. To overcome this limitation, we propose the use of a feature, called event-related (de)synchronisation (ERD/ERS), that measures the relative *changes* in spectral power in EEG sub-bands [56], for the characterization of various human factors that influence the perceived QoE of multimedia signals. As such, these features have proven useful for the evaluation of HIFs, such as users' fatigue caused by long-duration multimedia signals [30, 31]. It is hypothesized that such measures will provide more useful cues for continuous probing of brain activity (and HIFs) in response to long-duration multimedia signals.

The ERD/ERS features encode brain activity over a specific brain region. However, during the time course of experiencing long-duration multimedia signals, such as affective music video, several parts of the brain are activated to process and integrate the auditory and visual streams, as well as to evaluate emotional content via attentional and context updating mechanisms. Thus, comprising a complex flow and interplay of information between brain regions that may span different EEG frequency bands. Such interactions of neuronal networks can be quantified using features obtained from graph-theoretical analysis of brain activity [33]. Thus, in the context of complex multimedia signals, it is hypothesized that EEG-based graph-theoretical features would characterize HIFs more accurately.

Towards evaluating the proposed EEG-based features, we have explored two databases, namely the PhySyQX and DEAP [28] databases. The two databases differ based on the complexity of the stimuli used. As such, the PhySyQX database uses auditory signals produced by different TTS systems, whereas the DEAP database uses more complex audio-visual stimuli derived from affective

music videos, respectively. However, the PhySyQX database consists of multiple attitudinal and affective subjective dimensions, whereas the DEAP database only explores the affective factors.

The remainder of this chapter is organised as follows: Section 3.2 describes the databases, and features extraction procedures and classification methodology. Sections 3.3 and 3.4 provide the results and discussion for correlation and classification analyses. Lastly, conclusions are drawn in section 3.5.

3.3 Methods

This section presents the databases, features and classification strategies used to establish the validity of the features.

3.3.1 Experimental Setup:

TTS Stimuli: PhySyQX Database

The neurophysiological and subjective data used for characterization of synthesized speech QoE were obtained from the "database for physiological evaluation of synthesized speech QoE (PhySyQX)" initially described in [43] and made publicly available in [155]. The PhySyQX database has been described in Chapter 2. Here, we have only described the EEG signal preprocessing techniques that were used to clean the EEG signals.

The raw EEG data was pre-processed using EEGLAB [156], which is a freely available Matlab toolbox, following the typical pipeline. More specifically, the EEG signals were referenced to the 'Cz' electrode and bandpass filtered between 0.5 and 50 Hz with a finite impulse response filter. EEG epochs of 20-second duration were extracted covering the duration of the speech stimuli used in the experiment, time locked to the onset of the stimuli, plus a baseline rest period of 3 seconds. Epochs were extracted for each of the 44 speech stimuli, for each participant. Next, eye-blink and eye movement artifacts were removed via the widely-used independent component analysis (ICA)-based technique called ADJUST [112], a semi-automated procedure for the rejection of noisy independent components. Since the signals for the AF7 and AF8 positions were not available due to the fNIRS

montage, they had to be interpolated using the spherical interpolation technique implemented in EEGLAB.

Affective Video Stimuli: DEAP Database

The pre-processed EEG and subjective data used used for characterization of affective music videos were obtained from the publicly-available "database for emotion analysis using physiological signals (DEAP)" [28]. Here, only a brief description of the data is presented; the interested reader is referred to [28] for more details. Thirty-two healthy participants (50% females, average age = 26.9 years) were recruited and consented to participate in the study. Thirty-two channel EEG data were recorded using a Biosemi ActiveTwo system (Amsterdam, Netherlands). Data were recorded at a sampling rate of 512 Hz and electrodes were placed on the scalp according to the international 10-20 system.

The participants were presented with forty one-minute long music videos with varying emotional content. Before every video there was a baseline period of five seconds where the participants were asked to fixate at a cross in the middle of the screen. Following the presentation of each video, the participants were provided enough time to rate the music videos on a discrete 9-point scale for valence and arousal. Valence and arousal dimensions were scored using the self assessment manikins (SAM) to gauge users' emotional states [11].

The pre-processed EEG data was obtained from [157]. The initial pre-processing steps included common referencing, downsampling to 128 Hz, bandpass filtering between 4-45 Hz, and eye blink artifact removal via independent component analysis. The data was then epoched into forty 60 s long trials with a 3 s long pre-stimulus baseline. The pre-stimulus baseline was then subtracted from the preprocessed data.

3.3.2 Feature Extraction

Power Spectrum Features

For physiological assessment of multimedia QoE, the most widely-used EEG feature has been the so-called 'event-related potential' (ERP) [26, 27], which is a time- and phase-locked change in the

voltage amplitude of the EEG time series [25]. Certain neural activities, however, have been shown to be time-locked to the event, but not phase-locked [56]. As such, here we explore a more relevant feature, based on the EEG power spectrum, called event-related desynchronization/synchronization (ERD/ERS), which represents frequency-specific variations of EEG activity. This activity is indicated by the increase or decrease of the power in a given frequency band, and is termed ERS or ERD, respectively. Also, it is known that the ERD/ERS phenomenon is generated by changes in the parameters that control the oscillations in the neuronal networks, thus reflecting changes in the local interactions between neuronal assemblies [56].

The ERD/ERS phenomenon can be demonstrated in any EEG subband, but the most commonly occurring ERD/ERS phenomenon lies in the alpha subband where perceptual, judgment and memory information has been found [57, 56]. Moreover, increases in task complexity or attentional demands have also been shown to be directly related to alpha band ERD [158]. The alpha band ERD is a correlate of activation of cortical areas that are involved in sensory, cognitive or motor information processing [58], as increased cellular excitation leads to desynchronized EEG rhythms. Furthermore, an increased or widespread alpha band ERD is considered to be a result of the involvement of a larger neuronal network for information processing [56]. The factors that lead to the enhancement of alpha band ERD consist of increased complexity, more efficient task performance and more attentional demand [158, 56]. The alpha band can be divided into lower (8-10 Hz) and higher (10-12 Hz) subbands. The occurrence of lower alpha ERD is generally topographically widespread, and it reflects task demands and attentional processes. The locations of high alpha ERDs, however, are restricted topographically and are developed during sensory-semantic information processing [56]. Nonetheless, an increase in alpha band spectral peak or alpha ERS is based on synchronized behaviour of a large number of neurons that make active information processing difficult [56]. This indicates the deactivation of the cortical areas or their involvement in top-down inhibition and control [159]. Furthermore, these interpretations of alpha ERD/ERS have also been shown valid for the low-beta band [56].

In addition to alpha and beta band oscillations, induced oscillations can also be found in the theta and gamma bands. The theta band ERD/ERS reflects encoding of new information [160], whereas gamma band activities are related to binding of sensory information or sensory-motor information [161]. The alpha and beta band oscillations are too slow to act as carriers for binding at higher level processing [56]. However, gamma band oscillations are appropriate for the establishment

of rapid coupling between spatially separated neuronal assemblies [56], thus gamma band ERS reflects a stage of active information processing. As can be seen, the ERD/ERS phenomenon can characterize evoked activities that are time-locked but not necessarily phase locked to the stimulus. Furthermore, topographical ERD/ERS activities provide comprehensive information regarding the active involvement of different cortical regions of the brain in QoE assessment. The auxiliary information regarding cortical activation is useful in comprehensive understanding of perceived QoE and can help reduce the probing area for future QoE studies that leverage EEG as an assessment tool.

More recently, changes across several EEG subbands were also observed in response to different affective multimedia stimuli [28, 29]. Such findings suggest that EEGs can also be used to assess user affective states, and thus can be potentially useful for HIFs' characterization [30, 31]. Here, ERD/ERS features are also explored as correlates of human affective states.

Event related desynchronization features can be computed using three different methodologies: the classical method, the inter-trial variance (ITV) method and the Welch's method. The classical method involves band-pass filtering EEG epochs into the subbands of interest, followed by squaring the samples and averaging over the epochs. For the classical method, the instantaneous power $P_{(j)}$ can be computed as:

$$P_{(j)} = \frac{1}{N}\sum_{i=1}^{N} x_{(i,j)}^2, \tag{3.1}$$

where $P_{(j)}$ is the averaged power estimation of the subband signal, $x_{(i,j)}$ is the j-th sample of the i-th epoch of the subband signal and N is the total number of epochs. However, it has been found that in the classical method, a phase-locked power increase due to the ERP can mask the non-phase-locked power decrease (ERD) [59]. Therefore, to mitigate this effect the ITV procedure was introduced in [59], which requires the step of squaring the samples to be replaced with calculation of the point-to-point ITV, where ITV is computed as:

$$ITV_{(j)} = \frac{1}{N-1}\sum_{i=1}^{N}\{x_{(i,j)} - \bar{x}_{(j)}\}^2. \tag{3.2}$$

The parameter $\bar{x}_{(j)}$ corresponds to the mean of the data at the j-th sample, averaged over all subband epochs. However, Welch's method estimates the power spectral density based on the windowed computation of periodogram [60].

Features typically correspond to the percentage ERD, which can be quantified as the percentage of change of power or ITV ($A_{(j)}$) at each sample point or an average of some samples (in this analysis the samples are averaged in time over 2 seconds) relative to average power (or ITV) during a baseline period. More specifically,

$$ERD_{(j)} = \frac{B - A_{(j)}}{B} \times 100\%, \tag{3.3}$$

where B is the average power or ITV in the baseline interval (in this study we considered a baseline period that extended from 500 ms before the stimulus until the start of the stimulus for ERD computation), and is computed as:

$$B = \frac{1}{k} \sum_{j=n_0}^{n_0+k} A_{(j)}. \tag{3.4}$$

Another class of features, based on power spectral features, encodes asymmetric differences in power spectral density of EEG sub-bands and is called asymmetry index (AI). These are computed as the ratio of ERDs computed from asymmetrically placed right and left hemispheric electrodes, as follows:

$$AI = \frac{ERD_{right}}{ERD_{left}} \tag{3.5}$$

The AI-based features have proven useful in characterising affective states, as reported in [28].

The ERD components for EEG signals obtained from the DEAP database (for affective videos) were computed using Welch's method (described above), in accordance with [28]. The ERD was computed for 10 EEG subbands, namely: theta (4-8 Hz), low-alpha (8-10 Hz), high-alpha (10-12 Hz), alpha (8-12 Hz), low-beta (12-18 Hz), mid-beta (18-24 Hz), high-beta (24-30 Hz), beta(12-30 Hz), gamma (30-50 Hz) and full (4-50 Hz), as suggested in [32]. The ERD was computed for each electrode, for two-second-long non-overlapping windows, and was then averaged, thus resulting in 320 ERD features (32 electrodes $\times$ 10 sub-bands). Following that, AI features derived from 14 asymmetrically placed electrodes were computed for each sub-band, thus resulting in 140 AI features.

For the PhySyQX database (for TTS stimuli), to reduce the masking of non-phase locked ERD components, the ITV method for ERD computation was used in the experiments described herein. The ERD was computed for the 10 EEG subbands mentioned above. The ERD was computed for each electrode, for two-second-long non-overlapping windows covering 14 seconds (as the shortest speech stimulus was 14 seconds long) after stimulus presentation. Following ERD computation, five statistical features (i.e., mean, median, standard deviation, skewness and kurtosis) were derived across the 7 windows at each electrode, for each subband. This resulted in the development of 10 feature sets (each set corresponding to a different EEG subband), each comprised of 320 features (64 electrode $\times$ 5 statistical features). Furthermore, AI features derived from 27 asymmetrically placed electrodes were computed for each EEG sub-band, resulting in 270 AI features.

Cross Spectrum Features

Neuronal QoE and affect processing, induced by audio-video stimuli, is achieved through a variety of intermediate steps arising from the interpretation of both auditory and visual sensory information. These information streams have to be integrated in order to correctly realize the emotional significance of the stimuli and hence initiate an appropriate behavioral response. The so-called temporal correlation hypothesis addresses this issue of integrating brain signals separated over time and space in order to realize a complete unity, also known as the 'binding problem hypothesis' [162]. According to this hypothesis, neurons with similar feature properties can synchronize their discharges under certain specific circumstances. Previous studies have demonstrated the existence of local-scale synchrony/dependency [163, 164] between adjacent neuronal processes as well as large-scale synchrony/dependency [165, 166] between distant neurons. This measure of synchrony is then used to estimate the functional connectivity of the brain.

Different neuroimaging methods can be used (e.g., EEG, MEG) to quantify the synchronous activity of various interdependent brain regions using linear or non-linear metrics [167]. Here, we use the popular magnitude squared coherence (MSC) metric, based on electrode cross spectra, to quantify linear EEG synchrony in different frequency bands, as MSC has been shown to be associated with information regarding emotions [168]. These synchrony metrics were applied to all the possible pairs of EEG electrodes by splitting the recorded EEG data, for each trial, into 2-second long epochs (without overlap). The obtained values were then averaged over the epochs

in order to estimate the MSC metrics with more confidence. The value of these metrics denotes the strength/weight of the connection between two scalp regions (i.e., EEG electrode positions). Since the metrics are *non-directional*, the information transfer is independent of the direction of information flow between two regions.

The magnitude squared coherence (MSC) is a large scale measure of the underlying dynamic neuronal interactions; higher coherence values indicate greater functional interplay between the two underlying neuronal networks. In order to quantitatively measure coherence, cross-spectral analysis techniques are commonly applied [169], where the cross power spectral density ($|S_{xy}|^2$) of two signals x and y at frequency f is normalized by the product of each signal power spectral density (S_{xx} and S_{yy}), i.e.,

$$MSC_{x,y}(f) = \frac{|S_{xy}(f)|^2}{S_{xx}(f) * S_{yy}(f)}. \tag{3.6}$$

MSC values range from [0,1] with 0 signifying no correlation and 1 perfect correlation between the two signals under consideration. More specifically, the MSC values were calculated across ten EEG subbands.

The estimated functional connectivity patterns (also known as "connectomes") can be represented using graph theory. A graph is a set of nodes or vertices linked by connections or edges, thus forming an abstract representation of the interactions between the elements of a complex real world system [170]. For human brain networks, for example, nodes can be a representation of neurons (for fMRI) or electrodes (for EEG/MEG), whereas edges depict the connection representing the information being transferred (e.g., given by MSC). The edges of a graph can traverse in either direction, thus forming undirected graphs. Nevertheless, if the edges traverse only in one particular direction, a directed graph is formed. However, the graphs resulting from MSC metrics are *undirected graphs*. The weight of an edge corresponds to the strength of the interaction measure used, such as coherence or normalized mutual information. Weak (insignificant) edges have lower weights and likely represent spurious/noisy connections [171], thus graph thresholding is needed to generate unweighted/binary graphs.

A generic binary graph G consisting of N nodes and K edges is represented by an adjacency matrix (or connection matrix) a_{ij}, which is an $N \times N$ matrix. The elements of a_{ij} are 0s or 1s depending on the threshold used; 1 represents a connection between two nodes i and j and 0 represents the lack of a connection. In this study, symmetrical and undirected adjacency matrices

were computed for all 40 stimuli and for each of the 32 participants with eight different thresholds, ranging from $0.2-0.9$ in 0.1 increments. From these binary graphs, various metrics can be computed to quantify the structural properties of the networks. For example, the degree k_i represents the number of edges incident to the node i; the shortest path length d_{ij} indicates the distance between two nodes i and j quantified by the least number of connections that link them. Once d_{ij} is computed for all possible node pairs, several more complex graph metrics can be computed, as detailed below.

Global graph metrics directly measure the level of integration in a particular network [171]. One such measure is the characteristic path length L of a graph. It is defined as the average distance between two connected nodes of the graph and is given by:

$$L = \frac{1}{N(N-1)} \sum_{i \neq j} d_{ij}, \tag{3.7}$$

This measure is a direct reflection of the rate of information flow globally within a network. To better understand the concept of the measure, we can draw a parallel with a transportation network, where the shortest path length between two stations is the smallest sum of actual physical distances throughout all possible paths between them. Thus, paths having the shortest path length result in faster and more efficient transfer of goods in the network. However, L specifically measures the efficiency of information transfer for a *sequential* network [172], where only one "packet" of information travels through the network. A parallel network, on the other hand, can have all network nodes concurrently exchanging packets of information. Such scenario better represents the properties of the human brain connectome. Therefore, another measure of flow of global information (or communication) efficiency for parallel networks is used; such measure is called the global efficiency E_g of a graph. For a fully connected graph, $E_g = 1$, whereas for an empty graph, $E_g = 0$. The global efficiency is given by the arithmetic mean of the inverse of the distances (d_{ij}) between each electrode pair, i.e.,

$$E_g = \frac{1}{N(N-1)} \sum_{i \neq j} \frac{1}{d_{ij}}, \tag{3.8}$$

Local graph metrics directly measure the level of segregation in a particular network [171]. More specifically, they measure the fault tolerance of the network, or the efficiency of information transfer between the first neighbours of the node under consideration. A commonly used measure is the so-called clustering coefficient C, which is the average over all clustering coefficients $C_i, i = 1, \ldots, N$ in the network. Each C_i corresponds to the number of edges existing in the subgraph S_i, which

in turn, is the graph that remains after node i is removed from the main graph. The clustering coefficient C is mathematically given by:

$$C = \frac{1}{N} \sum_{i} C_i, \tag{3.9}$$

Alternately, the so-called local efficiency metric E_l measures the local properties of the graph and its tendency to form clusters. It quantifies the efficiency of information transfer in the specialized regions of the brain. Similar to E_g, for a fully connected graph, $E_l = 1$ and for an empty graph, $E_l = 0$. The local efficiency metric E_l is computed as follows:

$$E_l = \frac{1}{N} \sum_{i=1}^{N} E_g(S_i), \tag{3.10}$$

Commonly, three network types are considered: random, regular, and small-world. Random networks have high global efficiency, but very low local efficiency. Regular networks, on the other hand, have high local efficiency and low global efficiency. Lastly, small-world networks have both high global *and* local efficiencies [172]. Recently, in [171] it was proposed that human brain networks have likely evolved to balance both segregated and integrated processing of information, thus suggesting that the brain networks are small-world networks having high degree of both segregated as well as integrated processing. The highly segregated processing leads to a reduction in neuronal wiring "cost" as it reduces the number of unnecessary long distance connections. However, removing all the long distance connections would delay information transfer, thus causing an increase in energy cost to transfer information between distant regions. As a consequence, small-world networks balance this energy requirement by increased integrated processing, keeping important long distance connections intact. This characteristic feature of brain networks is also consistent with the so-called global workspace theory [173], which argues that synchronized oscillations emerge along large ensembles of widely distributed workspace neurons while attending to more salient stimuli [170]. Ultimately this leads to an overall increment in efficient information transfer in the brain.

To characterize the small-worldness of a graph, the clustering coefficient (C) and characteristic path length (L) of the graph under consideration are normalized, respectively, by the clustering coefficient (C_{rand}) and characteristic path length L_{rand} of a random graph with the same number of nodes, edges, and degree distribution as the network of interest. The small-worldness coefficient

S is thus given by [33]:

$$S = \frac{C/C_{rand}}{L/L_{rand}}, \tag{3.11}$$

where $S \geq 1$ indicates small-worldness.

The above mentioned features formed the graph-theoretic feature set that consisted of 45 features (5 graph metrics $\times$ 9 threshold levels) per EEG sub-band. Once these graph features were computed, we used them to solve the emotion classification problem. The details for classifier formulation and methodology are detailed in the following subsection.

3.3.3 Neural Correlates

Towards the understanding the neural underpinnings of HIFs and to quantify the relationship between the HIFs and EEG-based features, two approaches are taken. First we analyze the ERD_α and ERD_γ patterns, as these features have been found to be correlated with cortical activation [58, 56], for each subject to investigate neural activation patterns and explore their differences between varying quality of stimuli and then use Pearson's correlation statistic over all subjects to get a sense of correlates for classification. As in [28, 29], this analysis was performed for each subject individually and then, assuming independence, by pooling the resulting p-values, corresponding to each subject, per correlation direction (+ve and -ve) and feature into one p-value using Fisher's method [174]. The topographical maps of mean correlation values for ERD features, per sub-band, were then plotted for the PhySyQX database where significant EEG electrodes were highlighted. Furthermore, the features extracted from the significant channels could be important for classification. Also, to validate the cortical activation patterns, we extracted the EEG channels that showed significant differences in ERD_α and ERD_γ for high ($HQ = MOS \geq 3$) and low ($LQ = MOS < 3$) quality systems, as computed from unpaired t-test. Moreover, to reduce the false detection rate of significant channels, a Benjamini—Hochberg-–Yekutieli correction was implemented. Unfortunately, the DEAP database does not consist of quality scores for the stimuli, hence similar analysis was not possible for affective video stimuli. However, similar correlation analysis was performed in [28], for the DEAP database and hence, it is not reported here. The correlations for graph-theoretical features were computed for each of the nine investigated thresholds $(0.1, \ldots, 0.9)$ however, only the thresholds at which significance was achieved were used in the following classification analyses.

3.3.4 Classification Methodology

TTS Stimuli: PhySyQX Database

For physiological monitoring of multimedia QoE, typically binary classification of subjective dimensions is performed (e.g., high/low arousal) [28, 29, 45] where the subjective ratings from each participant are used as ground truth. In our experiments, the subjective ratings scored on a 9-point scale (e.g., valence and arousal) were binarized using a threshold at 5 and those using 5-point scales at a threshold of 3. Note that the subjective dimensions of dominance and speaking rate were excluded from classification analyses due to low reliability of the obtained scores, as reported in [43]. Pruned decision tree classifiers were trained using the Gini index and within a subject-wise classification task. As in [28, 29, 45], the developed classifiers were cross-validated using the leave-one-sample-out technique. Classifiers were obtained for each feature set separately, as well as fused using the equal-weighted fusion scheme, as proposed in [76]. Classifier performance is quantified using classification accuracy and the weighted F1-score; the latter allows more accurate estimation of classifier performance for imbalanced classes [61], as is the case with the PhySyQX dataset. To quantify significance of the obtained classifiers, comparisons with a random voting classifier was used, where random classes were assigned to the predicted classes.

Affective Video Stimuli: DEAP Database

Towards classifying affective states, support and relevance vector machine classifiers were implemented to solve four different binary classification problems: low/high valence, low/high arousal, low/high dominance and low/high liking. Towards this end, we used the features obtained from a graph theoretical approach (denoted as *Graph*) for quantifying brain connectomes. For comparison purposes, the ERD features along with asymmetry index (AI) features (as described in [28]) were also used for classification.

For classifying various emotional dimensions, each participant's subjective ratings were used as the ground truth values. In order to form the low and high classes, the subjective scores were thresholded at the mid-point of a 9-point scale, i.e., at 5. This approach resulted in unbalanced classes for each subjective rating, which was also reported in [28, 175]. In [28], the authors propose using the F1-score to reliably report the results while tackling the class imbalance. Therefore,

in this study we have reported F1-scores along with classification accuracies to quantify classifier performance.

Once the targets were obtained, the SVM and RVM classifiers were implemented using the Scikit-learn library in Python [176] and the Pattern Recognition Toolbox for MATLAB [177], respectively. During pilot studies several kernels were tested but the radial basis function (RBF) kernel showed improved performance for both the SVM and RVM classifiers, and thus is used throughout the remainder of this chapter. For classification, the data was split into two sets for each classification problem. The first is a 'development set' consisting of 5 randomly chosen trials from each class (high/low) for each subject. If 5 samples for each of the two classes were not available for a particular classification problem, then 10 samples were randomly chosen, from each subject, for this data set. The second is an 'evaluation set' consisting of the remaining data from each subject. Therefore, in total, the development set had 320 samples (32 x 10), whereas the evaluation set had 960 samples (32 x 30). The development set was used to search for the best parameters λ and γ_{RBF} over a grid of possible values which extended in this range $[10^{-10} : 10^{1}]$. Moreover in [175], the authors suggested using the inverse of the number of features as a good estimate of γ_{RBF}; therefore these values were also included in the grid. After obtaining the optimum parameter values, the evaluation set was used to determine the classifier performance metrics.

Towards this end, classifier metrics were determined per subject, following a leave-one-sample-out cross-validation scheme. The SVM classifier was implemented using the optimal λ and γ_{RBF} values determined from the development set. However, to harness the probabilistic outputs ($Prob$) of RVMs, the probabilities could be thresholded to assign weights to a particular class. For this, we developed two sets of RVM classifiers denoted as RVM_1 and RVM_2 where:

- RVM_1 assigns equal weights to both the classes (high/low), by thresholding the probabilities for both the classes at 50% (Rule: if $Prob_{high} \geq 50\%$, then predicted class = 'high', else predicted class = 'low').
- RVM_2 searches for the best probability threshold (T_{best}), over the development set, such that the F1-score is highest at that threshold. The threshold search for class 'high' extended from 51% to 60% incrementing in steps of 1%. This helps in prediction of class 'high' with more confidence and thus helps in tackling the unbalanced data sets. (Rule: if $Prob_{high} \geq T_{best}$, then predicted class = 'high', else predicted class = 'low').

RVM_1 is expected to perform well for classification problems with balanced classes, whereas RVM_2 is expected to perform well for classification problems with unbalanced classes. As reported previously [28, 175], the DEAP dataset is highly unbalanced, therefore RVM_2 is expected to outperform all other classifiers. For our experiments, multiple versions of the SVM, RVM_1 and RVM_2 classifiers were obtained. More specifically, ten versions of each classifier type were trained with graph features from each of the 10 frequency sub-bands listed at the end of section 3.3.2. Generally, the graphs created using very low threshold values (e.g., < 0.1) on MSC values are affected by volume conduction effects, whereas for a low number of nodes it is possible to create disconnected graphs with high threshold values (e.g., > 0.5) [33]. Such graphs cannot be interpreted as human brain functional connectivity networks, thus were not considered in this work. Therefore, each classifier was trained on a 20-dimensional feature set, corresponding to the five graph theory features described in section 3.3.2 (i.e., equations 3.7 - 3.11) computed using 4 different thresholds on the MSC values (i.e., 0.2-0.5 in 0.1 increments). We found in preliminary analyses that combining features extracted from the four different thresholds resulted in better overall performance than each individual threshold alone. For comparison purposes, the same classifier types were framed using conventional ERD, asymmetry index (AI) features, and their combination, computed for the same 10 frequency sub-bands, following the methodology proposed in [28].

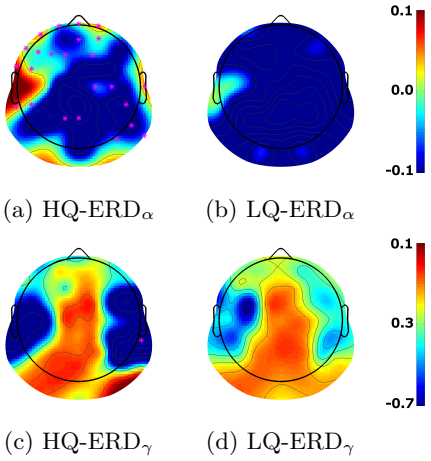
A secondary analysis performed consisted of decision-based fusion. Fusion of decisions from RVM_2 classifiers was also performed separately for each feature set using majority voting with equal weights to all the 10 classifiers, as suggested in [28]; in the case of a tie, a class was randomly chosen.

3.4 Experimental Results

3.4.1 Neural Correlates

TTS Stimuli: PhySyQX Database

Towards understanding the neural underpinnings behind human QoE perception and the effects of HIFs on QoE, ERD_α and ERD_γ patterns were analysed, as shown in Fig. 3.1. As can be seen,



(a) HQ-ERD$_\alpha$ (b) LQ-ERD$_\alpha$

(c) HQ-ERD$_\gamma$ (d) LQ-ERD$_\gamma$

Figure 3.1 – Average ERD$_\alpha$ and ERD$_\gamma$ for high ($HQ = MOS \geq 3$) and low ($LQ = MOS < 3$) quality systems. The electrodes with significant ($p < 0.05$) differences between HQ and LQ are highlighted with a magenta coloured '$*$' symbol.

higher ERD$_\alpha$ and lower ERD$_\gamma$ values are seen over temporal regions of the scalp for HQ systems as compared to LQ systems, indicated by highlighted significant channels.

Moreover, neural correlates of QoE percepts were analyzed using topographical maps for average correlations between neural features and all subjective dimensions of QoE to validate the findings from Fig. 3.1. Furthermore, as the majority of the subjective dimensions followed similar topographical correlation patterns, the topographical correlation maps are shown only for the 'overall impression' QoE dimension and EEG in Fig. 3.2. From Fig. 3.2, it is evident that increasing speech quality induced increases in ERD_θ, $ERD_{l-\alpha}$, $ERD_{h-\alpha}$, ERD_α, $ERD_{l-\beta}$ and ERD_β in the left and right fronto-temporal regions of the brain. However, this effect was significant only for $ERD_{h-\alpha}$, ERD_α, $ERD_{l-\beta}$ and ERD_β, as indicated by the highlighted electrodes with significant correlation. A significant correlation over left and right fronto-temporal regions for $ERD_{l-\alpha}$, $ERD_{h-\alpha}$, ERD_α and $ERD_{l-\beta}$ was consistently obtained for all subjective dimensions.

Moreover, $ERD_{h-\beta}$ and ERD_γ consistently showed significant correlations over occipital and temporal regions of the brain, with all subjective dimensions. In Fig. 3.2, $ERD_{h-\beta}$ showed significant positive correlations with overall impression of the signal quality in the frontal region of the brain, whereas ERD_γ showed significant positive correlation with overall impression in the occipital region of the brain. Also, it can be observed that $ERD_{h-\beta}$ and ERD_γ in the fronto-temporal regions decreased with decreasing signal quality, however this effect was not significant. In general, high frequency bands, such as γ and $h-\beta$ bands, showed inverse patterns compared to the low

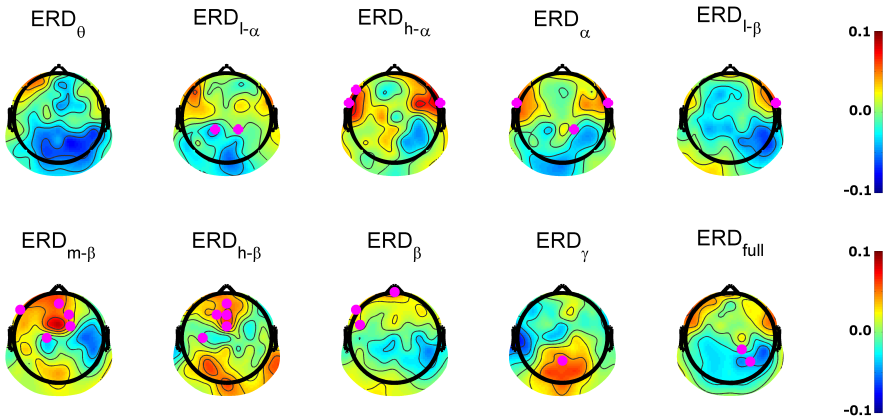


Figure 3.2 – Topographical maps of the average correlation between ERD and the MOS rating. Electrodes with significant correlations ($p < 0.05$) are highlighted with a magenta coloured '•' symbol.

frequency bands, such as θ, $l-\alpha$, $h-\alpha$, α and $l-\beta$ bands. Furthermore, the ERD_{full} band showed significant correlations in fronto-temporal and occipital regions of the brain, depending on the subjective dimension. For overall impression, the ERD_{full} band showed significant negative correlation in the occipital region.

Similar correlation analysis with AI features revealed significant correlations with subjective dimensions. The significant subjectwise correlation coefficients ranged between 0.09 and 0.12 and were distributed over all EEG sub-bands. Moreover, it was noticed that the AI features derived from the α sub-band over the frontal region of the scalp were positively correlated with subjective dimensions, such as valence and overall impression. The AI features derived from the γ and $full$ sub-bands, computed over centro-parietal region, were negatively correlated with valence and overall impression.

The subjectwise correlation analysis between subjective dimensions and graph theoretic features, in turn, revealed many significantly correlated features between thresholds of 0.2 and 0.6. The majority ($> 50\%$) of the features represented segregated processing metrics (local efficiency and clustering coefficient). Moreover, for several subjective dimensions, the graph metrics derived from high frequencies ($> m-\beta$) and full bandwidth signal showed highest average correlation coefficients ($r_{subjectwise} = 0.09$). The graph metrics, such as E_l, E_g, C_{mean} and S were positively correlated with subjective dimensions (except CP), whereas L showed negative correlation with subjective dimensions (except CP).

Affective Videos Stimuli: DEAP Database

The subjectwise correlation analysis for the affective video dataset revealed significant correlations between affective dimensions and ERD and AI features as reported in [28]. In a similar vein, subjectwise correlations were computed between affective dimensions and graph-theoretic features. It was observed that significance was only attained for features computed at thresholds between 0.2 and 0.5, thus resulting in 200 features that were used for classification analyses. Amongst the significantly correlated features, measures of segregation, integration and small-worldness were equally distributed. However, positive correlation coefficients were observed for E_l, E_g, C_{mean} and S, and negative correlations were obtained for L in most of the EEG sub-bands. For the arousal subjective dimension, significant correlations were obtained for most of the EEG sub-bands whereas, for valence, significant correlations were obtained for higher sub-bands ($> l - \beta$).

In order to visualize the significant differences between the 'high' and 'low' categories, as well as the interplay between the integration and segregation modules, the data was tested for normality using Shapiro-Wilk test followed by t-tests between 'high' and 'low' category samples for E_l and E_g. The data for different bands was found to be normal with $W < 0.9$ and $p > 0.05$. The measures of local and global efficiencies (computed at an empirically selected threshold $T = 0.5$) were found to be significantly higher for stimuli that invoked high valence at higher frequency bands ($\geq$l-β), as shown in Fig. 3.3. However, the measures of local and global efficiencies were found to be significantly higher for stimuli that invoked higher arousal at both lower and higher frequency bands (θ-γ bands), as shown in Fig. 3.4.

3.4.2 Classification Results

TTS Stimuli: PhySyQX Database

Table 3.1 reports average accuracies and F1-scores over all participants for each feature set for EEG, for each subjective dimension. The table indicates the best performing classifiers from individual EEG sub-bands along with the decision fusion for each feature set. The performance for each classifier was tested using a two-sided repeated samples t-test over the concatenated results from each rating scale and participant, as suggested by [28]. It was observed that most of the classifiers performed significantly above chance; however, the classifiers that did not result in significantly

above chance performance are indicated by a superscripted bullet. Moreover, it was observed that the decision fusion-based classifiers performed better than classifiers based on best performing individual EEG sub-bands. However, this observation was significant (indicated using superscripted '$*$' in Table 3.1) only for MOS and CP dimensions using the ERD feature set and CP dimension for the AI feature set, as tested using a paired t-test. Furthermore, comparing performances of different EEG-based feature sets, using ANOVA, resulted in no significant differences. From Table 3.1, it is evident that, using a combination of individual feature sets, it is possible to classify *all* subjective QoE dimensions (except the voice emotion dimension using graph features) with above-chance performance. For comparison, the table also reports the average performance of a random classifier developed using random voting, as well as the average percentage and standard deviation of positive class labels ('label 1') to quantify class imbalance. Finally, it was found that the fusion of graph, ERD and AI-based classifiers resulted in $1-2\%$ increase in performance for each dimension; however, this observation was not significant.

Affective Videos Stimuli: DEAP Database

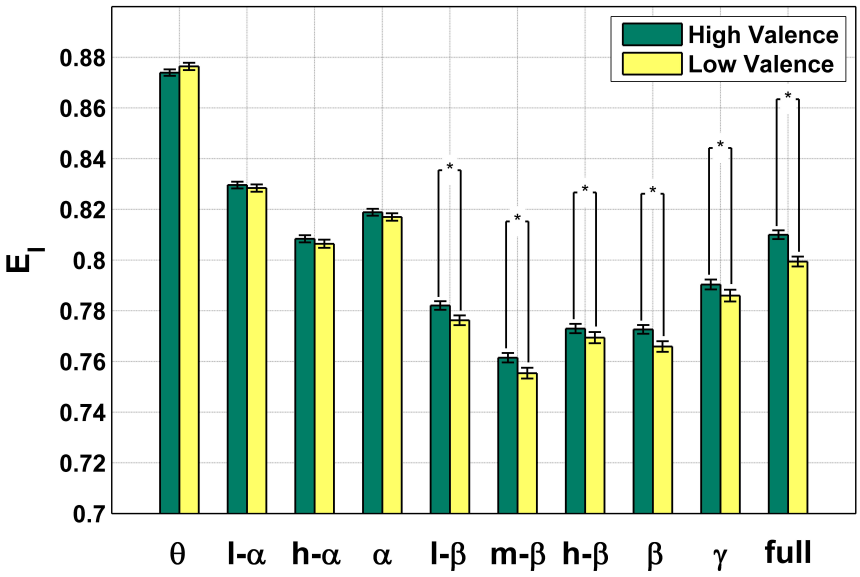
Table 3.2 reports the F1-score and classification accuracy for the classifier corresponding to the frequency sub-band with maximum F1-score chosen amongst the 10 individual classifiers for the arousal and valence categories. The computed F1-scores were then tested for significance against a random chance level of 50% using an independent one-sample t-test, as proposed in [28]. It can be observed that graph, spectral power and AI features perform significantly better than chance in classifying users' emotional states. Following that, repeated measures ANOVA with Bonferroni-Holm correction for multiple comparisons was used to test significant differences between the performance of classifiers obtained from graph features and traditional features (such as ERD and AI). There was a significant ($p < 0.05$) improvement in classification accuracy of affective dimensions using the proposed graph-theoretical features compared to traditional features, as shown in Table 3.2. Moreover, similar significance tests between the classifiers developed from different sub-bands were carried out for each affective dimension. This revealed significant differences for the valence dimension ($F(9, 310) = 33.5, p < 0.05$). The post-hoc multiple comparison tests for valence revealed better performance of classifiers developed using graph features computed at higher frequency bands ($\geq$l-β) compared to lower frequency bands ($\leq$l-β). Similar comparative tests using graph features for arousal classification resulted in insignificant ($p > 0.05$) differences. Also, similar tests using ERD

Table 3.1 – Average accuracies (Acc) and F1-scores (F1) over participants, for each EEG-based feature set. Superscripted bullets (•) indicate results that are not significantly higher than chance according to an independent one-sample t-test (• = $p < 0.05$). Superscripted asterisks ($*$) indicate the decision fusion-based classifiers that perform significantly better than the best performing classifiers based on individual EEG sub-bands ($* = p < 0.05$). The penultimate row benchmarks the system based on a random voting classifier. The last row presents the mean and standard deviation of the percentage of positive class labels across subjects.

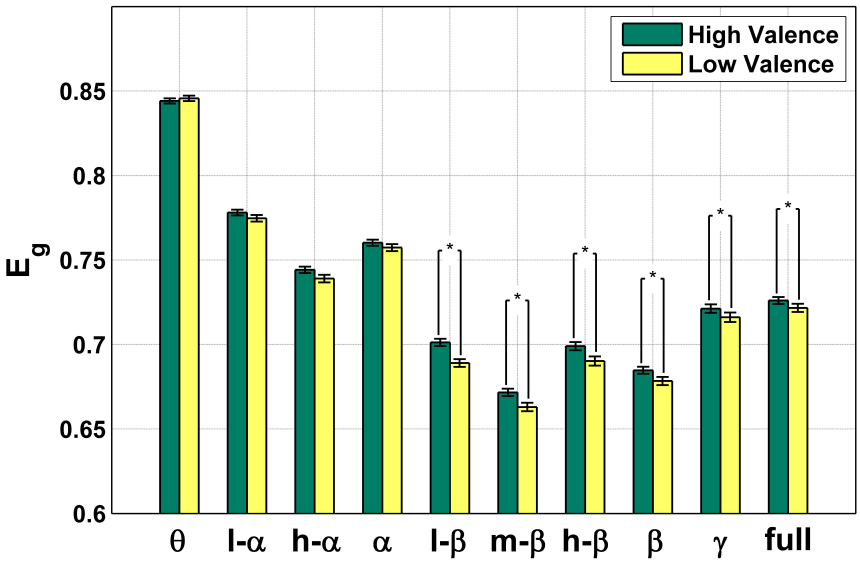
Feature	Bands	Metric	MOS	VP	Ac	Int	Nat	LE	VE	Val	Ar	CP
ERD	Individual	Acc	0.58	0.58	0.59	0.60	0.59	0.57	0.56	0.59	0.56	0.61
		F1	0.58	0.58	0.59	0.60	0.59	0.57	0.56	0.58	0.56	0.61
		Band	$full$	$l-\beta$	$full$	$h-\alpha$	$full$	$h-\beta$	$full$	β	β	β, $full$
	Fusion	Acc	0.64^{*}	0.59	0.64	0.60	0.62	0.58	0.57	0.60	$0.52^{\bullet}$	0.67^{*}
		F1	0.62	0.57	0.61	0.57	0.59	0.55	0.56	0.58	$0.50^{\bullet}$	0.63
AI	Individual	Acc	0.56	0.60	0.59	0.57	0.57	0.56	0.55	0.58	$0.53^{\bullet}$	0.64
		F1	0.56	0.60	0.59	0.57	0.57	0.55	0.55	0.57	$0.53^{\bullet}$	0.65
		Band	β	β	$full$	β	θ	$m-\beta$	$l-\alpha$	$full$	$m-\beta$	$m-\beta$
	Fusion	Acc	0.59	0.59	0.61	0.61	0.60	0.61	0.57	0.56	$0.54^{\bullet}$	0.70^{*}
		F1	0.58	0.57	0.58	0.59	0.57	0.59	0.55	0.55	$0.53^{\bullet}$	0.67
ERD+AI	Individual	Acc	0.61	0.63	0.64	0.59	0.63	0.62	0.59	0.61	0.58	0.67
		F1	0.61	0.63	0.64	0.59	0.62	0.61	0.59	0.60	0.58	0.67
		Band	$full$	$l-\beta$	$full$	$h-\alpha$	γ	$m-\beta$	$l-\alpha$	β	$l-\beta$	α
	Fusion	Acc	0.62	0.62	0.64	0.61	0.64	0.61	0.57	0.59	$0.54^{\bullet}$	0.69
		F1	0.58	0.59	0.59	0.56	0.59	0.56	$0.54^{\bullet}$	0.56	$0.51^{\bullet}$	0.63
Graph	Individual	Acc	0.55	0.56	0.59	0.56	0.59	0.57	$0.53^{\bullet}$	0.56	$0.53^{\bullet}$	0.62
		F1	0.55	0.55	0.60	0.56	0.58	0.57	$0.53^{\bullet}$	0.56	$0.53^{\bullet}$	0.63
		Band	$full$	$full$	$h-\beta$	$full$	$h-\beta$	α	α	α	β	α
	Fusion	Acc	0.58	0.60	0.60	0.59	0.59	0.58	$0.54^{\bullet}$	0.57	0.55	0.67
		F1	0.55	0.58	0.56	0.56	0.56	0.57	$0.52^{\bullet}$	0.55	0.55	0.63
Random Voting	-	Acc	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50
		F1	0.50	0.50	0.50	0.50	0.51	0.51	0.50	0.50	0.50	0.52
+ve Class Percentage	-	Mean	0.50	0.46	0.40	0.49	0.39	0.44	0.44	0.47	0.49	0.35
		St. Dev.	0.13	0.13	0.13	0.14	0.13	0.15	0.12	0.13	0.11	0.19

and AI features for valence and arousal classification revealed insignificant ($p > 0.05$) differences. The significant sub-bands are reported in Table 3.2.

Table 3.3, in turn, shows the performance achieved with decision-level fusion of the individual RVM_2 classifiers and majority voting. Again, the computed F1-scores were tested for significance against a random chance level of 50% using an independent one-sample t-test. Moreover, repeated measures ANOVA with Bonferroni-Holm correction was carried out for the different classifier groups, and the pairs with significant differences are reported in Table 3.3. From Table 3.3 it can be noticed that all the classifiers performed significantly better than chance, and classifiers corresponding to a decision fusion of graph features performed significantly better than the corresponding classifiers utilizing traditional features. Furthermore, comparing the individual classifiers and decision fusion-based classifiers, the decision fusion classifiers performed better than individual classifiers; however this observation was not significant, as tested using a t-test. Finally, as observed for the TTS stimuli,



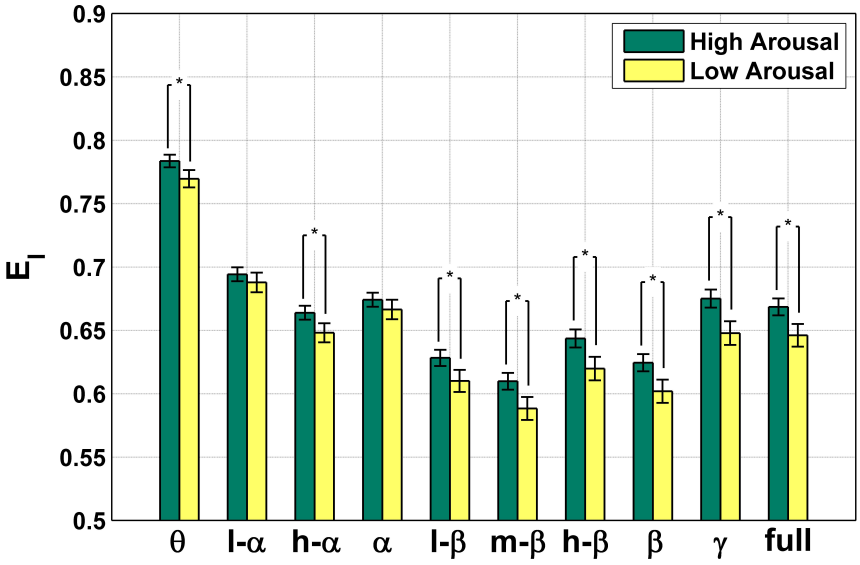
(a) Local efficiency (E_l).



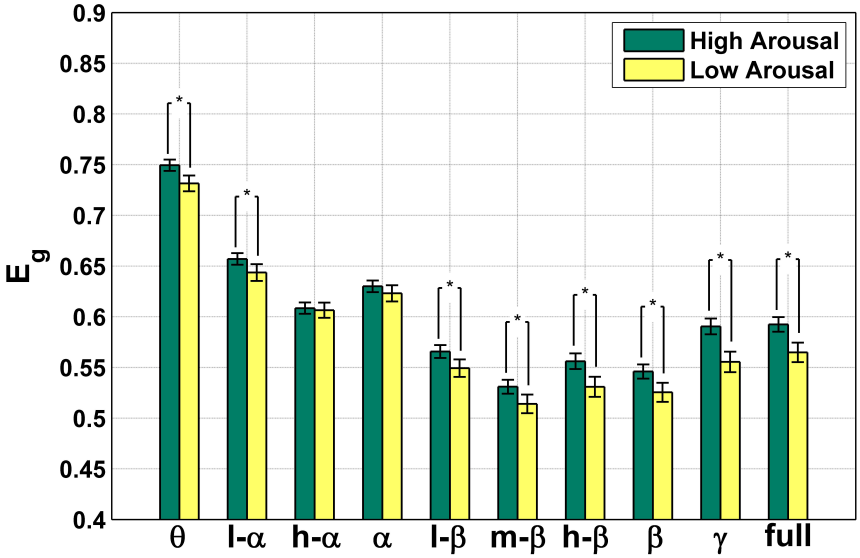
(b) Global efficiency (E_g).

Figure 3.3 – Changes in local efficiency and global efficiency, across different frequency sub-bands with high and low valence stimuli. The significant pairs ($p < 0.05$) tested using a t-test are represented with a '$\star$'.

the fusion of graph, ERD and AI-based classifiers resulted in $1-2\%$ increase in performance for each dimension.



(a) Local efficiency (E_l).



(b) Global efficiency (E_g).

Figure 3.4 – Changes in local efficiency and global efficiency, across different frequency sub-bands with high and low arousal stimuli. The significant pairs ($p < 0.05$) tested using a t-test are represented with a '$\star$'.

Table 3.2 – Performance comparison of different classifier types using F1-scores ('F1') and accuracy ('Acc'). Reported results are for the classifiers with the highest accuracy scores. Superscripted stars indicate whether the F1-score distribution over subjects is significantly higher than chance according to an independent one-sample t-test ($** = p < 0.01$, $* = p < 0.05$). To denote a significantly ($p < 0.05$) higher F1-score of graph features in comparison to ERD, AI or, combined ERD and AI features, as obtained from repeated measures ANOVA, a subscripted '†', '$\star$' or '$\bullet$' were used, respectively. Subscripted '‡' was used when Graph features performed significantly better than all other features sets. Row 'T' reports the optimal threshold found for the RVM_2 classifier. Also, sub-bands which resulted in significantly better performing classifiers are denoted with superscripted '$\diamond$'.

Subjective Dimensions	Metric	Graph Features			ERD			AI Features			ERD + AI		
		SVM	RVM_1	RVM_2	SVM	RVM_1	RVM_2	SVM	RVM_1	RVM_2	SVM	RVM_1	RVM_2
Valence	Acc	64	64	65	58	57	58	57	57	57	57	58	58
	F1	$63^{**}_{\star\bullet}$	$63^{**}_{\ddagger}$	$65^{**}_{\ddagger}$	58^{*}	57^{*}	59^{*}	57^{*}	56^{*}	58^{*}	57^{*}	56^{*}	56^{*}
	T	-	50	51	-	50	58	-	50	51	-	50	52
	Band	$\beta^{\diamond}$	l-$\beta^{\diamond}$	$\gamma^{\diamond}$	β	l-β	l-β	β,γ	*full*	*full*	β	l-β	l-β
Arousal	Acc	64	68	68	58	60	61	57	58	61	58	60	61
	F1	$63^{**}_{\star}$	$68^{**}_{\ddagger}$	$68^{**}_{\ddagger}$	58^{*}	60^{*}	61^{*}	57^{*}	58^{*}	61^{*}	58^{*}	60^{*}	60^{*}
	T	-	50	51	-	50	51	-	50	51	-	50	52
	Band	h-β	*full*	*full*	β	β	β	m-β	l-β	l-β	β	l-β	l-β

Table 3.3 – Performance comparison after decision level fusion of the RVM_2 classifiers using F1-scores ('F1') and accuracy ('Acc'). Superscripted stars indicate whether the F1-score distribution over subjects is significantly higher than chance according to an independent one-sample t-test ($** = p < 0.01$, $* = p < 0.05$). To denote significantly ($p < 0.05$) higher F1-score of graph features in comparison to ERD, AI or, combined ERD and AI features, as obtained from repeated measures ANOVA, a subscripted '†', '$\star$' or '$\bullet$' were used, respectively. Subscripted '‡' was used when Graph features performed significantly better than all other features sets. Row 'T' reports the optimal threshold found for the RVM_2 classifier.

Subjective Dimensions	Metric	Graph Features	ERD	AI Features	ERD + AI Features
Valence	Acc	67	60	60	60
	F1	$67^{**}_{\ddagger}$	59^{*}	60^{*}	60^{*}
Arousal	Acc	69	60	64	65
	F1	$68^{**}_{\dagger}$	60^{*}	64^{**}	65^{**}

3.5 Discussion

3.5.1 Neural Correlates

TTS Stimuli: PhySyQX Database

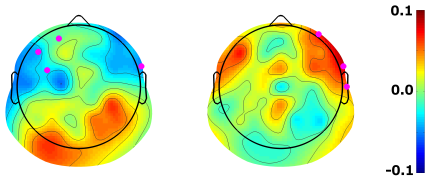
There are several neurophysiological indicators of cortical activation. For EEG, an increase in ERD_{α} or $ERD_{l-\beta}$ or a decrease in ERD_{γ} or $ERD_{h-\beta}$ (suggesting ERS in γ and $h-\beta$) have been shown to indicate cortical activation [56]. In fact, ERD_{α} is thought to be a prerequisite to evoke ERS_{γ}, and an inverse relationship between the two suggests cortical activation in the brain region under test [178]. Previous research with simultaneous MRI and EEG recordings has shown direct relationships between ERD_{α} and BOLD responses, as well as ERS_{γ} and BOLD responses from the

same regions of the brain [178]. Such insights shed light into the cortical regions activated during a specific experimental task. In the present case, this corresponded to a multi-attribute TTS QoE perception task.

Having this said, the cortical activity patterns depicted by Fig. 3.1 and the topographical correlation plots depicted by Fig. 3.2 indicate higher left and right fronto-temporal activation with increasing speech quality. The significant negative correlation for ERD_γ in left temporal and fronto-temporal regions, for different subjective dimensions, indicate activation of this region with increasing speech quality. Also, a parieto-occipital deactivation is evident from significant increase in ERD_γ and decrease in ERD_α with increasing speech quality.

The left and right temporal regions of the brain are considered important for speech perception [62]. The left temporal region is considered to be responsible for speech comprehension and has been found to have increased activation with increasing speech intelligibility [63]. Previous studies have reported an increase of ERD_α with increasing intelligibility of speech [179], specifically in the anterior temporal region [180], reflecting less effortful speech processing and more attentive cognitive processing of the speech signals. Also, in [181], an increase in ERS_γ was observed in the left temporal for speech perception tasks. These previous findings corroborate our observations regarding increasing ERD_α and decreasing ERD_γ with increasing speech quality, comprehension, and decreasing listening effort. These observations are also aligned with the findings in [84], where higher quality speech signals evoked higher P300 amplitude compared to lower quality signals reflecting more attentive processing of high quality signals. Also, in [30, 31], authors demonstrated increase in alpha band spectral power or decreased ERD_α for low quality multimedia signals, thus further corroborating our findings.

Moreover, the left temporal region has been proposed to be the area responsible for processing linguistic prosody [62] and temporal microstructure of the sound [64]. The right temporal region, in turn, has been associated with affective prosody [62], as well as pitch, direction of pitch perception, and spectral processing [64]. Since synthetic speech prosody (or intonation) is known to affect the perception of naturalness [65] and modulates the emotional state associated with the produced speech excerpt [66], right temporal region activation was expected for TTS systems of varying quality. In fact, the topographical correlation plots in Fig. 3.5 between $ERD_{h-\alpha}$ and the CP and Int dimensions show significant correlations in the left fronto-temporal and right fronto-temporal region,



(a) CP (b) Int

Figure 3.5 – Topographical correlation maps for $ERD_{h-\alpha}$ with comprehension problems (CP) and intonation (Int) dimensions.

respectively, thus corroborating the previously-reported asymmetry associated with intelligibility and prosody.

Another major observation from the correlation plots is the increased ERD_{γ} and decreased ERD_{α} over occipital and parieto-occipital regions (i.e., over the visual cortex). Such findings are indicative of a relative deactivation or top-down inhibition [159] of these areas due to participants focussing on speech stimuli in the absence of any visual stimulus. Similarly, fNIRS showed deactivation in the temporo-parietal region that has been associated with audio-visual integration [182, 183]. The task in the current study was not designed to evoke audio-visual integration, thus led to deactivation of the temporo-parietal regions. In summary, the present neural correlate study involving TTS systems concurred with previous studies exploring speech perception, thus validating the extracted features as correlates of synthesized speech QoE perception.

Furthermore, positively correlated α AI features and subjective dimensions, over the frontal regions of the scalp, corroborate results reported in [184, 185]. The positive correlation indicates higher α activity in the left frontal cortex as compared to the right frontal cortex, thus indicating higher activation of the left frontal cortical region with increasing signal quality. Moreover, a negative correlation between γ AI features and subjective dimensions, over the centro-parietal regions of the scalp, also indicate higher left centro-parietal activation with increasing signal quality (as the γ band ERD is directly proportional to cortical activation). This observation is in corroboration with the findings of [186], where the authors found an increased left cortical activation in the centro-parietal region with increasing valence of the stimuli.

The correlation analysis using graph features found significant correlations between thresholds of 0.2 and 0.6. This can be due to volume conduction effects for thresholds below 0.2 and disconnected graphs for thresholds above 0.6 [33]. Moreover, positive correlations for E_l, E_g, C_{mean} and S, and negative correlation for L indicate increase in local and global efficiencies with increasing signal

quality. However, a higher percentage of significantly correlated local efficiency features is indicative of increase in local information processing (in the context of brain networks) with increasing quality of TTS speech stimuli.

Affective Videos Stimuli: DEAP Database

The correlation analysis between affective dimensions and graph metrics revealed that L is inversely related to the subjective ratings, whereas E_g, E_l, C and S are all positively correlated. A decrease in L and an increase in E_g points towards an increase in sequential and parallel global information flow, thus leading to greater integration of information in brain connectomes [33]. On the other hand, an increase in C and E_l suggests an increase in efficiency of local information flow or segregation in brain connectomes.

In fact, salient stimuli are known to induce high arousal levels [67], thus leading to more integrated processing of information via the so-called 'workspace neurons,' as proposed by the global workspace theory [68]. Also, the increase in segregation for brain connectomes may be due to the fact that increased salience may lead to an increase in processing of cognitive states, such as attention, which have a top-down effect on the sensory information processing [187]. Combined, these two results lead to an overall increase in small-worldness properties with increasing arousal. The subjective arousal rating, specifically, showed positive correlations in higher frequency bands (e.g., $f \geq 18Hz$) using MSC, along with moderate correlations for $f < 18Hz$. The higher involvement of faster rhythms in high arousal states has been observed in previous studies [188, 189].

We found a significant increase in local properties of the brain networks in the mid-beta frequency range with increasing valence. Various studies have shown that the beta band (and its sub-bands) encodes affect-related information. For example, in [190] a decreased intra-hemispheric left coherence in the low beta band with negative affect was reported and in [191] an asymmetric activation of the beta band while attending to affective visual stimuli was shown. Thus, we can state that the graph features encode meaningful affect-related information as the observed correlations partially concur with some of the previous studies, and can be used as valid features for affective state recognition.

3.5.2 Classification

TTS Stimuli: PhySyQX Database

The feasibility of characterizing several QoE percepts using EEG signals is evident from the classification results presented in Table 5.1, where results significantly better than chance were obtained. The decision fusion of feature sets within EEG were shown to lead to significant improvements, thus suggesting that decision fusion from different sources of information from within EEG improves the characterization of QoE percepts. Furthermore, it was observed that the performance of features derived from the power spectrum (e.g., ERD, AI) and cross spectrum (e.g., Graph features) did not differ significantly. Also, it was noticed that the classifiers based on the fusion of individual bands of ERD and AI performed better than classifiers based on graph features derived from individual bands. However, this observation was not significant. Moreover, the performance of decision fusion classifiers based on fusion of ERD and AI, and decision fusion classifiers based on graph features did not differ much. This indicates that power spectrum features and cross spectrum features encoded equivalent QoE-related information.

Affective Videos Stimuli: DEAP Database

From Tables 3.2, it can be observed that graph, ERD and AI features perform significantly better than chance in classifying users' emotional states, thus suggesting their utility in affect classification. Also, it can be said that the classifiers developed using the graph theoretical treatment of EEG data produced significantly ($p < 0.05$) better classification metrics for emotional dimensions, specifically for valence and arousal, as compared to the traditionally used ERD and asymmetry features. This shows the utility of the proposed graph features in characterizing users' emotional states more accurately than previously used EEG features. Moreover, the improved performance seen for the valence classification problem at higher frequency bands ($\geq$l-β) could be explained by the higher long and short distance coherence in such bands in response to higher valence, as observed in [168]. This finding is further corroborated by the fact that graph features, encoding local and global information transfer, were significantly higher for frequencies above the l-β band ($> 12Hz$), as shown in Fig. 3.3. However, classification performance for arousal did not differ significantly between different frequency sub-bands. This could be attributed to significantly different local and

global graph metrics for high or low arousal-inducing stimuli at both low and high frequency sub-bands, as shown in Fig. 3.4. Also, classification performance for any of the affective dimensions did not differ significantly for traditional features computed from different frequency sub-bands. This can be attributed to the relevance of each frequency sub-band in affect classification as reported in [28], where the authors show significant correlations between different sub-band powers and affective dimensions. It can also be observed that the RVM_2 performed better than RVM_1. This could be attributed to the fact that the emotion classification problems were unbalanced, hence searching for the optimal weights for classes improved the classification performance. Also, RVM_2 performed better than SVM for all classification problems, likely owing to their probabilistic nature.

The classifiers developed using decision fusion of graph metrics-based classifiers performed significantly better than the ones that used ERD or AI features, particularly for valence. However, for arousal classification, decision fusion of the graph theoretic metrics was only significantly better than ERD features. These results further strengthen the utility of graph features in solving the emotion classification problem. Also, in this study we observed that the decision-level fusion (Table 3.3) performed at par with the graph metrics-based standard (individual) classifiers, as reported in Table 3.2. This observation is in corroboration with previous studies, where the decision fusion scheme provides the best classification performance [28]. However, comparison between standard and decision fusion classification techniques did not reveal significant difference between the techniques.

Moreover, these results showed higher performance than the results reported using traditional EEG features (such as ERD and AI), peripheral, multimedia content analysis or the fusion of these features, in [28], or DT-CWPT features, in [175]. Specifically, the % increase in F1-scores, using decision fusion as compared to best previously reported results in [28], are as follows: Valence (11%) and Arousal (10%), whereas the gains over conventional features, found from Tables 1-3, were: Valence ($7-9\%$) and Arousal ($3-8\%$). This provides sufficient evidence of the superiority of the proposed methods for the task at hand.

Furthermore, graph-theoretic features were observed to outperform power spectrum-based features (e.g., ERD and AI), in classifying affective dimensions, for affective videos whereas, for TTS stimuli, graph-theoretic features and power spectrum-based features showed equivalent performance. This observation can be attributed to the difference in the nature of stimuli that induced changes

in affective states. The DEAP database used affective music videos as stimuli that involve visual as well as audio information processing, whereas the PhySyQX database used TTS as stimuli that involve only auditory processing. This is also corroborated by the fact that mostly segregated information processing metrics were correlated with QoE (and affective) dimensions for TTS stimuli, whereas for video stimuli, both segregated and integrated information processing metrics were correlated with affective dimensions. Furthermore, as reported in [28], the video stimuli spanned all four quadrants of the valence-arousal scale (i.e., HAHV, HALV, LALV and LAHV); however, the TTS stimuli spanned just the HAHV and LALV quadrants. This can lead to different dynamics of cortical information processing.

3.6 Conclusions

In this chapter, we have taken first steps towards using EEG-based BCIs for characterising the human factors that influence the QoE of long-duration multimedia signals. Towards this end, we have explored the use of EEG-based features, such as event-related desynchronization and graph-theoretical features, for characterising HIFs. Several such metrics were shown to reliably discriminate between low and high levels of subjective HIFs. These findings suggest that objective affective characterisation is possible using the EEG-based BCIs. This also indicates that hybrid BCIs that involve EEG can be useful for objective QoE assessment, which will be explored in later chapters.

Chapter 4

Characterization of HIFs Using fNIRS-based BCIs

4.1 Preamble

This chapter is compiled from material extracted from manuscripts published in the Proceedings of 2013 Workshop on Perceptual Quality of Systems [47] and a manuscript that is under review in the IEEE Journal for Selected Topics in Signal Processing: Special Issue on Measuring Quality of Experience for Advanced Media Technologies and Services [46].

4.2 Introduction

Recently, fNIRS has emerged as an alternative neuroimaging modality providing complementary information to EEG for studying long-duration multimedia signals [36]. fNIRS provides good spatial resolution, thus overcoming a major limitation of EEG. Such qualities of fNIRS indicate that using fNIRS-based BCIs could be effective for objective QoE assessment. In general, fNIRS-based features capture cortical activation and are generally better suited for observing long-term users' states than an instantaneous change in their states [36]. Towards describing long-term users' states using fNIRS, various features have been investigated recently. One of the most common features is the temporal dynamics of the $\Delta[HbO]/\Delta[HbR]$ amplitude itself. For example, [36] used it to characterize the

users' preference of a movie based on its likeability. In [37], the authors have described the laterality (such as lateral absolute mean difference between two brain hemispheres, lateral slope ratio) and single channel-based features (e.g., stimuli period mean, stimuli period slope) to characterize users' emotional states. Various other features, along with their applications in the user experience (UX) domain (which is closely related to QoE [123]), have been studied and are listed in [124]. However, fNIRS as a technique is in its infancy in the QoE assessment domain and could prove to be very effective alongside EEG, in developing better objective models to characterize QoE. Towards this end, we have explored alternative fNIRS derived features for the characterization of HIFs. This was achieved by probing the prefrontal cortex and the full head in the preliminary study and the PhySyQX database, respectively, which are described in this chapter.

The remainder of this chapter is organised as follows: Section 4.2 and 4.3 describe the methodology, results and discussion for preliminary and main study, respectively. Finally, in section 4.4 conclusions are drawn.

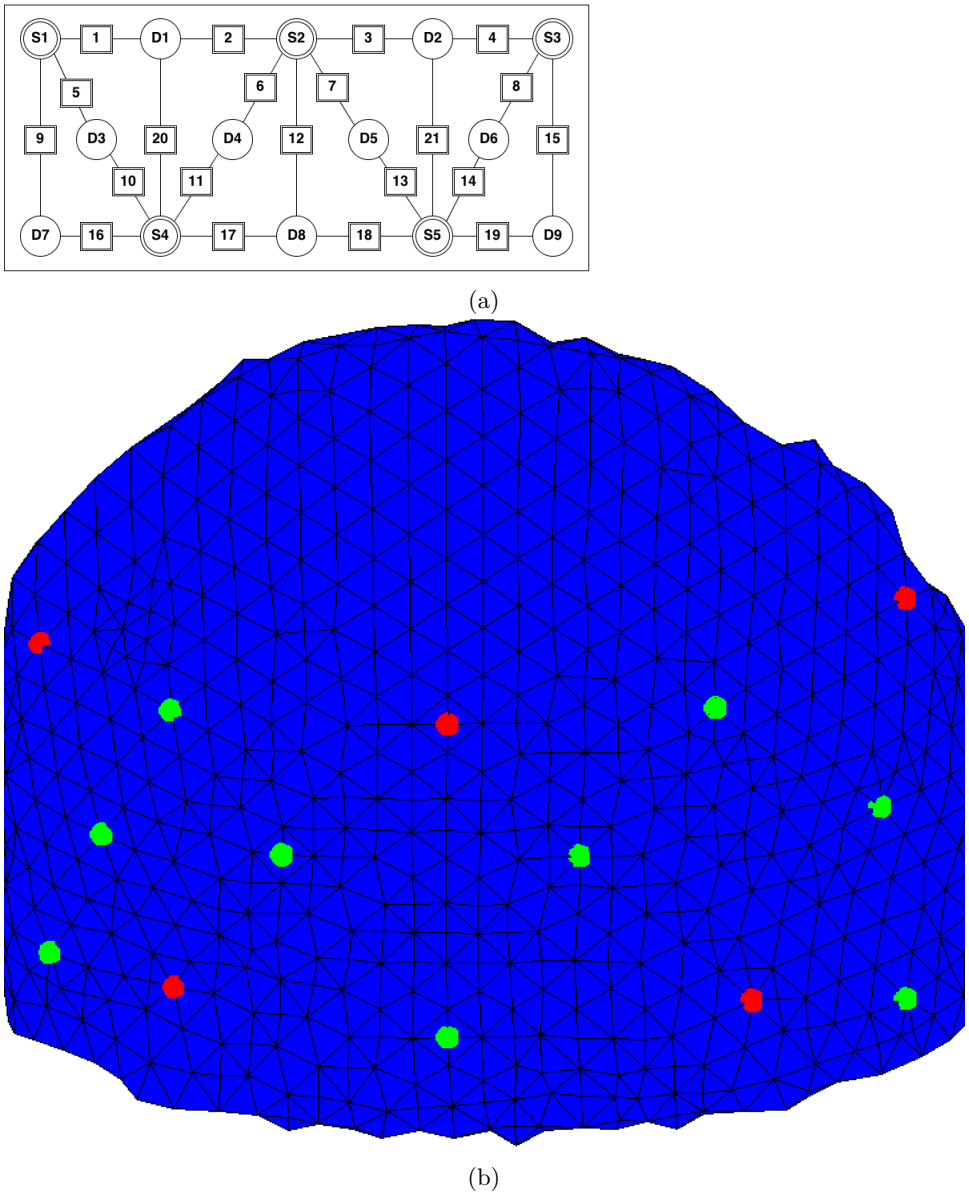
4.3 Prefrontal Cortex Probing - Preliminary Study

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 [192] and decision making [72]. More distinctly, a region of the PFC called orbito-frontal cortex (OFC) has been found to be activated during decision making tasks [69]. According to the so-called neuro-economics literature, the OFC is also responsible for the valuation and outcome evaluation processes involved in decision making [70, 193, 71, 194, 195]. Such findings have motivated our research of probing the PFC/OFC regions to obtain insights into the human speech quality perception processes.

4.3.1 Materials and Methods

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



(a)

(b)

Figure 4.1 – 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.

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.

Synthesized speech stimuli

In order to utilize synthesized speech stimuli representative of existing systems, data from the 2009 Blizzard TTS Challenge were used [50]. 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 that 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.

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.

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 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. 4.1a and 4.1b, 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 [196]. 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) [197].

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. 1.7. 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 [198, 122, 199]. Moreover, to account for the variation in these features, the coefficient of variation, which is the ratio between the standard deviation and the mean for a particular feature, was calculated.

4.3.2 Results

Subjective Data Analysis

As shown in Fig. 4.2, 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 4.1, 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.

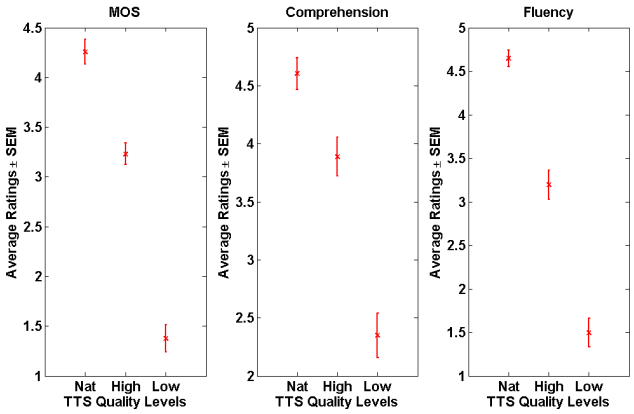


Figure 4.2 – Summary of obtained subjective ratings.

Table 4.1 – 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

Physiological Data Analysis

As the literature suggested, increased activation of the OFC was expected with increasing perceptual quality of the stimuli [70]. This effect is clearly visible in Fig. 4.3, 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. 4.4, 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 ($p = 0.267$). 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 4.2. The 'η^2' was maximum (0.44) for the $\Delta[HbO]$ peak for channel 14 located on the right hemisphere of the OFC,

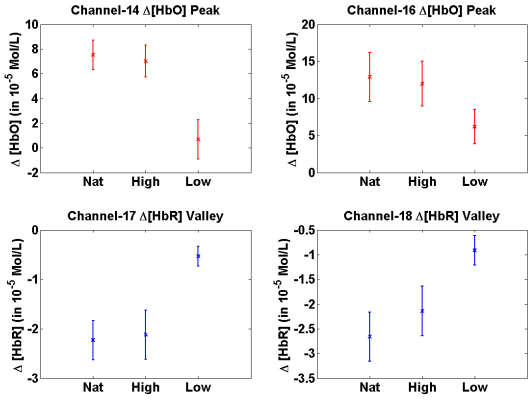


Figure 4.3 – Physiological features: Mean± Standard Error of Mean (SEM).

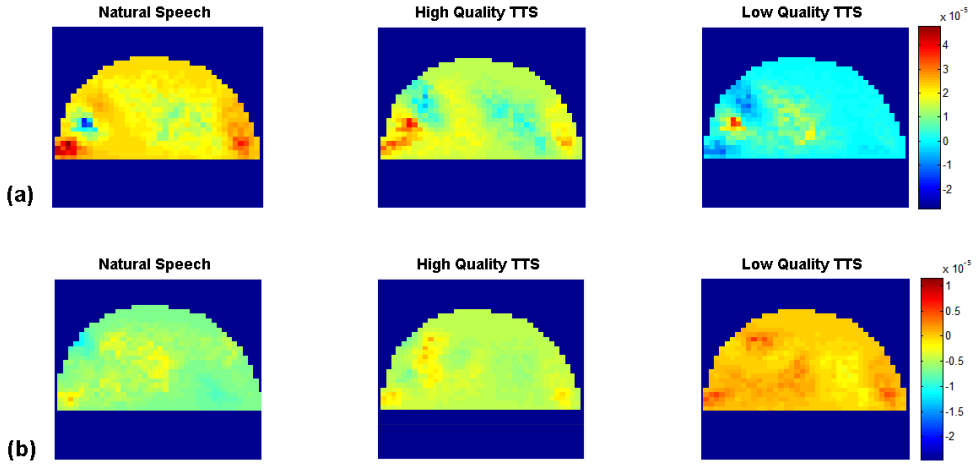


Figure 4.4 – fNIRS-based reconstructed image of the Prefrontal Cortex (Coronal View) for: (a)$\Delta[HbO]$ Peak and (b)$\Delta[HbR]$ Valley

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 4.2 – 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 4.3. This confirmed the linear relationship between the fNIRS features and subjective quality ratings of the stimuli. Also, the centrality of these features

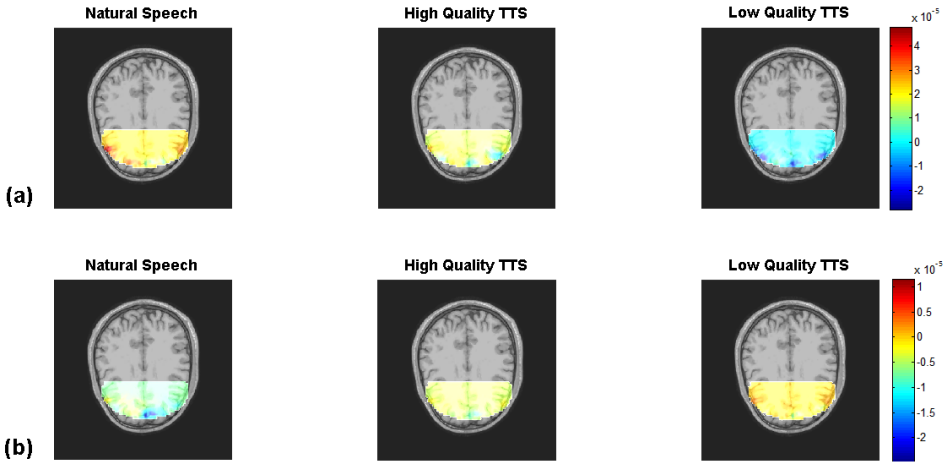


Figure 4.5 – fNIRS-based reconstructed image overlaid on MRI scan of the Prefrontal Cortex (Sagittal View) for: (a) $\Delta[HbO]$ Peak and (b) $\Delta[HbR]$ Valley

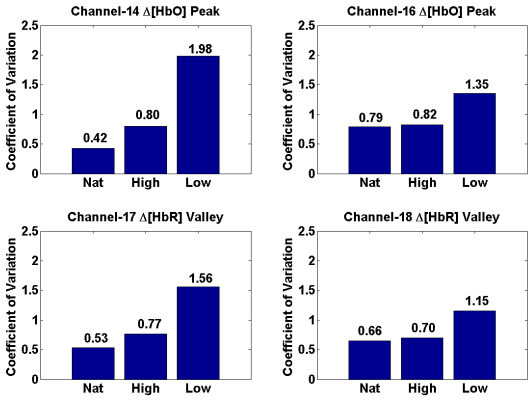


Figure 4.6 – Coefficient of Variation for the Physiological Features.

was determined by measuring their Mean ± its Standard Error of Mean (SEM) as shown in Fig. 4.3, 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. 4.6. In order to test the significance of this trend in the coefficient of variation, Levene's test was carried out [200]. 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 4.3 – 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

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. The $\Delta[HbO]$ 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 4.4.

Table 4.4 – fNIRS Correlates of Subjective Quality Metrics.

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

4.3.3 Discussion

It is quite intuitive that, as the quality of an 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 [71]. 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 [70], 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 [71]. However, there are indications of existence of deeper located neural systems for valuation [201]. 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.

4.4 Full Head Probing - the PhySyQX Database

In the previous section, we have shown the importance of fNIRS and cortical haemodynamics for TTS QoE measurement, particularly in the pre-frontal cortex (PFC) region of the brain, which has been proven to be associated with cognition and decision making [72]. Here, we expand that work by probing the entire head in an attempt to decode more complex cortical interactions involved in the QoE perception process. Moreover, fNIRS probing simultaneously with EEG along the same cortical regions can provide validation to the observations from EEG. Furthermore, physiological measures (e.g., heart rate) have also been shown to be useful in affective state monitoring, particularly within a multimodal setup [28, 29]. Here, we attempt to extract heart rate-based features, from fNIRS signals, to characterize HIFs using the the PhySyQX database described in Chapter 2.

4.4.1 Methodology

Pre-processing

Raw fNIRS signals comprised data from 384 channels (16 sources $\times$ 24 detectors) from two wavelengths: 760 nm and 850 nm. The raw fNIRS signals were pre-processed using the nirsLAB toolbox [202]. Functional channels were determined from the 384 raw channels by analyzing the inter-optode (source-detector) distances and keeping only the ones with distances of approximately 3 cm, which have been shown to be optimal for cortical activation characterization [203]. As such, 60 channels were kept and deemed functional for analysis (see Fig. 2.2). Next, signals from the 60 functional channels were band-pass filtered between 0.005-0.1 Hz using a third order Butterworth filter, to remove the noise due to physiological processes, such as heart beats and respiration. However, it should be noted that this filtering step was carried out to extract fNIRS-based features and not heart rate-based features. Next, 25 seconds-long epochs comprised of speech stimuli duration plus 10 seconds of pre-stimulus baselines were extracted from each of the 44 speech stimuli, per participant. Finally, these signals were used to measure $\Delta[HbO]$ and $\Delta[HbR]$ concentrations using the MBLL [34].

Feature extraction

Motivated by the promising results reported in [47] for TTS QoE assessment and those in [124], here we explore the use of several features extracted from all 60 functional fNIRS channels, which were simultaneously recorded with EEG. First, we computed the peaks, valleys, rise amplitude, decrease amplitude, and zero crossing rate computed from the $\Delta[HbO]$ and $\Delta[HbR]$ curves, as shown in Fig. 1.7. This resulted in 300 features (60 channels $\times$ 5 features) for each of the $\Delta[HbO]$ and $\Delta[HbR]$ curves. Henceforth, these two feature sets will be referred to as HbO_{temp} and HbR_{temp}, respectively. Moreover, we also computed statistical measures, such as mean, median, standard deviation, skewness and kurtosis, over four 5-second non-overlapping windows (lasting 20 seconds post stimulus) of $\Delta[HbO]$ and $\Delta[HbR]$ curves from all 60 channels. This resulted in 1200 features (60 channels $\times$ 5 features $\times$ 4 windows) for the $\Delta[HbO]$ and $\Delta[HbR]$ curves. Henceforth, these features are termed HbO_{stats} and HbR_{stats}, respectively.

fNIRS-derived heart rate extraction

As mentioned previously, raw fNIRS signals are corrupted by a cardiac “noise" if the acquired signal is sampled at a rate $\geq$ 3.5 Hz (which is twice the frequency of an average heart rate) [35]. Typically, this interference is filtered during the pre-processing stage, as mentioned previously. Heart rate and and heart rate variability (HRV) parameters, however, have proven useful in characterizing human affective states evoked in response to multimedia content [28, 29]. Therefore, we propose to extract heart rate information from the functional fNIRS channel that offered the best possible heart rate signal quality for each subject. In fact, previously, fNIRS has been used to extract infant heart rate [204]. Here, the quality of the heart rate signal was determined using the criterion explored in [205], which detected a heart rate if a peak was identified in the log-power spectrum ($p(f)$) obtained using the Welch's method for the frequency band f between 0.8 and 1.7 Hz, i.e.,

$$max\{p(f)|f \in [0.8, 1.7]\} - mean\{p(f)|f \in [0.8, 1.7]\} > 0.5. \tag{4.1}$$

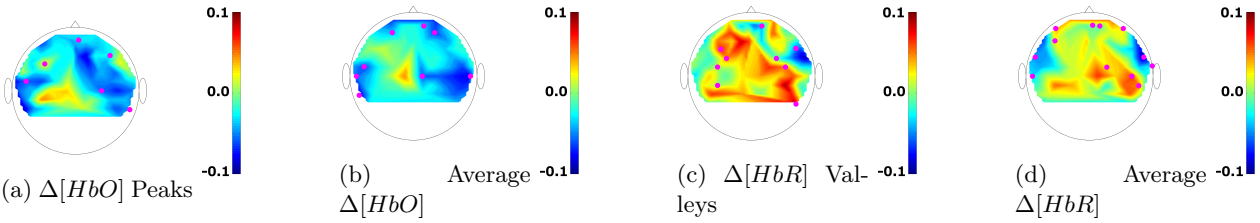
Once a good quality signal was found, it was band-pass filtered between 0.05-2 Hz using a third order Butterworth filter to create a heart rate and HRV time series [73]. From these two time series, two feature sets were extracted. The first set, henceforth called ‘HRF1’, was computed according

to [28] and consisted of 1) statistical measures (that included mean, median, standard deviation, skewness, kurtosis, minimum and maximum) of HR and HRV time series, 2) spectral power bands, extracted from HRV time series, for the frequency ranges between 0.04-0.15, 0.15-0.40, 0.15-0.25, and 0.25-0.35 Hz, and 3) the energy ratio between the 0.04-0.15 and 0.15-0.40 Hz frequency bands, as well as the 0.15-0.25 and 0.25-0.35 Hz bands. Overall, a total of 20 features are available in HRF1. The second feature set, henceforth called 'HRF2', was computed according to [73] and consisted of 1) average inter-beat-intervals (IBI), 2) standard deviation of all normal to normal (NN) intervals or SDNN, 3) standard deviation and the absolute value of the first derivative of IBI, 4) the number of successive NN intervals differing by more than 50ms (or NN50), 5) the ratio of NN50 to NN, 6) the standard Poincaré descriptors (SD1 and SD2) and their ratio (SD1/SD2), and 7) the sample entropy of the HRV time series, as proposed in [206]. Overall, a total of 9 features is available in HRF2.

4.4.2 Results

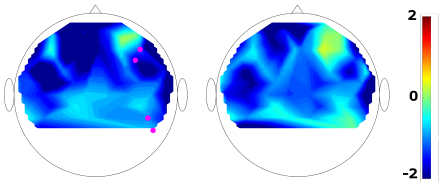
Neural Correlates

From Fig. 4.7 it was observed that $\Delta[HbO]$ peaks showed significant positive correlations in the left and right fronto-temporal and temporo-central regions of the brain, with increasing signal quality. Moreover, $\Delta[HbO]$ peaks showed significant negative correlations in the left and right parieto-temporal and mid-frontal regions of the brain. This was observed consistently over different subjective dimensions. Furthermore, $\Delta[HbO]$ average concentration, during stimulus presentation, showed similar correlation patterns as observed for $\Delta[HbO]$ peaks. The $\Delta[HbR]$ average concentration showed significant negative correlations, in the left and right fronto-temporal regions with increasing signal quality, consistently for all subjective dimensions. Also, $\Delta[HbR]$ average concentration showed significant positive correlations for some subjective dimensions in the central and temporo-central regions. The correlation patterns for $\Delta[HbR]$ valleys closely followed the $\Delta[HbR]$ average concentration patterns. These results were corroborated by the significant differences in the $\Delta[HbR]$ response towards high and low quality systems, as shown in Fig. 4.8. However, no significant channels were found for $\Delta[HbO]$ and therefore, the plots for $\Delta[HbO]$ are not shown.



(a) $\Delta[HbO]$ Peaks (b) Average $\Delta[HbO]$ (c) $\Delta[HbR]$ Valleys (d) Average $\Delta[HbR]$

Figure 4.7 – Topographical maps of the average correlation between different fNIRS features and the MOS rating. Channels with significant correlations are highlighted with a magenta coloured '•' symbol.



(a) HQ-Δ[HbR] (b) LQ-Δ[HbR]

Figure 4.8 – Average Δ[HbR] for high ($HQ = MOS \geq 3$) and low ($LQ = MOS < 3$) quality systems. The channels with significant ($p < 0.05$) differences between HQ and LQ are highlighted with a magenta coloured '$*$' symbol.

Table 4.5 – Average accuracies (Acc) and F1-scores (F1) over participants, for each fNIRS-based feature set. Superscripted bullets (•) indicate results that are not significantly higher than chance according to an independent one-sample t-test (• = $p > 0.05$). Superscripted asterisks ($*$) indicate the decision fusion-based classifiers that perform significantly better than the best performing classifiers based on individual EEG sub-bands ($* = p < 0.05$). The penultimate row benchmarks the system based on a random voting classifier. The last row presents the mean and standard deviation of the percentage of positive class labels across subjects.

Modality	Feature	Metric	MOS	VP	Ac	Int	Nat	LE	VE	Val	Ar	CP
fNIRS	Individual	Acc	0.60	0.55	0.59	0.55	0.60	0.60	$0.53^{\bullet}$	0.55	$0.53^{\bullet}$	0.64
		F1	0.60	0.55	0.59	0.55	0.60	0.60	$0.53^{\bullet}$	0.55	$0.53^{\bullet}$	0.64
		Feat	HbR_t	HbO_s	HbO_t	HbO_s	HbO_t	HbR_s	HbR_t	HbO_s	HbR_s	HbR_s
	Fusion	Acc	0.63	0.58	0.64^{*}	0.62^{*}	0.65^{*}	0.62	0.60^{*}	0.56	$0.55^{\bullet}$	0.69
		F1	0.62	0.57	0.63	0.60	0.64	0.61	0.59	0.56	$0.55^{\bullet}$	0.67
HR	Individual	Acc	$0.55^{\bullet}$	0.57	0.56	$0.55^{\bullet}$	0.56	0.56	$0.55^{\bullet}$	0.56	$0.52^{\bullet}$	0.59
		F1	$0.55^{\bullet}$	0.57	0.56	$0.55^{\bullet}$	0.56	0.56	$0.55^{\bullet}$	0.56	$0.52^{\bullet}$	0.59
		Feat	HRF_1	HRF_1	HRF_1	HRF_1	HRF_1	HRF_1	HRF_1	HRF_1	HRF_2	HRF_2
	Fusion	Acc	0.58^{*}	0.58	0.59^{*}	0.58	0.59^{*}	0.57	0.56	0.58	$0.54^{\bullet}$	0.60
		F1	0.58	0.58	0.59	0.58	0.59	0.57	0.56	0.58	$0.53^{\bullet}$	0.61
Random Voting	-	Acc	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50
		F1	0.50	0.50	0.50	0.50	0.51	0.51	0.50	0.50	0.50	0.52
+ve Class Percentage	-	Mean	0.50	0.46	0.40	0.49	0.39	0.44	0.44	0.47	0.49	0.35
		St. Dev.	0.13	0.13	0.13	0.14	0.13	0.15	0.12	0.13	0.11	0.19

Classification

Table 4.5 reports average accuracies and F1-scores over all participants for each feature set for fNIRS, for each subjective dimension. The Table indicates the best performing classifiers from individual fNIRS features along with the decision fusion of each classifier developed using individual

fNIRS features. The performance for each classifier was tested using a two-sided repeated samples t-test over the concatenated results from each rating scale and participant, as suggested by [28]. It was observed that most of the classifiers performed significantly above chance; however, the classifiers that did not result in significantly above chance performance are indicated by a superscripted bullet. Moreover, it was observed that the decision fusion-based classifiers performed better than classifiers based on best performing individual fNIRS and HR features. However, this observation was significant (indicated using superscripted '$*$' in Table 4.5) for Ac ($t(20) = 2.18, p < 0.05$), Int ($t(20) = 3.24, p < 0.05$), Nat ($t(20) = 2.17, p < 0.05$) and VE ($t(20) = 2.28, p < 0.05$) dimensions using fusion of fNIRS feature sets; and the MOS ($t(20) = 2.15, p < 0.05$), Ac ($t(20) = 2.16, p < 0.05$) and Nat ($t(20) = 2.18, p < 0.05$) dimensions for fNIRS-derived heart rate feature sets, as tested using a paired t-test. Furthermore, comparing performances of individual fNIRS and HR features, it was observed that fNIRS-based features performed better than HR; however this finding did not reveal any significant differences. Also, comparing the fusion classifiers developed using individual fNIRS and HR features, it was observed that the fNIRS-based fusion classifiers performed better than HR and this finding was significant for the CP dimension. For comparison, the table also reports the average performance of a random classifier developed using random voting, as well as the average percentage and standard deviation of positive class labels ('label 1') to quantify class imbalance.

4.4.3 Discussion

In this section we discuss the results obtained from the study and their potential future role in QoE perception modelling.

4.4.4 Neural correlates of QoE perception

There are several neurophysiological indicators of cortical activation. For fNIRS, an increase in average $\Delta[HbO]$ concentration and $\Delta[HbO]$ peaks along with a decrease in average $\Delta[HbR]$ concentration and $\Delta[HbR]$ valley reflect increases in cortical activation [199, 122]. Such insights shed light into the cortical regions activated during a specific experimental task. In the present case, this corresponded to a multi-attribute TTS QoE perception task.

Having this said, the cortical activation maps shown in Fig. 4.8 and the topographical correlation plots depicted by Fig. 4.7 indicate higher left and right fronto-temporal activation with increasing speech quality; such findings were consistent with the findings from EEG, as reported in Section 3.4.1. However, a parieto-occipital deactivation is evident from significant increase in ERD_γ and decrease in ERD_α with increasing speech quality. This observation could not be corroborated with fNIRS, as this region of the cortex was not probed using fNIRS optodes. Furthermore, a left and right deactivation of temporo-parietal areas is evident from fNIRS. However, an increase in ERD_γ and decrease in ERD_α only corroborate deactivation of right temporo-parietal areas with increasing speech quality.

Furthermore, fNIRS showed deactivation in the temporo-parietal region that has been associated with audio-visual integration [182, 183]. The task in the current study was not designed to evoke audio-visual integration, and thus led to deactivation of the temporo-parietal regions. In summary, the present neural correlate study involving TTS systems concurred with previous studies exploring speech perception, thus validating the extracted features as correlates of synthesized speech QoE perception.

4.4.5 Classification

The feasibility of characterizing several QoE percepts using physiological signals is evident from the classification results presented in Table 4.5, where results significantly better than chance were obtained. Interestingly, fNIRS-derived heart rate features also resulted in above-chance performance, thus indicating that so-called physiological noise from fNIRS signals can be used for QoE perception modelling. Moreover, fusion of feature sets within a specific modality were shown to lead to significant improvements for fNIRS and HR, thus suggesting that decision fusion from different sources of information from within a modality improves the characterization of QoE percepts. Furthermore, fNIRS-based classifiers were observed to outperform HR-based classifiers, thus suggesting that fNIRS signals provide richer information regarding QoE dimensions. Also, comparing the performance of EEG and fNIRS-based features, there was no significant difference between the two modalities. This indicates both modalities provide equal amounts of information for the characterization of HIFs.

4.5 Conclusions

We have successfully established the use of fNIRS-based BCIs to obtain insights into the neural processes involved in TTS system QoE perception. The findings of our studies point towards significant correlations between the fNIRS features with HIFs. In our preliminary study, 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, such as temporal, central and frontal, were also investigated in the PhySyQX database, simultaneously with EEG. This simultaneous probing validated the activation of cortical regions, in response to TTS stimuli, detected using EEG. Furthermore, we successfully extracted the heart rate from fNIRS signals and leveraged it for characterising HIFs. It was observed that fNIRS-based features outperformed HR-based features while classifying HIFs. However, fusing information from different neurophysiological modalities, to develop a truly hybrid BCI, and comparing their performance against the individual modality based BCIs will be explored in the next chapter.

Chapter 5

Characterization of HIFs using Multimodal Fusion

5.1 Preamble

This chapter is compiled from material extracted from a manuscript that is under review in the IEEE Journal for Selected Topics in Signal Processing: Special Issue on Measuring Quality of Experience for Advanced Media Technologies and Services [46].

5.2 Introduction

QoE perceptual processes are complex and are not directly observable and render the task of modelling QoE challenging. This is recognized and represented by an increasing interest in BCI-based methods, such as EEG or fNIRS, to investigate neural processes and HIFs involved in QoE perception [47, 84]. However, in the field of human factors research, specifically in the domain of affective computing, researchers have explored fusion of measurements from different sensor modalities through various pattern recognition methods [74]. It is noted in [75] that a multimodal approach towards affect recognition leads to more accurate results. Therefore, towards characterising HIFs more accurately, we have explored fusing information from different neurophysiological modalities to develop hybrid BCIs.

The general approaches for modality fusion can be categorised into two classes, namely, feature fusion and decision fusion [207]. However, there have been attempts to combine both fusion methodologies using hybrid fusion methods [208]. The feature fusion technique involves concatenation of feature vectors from each modality to form a composite feature vector, which is then used to train a classifier. However, in the decision fusion, features from each modality are used to develop individual classifiers and the outputs of the classifiers are combined to obtain the final result. In general, feature fusion considers synchronous characteristics of the modalities, whereas decision fusion considers the asynchronous characteristics of the modalities [76]. Previous research has attempted to fuse decisions from various modalities, such as EEG, face, eye gaze and peripheral physiological signals [28, 29], towards affect recognition.

In a similar vein, we attempted to develop accurate objective measures of HIFs by exploring the decision fusion of multiple signal modalities from the PhySyQX database. More specifically, EEG, fNIRS and fNIRS-derived heart rate measures are combined with decision tree classifiers to measure several subjective QoE dimensions, namely: overall impression, voice pleasantness, acceptance, comprehension problems, intonation, naturalness, listening effort, emotions, valence and arousal. It is hypothesised that significant classification accuracy can be achieved with EEG, fNIRS, and fNIRS-derived heart rate features. Such findings corroborate the importance of HIFs in overall objective QoE assessment of TTS systems.

5.3 Methodology

The fusion of classifier decisions obtained from the different modalities (i.e., EEG, fNIRS and fNIRS-derived heart rate) was implemented using the the weighted decision fusion scheme proposed in [76]. According to this technique, the fusion classification probability $p_0^x \in [0, 1]$ for each class $x \in \{1, 2\}$ can be denoted by

$$p_0^x = \sum_{i=1}^{N} \alpha_i p_i^x t_i, \tag{5.1}$$

where i is the index of a particular modality used for assessment, N is the number of modalities used, and α_i are the weights corresponding to each modality ($\sum_{i=1}^{N} \alpha_i = 1$). The parameter t_i is the normalized training set performance for a particular modality, such that the fusion probabilities for

all classes sum up to unity [76], and is given by:

$$t_i = \frac{F_i}{\sum_{i=1}^{N} \alpha_i F_i}, \tag{5.2}$$

where F_i is the F1-score obtained on the training set using a particular modality and $F_i \in [0, 1]$.

Using this formulation, two decision fusion strategies can be implemented: (i) equally weighted or (ii) optimally weighted decision fusion. The equally weighted decision fusion assigns equal weights ($\alpha_i = 1/N$) to each modality. Optimally weighted decision fusion, on the other hand, relies on optimal weights for each modality, which are found by searching for the α_i values that result in the best performance in a validation set. The latter typically relies on larger available datasets. However, both fusion strategies were implemented to develop a multimodal classifier.

The fusion of different modalities was carried out in two phases. In the first phase, all feature sets from each modality were used for fusion, while equally weighting each feature. However, for the second phase of decision fusion, only the best performing feature sets (on the training set) were used from each modality.

5.4 Results

Table 5.1 provides classification performance levels for decision-level fusion of *all* feature sets available for a given modality and different modality combination strategies. It was observed that the combination of different physiological modalities increased the performance of individual modalities for all subjective dimensions, except Ar, as compared to classifiers developed using best performing features from individual modalities. However, ANOVA (as there were more than 2 modalities to evaluate) revealed significant main effect only for combination 3 (fNIRS and HR) and 4 (EEG, fNIRS and HR) with Nat ($F(2, 60) = 4.23, p < 0.05$) and MOS ($F(2, 60) = 2.76, p < 0.05$), respectively. The post-hoc tests revealed significant differences between classifiers based on heart rate features and decision fusion. Lastly, comparing the classifiers developed using fusion of modalities (reported in Table 5.1), and fusion of features of each modality (reported in Table 3.1 and Table 4.5) resulted in no significant difference.

Table 5.1 – Average accuracies (Acc) and F1-scores (F1) over participants for fusion of features sets from different modalities.

Modality	Metric	MOS	VP	Ac	Int	Nat	LE	VE	Val	Ar	CP
EEG, fNIRS	Acc	0.64	0.64	0.64	0.62	0.64	0.63	0.59	0.60	0.57	0.71
	F1	0.59	0.60	0.58	0.56	0.58	0.57	0.54$^{\bullet}$	0.55	0.52$^{\bullet}$	0.64
EEG, HR	Acc	0.63	0.64	0.64	0.60	0.63	0.64	0.56	0.59	0.54$^{\bullet}$	0.70
	F1	0.58	0.60	0.58	0.54$^{\bullet}$	0.57	0.58	0.52$^{\bullet}$	0.55$^{\bullet}$	0.50$^{\bullet}$	0.63
fNIRS, HR	Acc	0.62	0.58	0.65	0.62	0.65	0.63	0.60	0.60	0.55$^{\bullet}$	0.67
	F1	0.61	0.57	0.63	0.60	0.63	0.61	0.58	0.58	0.53$^{\bullet}$	0.65
EEG, fNIRS, HR	Acc	0.64	0.64	0.65	0.63	0.64	0.63	0.59	0.60	0.57	0.70
	F1	0.59	0.59	0.59	0.56	0.58	0.58	0.54	0.55	0.52$^{\bullet}$	0.63

Table 5.2 – Average accuracies (Acc) and F1-scores (F1) over participants for fusion of best performing feature sets within each modality.

Modality	Metric	MOS	VP	Ac	Int	Nat	LE	VE	Val	Ar	CP
EEG, fNIRS	Acc	0.63	0.66	0.68	0.62	0.62	0.62	0.59	0.61	0.63	0.68
	F1	0.63	0.66	0.67	0.61	0.61	0.61	0.58	0.60	0.62	0.66
EEG, HR	Acc	0.63	0.65	0.67	0.64	0.63	0.65	0.61	0.62	0.63	0.69
	F1	0.63	0.65	0.66	0.63	0.62	0.64	0.61	0.61	0.63	0.67
fNIRS, HR	Acc	0.60	0.64	0.62	0.60	0.60	0.58	0.60	0.61	0.61	0.67
	F1	0.60	0.64	0.62	0.60	0.59	0.58	0.60	0.61	0.60	0.67
EEG, fNIRS, HR	Acc	0.61	0.65	0.65	0.60	0.59	0.62	0.58	0.59	0.59	0.66
	F1	0.59	0.63	0.64	0.58	0.57	0.60	0.57	0.57	0.57	0.62
ANOVA	F-stat	0.51	1.72	2.50*	1.90	1.03	0.78	1.53	1.25	6.46*	0.32

Finally, Table 5.2 reports the classification performance for combinations of fusion of best performing (on the training set) feature sets from each physiological modality. As can be seen, this fusion strategy improved classification performance relative to using all feature sets from each modality for decision fusion (i.e., as reported in Table 5.1). Comparing the performance of the classifiers reported in Table 5.2 with those in Table 5.1, an ANOVA was conducted, as reported in the last row of Table 5.2. As evident, the comparison was significant for Ac and Ar subjective dimensions. Also, post hoc tests revealed significant differences between combination 1, from Table 5.2; combination 1-4, from Table 5.1 for Ac; combinations 1-2 from Table 5.2, and combination 1-4 from Table 5.1 for Ar. Furthermore, it was observed that the optimal weights decision fusion of EEG, fNIRS and HR modalities did not show any significant improvements to the classification performance.

5.5 Discussion

The hybrid BCI approach lead to significant improvements as compared to using single modality based BCIs, thus suggesting that decision fusion from different sources of information improves the characterization of QoE percepts. This also highlights the need for multimodal physiological monitoring for QoE perception modelling. Overall, combination of EEG with fNIRS or HR resulted in better performing classifiers compared to classifiers developed by combining fNIRS and HR, however this observation was not significant. Also, the optimal fusion of three modalities did not result in significant improvement which could be attributed to the limited size of the PhySyX dataset. Finally, fusing the best performing feature sets, from each modality, lead to significant improvements in classifier performance for Ac and Ar dimensions, thus emphasizing the need to incorporate optimal feature sets, from each modality, while developing decision fusion classifiers for HIFs characterization.

5.6 Conclusions

Findings from the study validate the importance of hybrid BCIs in characterising the HIFs. Fusion of EEG, fNIRS, and HR modalities showed to accurately classify multiple QoE dimensions, thus highlighting their compatibility for objective QoE modelling. Moreover, the fusion of neuro-physiological modalities in the second phase was shown to more accurately represent QoE percepts, indicating the superiority of the technique. The classification results presented here validate the use of hybrid BCIs for characterising HIFs. As such, it is expected that the pre-existing state-of-the-art objective QoE models can benefit a lot from the incorporation of such hybrid BCI-based techniques. This application of hybrid BCIs will be explored in the following chapter.

Chapter 6

QoE characterization using a passive hybrid BCI

6.1 Preamble

This chapter is compiled from material extracted from manuscripts published in the Springer journal on Human-centric Computing and Information Sciences [48] and a manuscript that has been accepted at 2016 IEEE Workshop on Multimedia Signal Processing [49].

6.2 Introduction

The majority of state-of-the-art objective QoE models rely on technological and contextual aspects of the service [7, 77]. However, in order to develop truly 'user-centric' QoE models, HIFs, such as user emotions and attitudes, also need to be incorporated. For the development of such objective QoE models, BCIs can be used. As such, previous chapters have established the use of BCIs for characterising HIF-related information from the users, using binary classifiers. However, the goal of state-of-the-art objective QoE models is to quantify the QoE. Therefore, towards developing hybrid BCI-based objective models that quantify the QoE, while incorporating HIFs, a regression analysis must be undertaken to estimate the QoE on a continuous scale.

As a first step towards incorporating HIFs into objective QoE models, we explored the use of the so-called passive hybrid BCIs. Passive hybrid BCIs characterise users' affective states using fusion of multiple neurophysiological and physiological modalities [78]. Specifically, the affective human factors were chosen as they are significant indicators of the two perceptual constructs, 'listening pleasure' and 'prosody', of TTS QoE, as established in Section 2.5. Fundamentally, affective dimensions of valence and arousal indicate the changes in 'listening pleasure' and 'prosody', respectively.

Here, we are interested in exploring the use of hybrid BCIs, based on fusion of EEG, fNIRS and fNIRS-derived heart rate, to measure HIFs, and incorporate them into objective models of speech QoE perception. A scenario based on text-to-speech (TTS) systems, based on the PhySyQX database, is explored. The general scheme of the proposed hybrid BCI system for user QoE perception monitoring is shown in Fig. 6.1. Within this framework, the audio signal, generated by a TTS system, is used to extract technological influence factors, whereas the hybrid BCI collects implicit human influential factors. For simplicity and without loss of generality, this study does not investigate the effects of contextual factors. As such, the final user-perceived QoE model is obtained as a combination of the measured TIF and HIFs. The measured QoE provides service providers with invaluable information that allows them to fine tune service parameters, thus resulting in improved user experience. Therefore, here, we have proposed a novel BCI-based approach that incorporates neurophysiology-based objective measures of HIFs into state-of-the-art objective models for the measurement of QoE, thus resulting in significant gains in performance. Furthermore, this chapter compares the performance of objective BCI models, based on individual neurophysiological modalities, and a hybrid BCI model, developed using fusion of individual modalities.

The remainder of this chapter is organized as follows. Section 6.3 provides an overview of the methodology and experimental setups used in the studies. Sections 6.4 and 6.5 describe the experimental results and discussion, respectively. Lastly, conclusions are presented in Section 6.6.

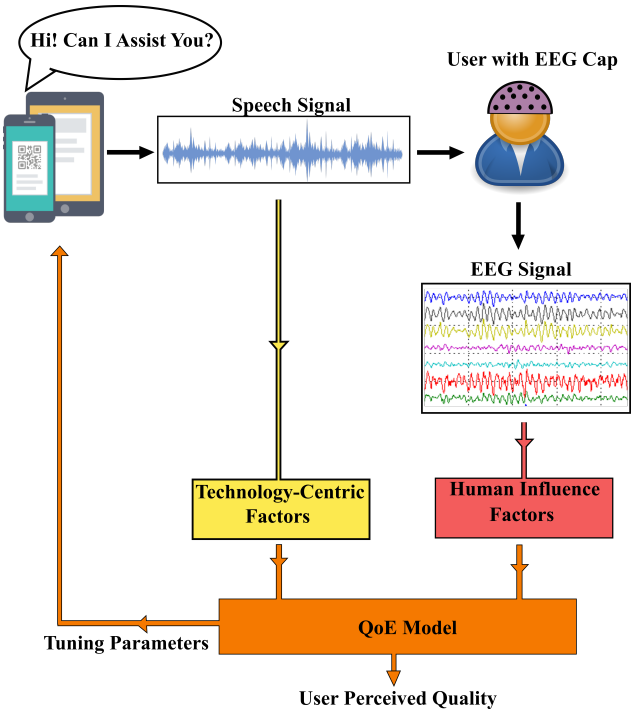


Figure 6.1 – The figure shows an overview of the hybrid BCI approach for monitoring user QoE.

6.3 Methods and Materials

In the following subsections, we only describe the objective assessment techniques that were used to develop the passive hybrid BCI-based objective QoE assessment model, as the subjective assessment techniques have been described in Section 1.1.1.

6.3.1 Objective Assessment Methods

The state-of-the-art objective QoE assessment models are technology-centric, as they aim to replace the human rater by a computer algorithm that has been developed to extract relevant features from the analyzed signal (speech, audio, image, or video) and map a subset/combination of such features into an *estimated* QoE value. As such, for TTS systems, studies have shown the importance of signal-based metrics [15], such as prosody and articulation [79]. Recently, two quantitative parameters were shown useful [79], thus are used in our TTS study: 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 the macro-prosodic or intonation-related properties of speech, $MFCC_2$ models articulation-related properties

[79]. In our experiments, the openSMILE toolbox [80] was used to extract these features using the default window length of 25 ms and frame shift of 12.5 ms.

Towards developing hybrid BCIs, we leveraged data acquired from EEG and fNIRS. Typical EEG-BCI features involve the calculation of specific EEG frequency subband powers, such as delta, theta, alpha, beta, or gamma sub-bands, as well as their interactions [209], whereas fNIRS-based BCIs rely on cortical haemodynamics related features, such as average and peak values for $\Delta[HbO]$ and $\Delta[HbR]$ concentrations [47]. Furthermore, features from other physiological modalities, such as heart rate monitors, have also proven useful in characterising valence [28]. Therefore, based on the correlation analysis results in Section 3.4.1 and Section 4.4.2, the features that showed maximum correlation with the subjective valence and arousal were used as neurophysiological correlates of affective states. As such, for EEG signals, local efficiency (E_l) metrics derived from high-beta subband, and medial beta power (MBP) [48] were used to model valence and arousal, respectively. For fNIRS signals, 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. Finally, for heart rate signal, mean HRV correlated with valence and none of the heart rate-derived features showed correlation with arousal. Therefore, the above mentioned neurophysiological features were used to develop BCI models based on each individual modality, as well as hybrid BCI model.

6.3.2 QoE Model Performance Assessment

In order to assess QoE model performance, six tests were conducted. First, we explored the goodness-of-fit (r^2) and root mean-squared error (RMSE) achieved by using only the technology-centric speech metric as a correlate of the QoE score reported by the listeners (denoted as QoE_{Tech}). 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 (denoted as QoE_{HIF}). Gains in the goodness-of-fit metric should indicate the benefits of including HIFs into QoE perception models. Next, we replaced the ground-truth HIFs by the BCI features, based on each individual modality, that are used as correlates of the listener's emotional states (denoted as QoE_{EEG}, QoE_{fNIRS} and QoE_{HR}). It is expected that the r^2 achieved will lie between those

achieved without and with HIFs, thus signalling the importance of BCIs in QoE perception modelling. Finally, we replaced the ground-truth HIFs by the hybrid BCI features (combination of individual BCI features). It is expected that an increased r^2 can be achieved using hybrid BCI models, thus, establishing the use of hybrid BCI models.

Towards this end, the goodness-of-fit measures were obtained by developing linear regression equations for each of the six proposed tests ($i = 1, \ldots, 3$). Linear regression model 'i' had dependent variable y_i as a linear combination of 'p' independent variables (or regressors, x_{ip}) weighted by regression coefficients (β_p) and error (ϵ_i). The linear regression is formulated as follows:

$$y_i = \epsilon_i + \beta_1 x_{i1} + \beta_2 x_{i2} ... + \beta_p x_{ip} = \boldsymbol{x_i^T \beta} + \epsilon_i, \tag{6.1}$$

The values of β and ϵ are estimated using least squares fitting on training data. The test set was formed by randomly selecting at least two stimuli from each of the 11 speech systems, whereas the remaining stimuli formed the training set. Finally, the model performances were compared using a F-test [210]:

$$F = \frac{(SS_1 - SS_2)/(df_1 - df_2)}{SS_2/df_2}, \tag{6.2}$$

where SS_1 and SS_2 are the sum of squared errors for the two models and df_1 and df_2 are the degrees of freedom for the two models.

6.4 Experimental Results

In this section, we report the experimental results obtained from the subjective and objective methodologies used.

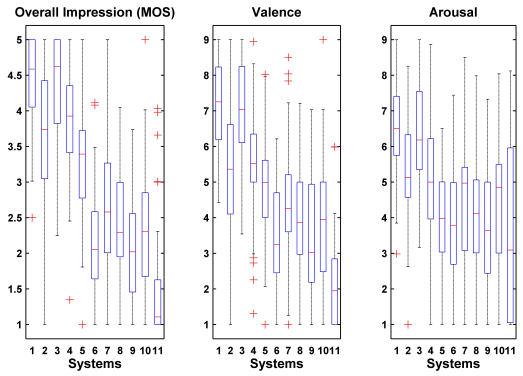


Figure 6.2 – Box plots for the subjective overall impression (QoE-MOS), valence and arousal scores.

6.4.1 Subjective Data Evaluation

Initially, the impact of varying TTS system quality on affective factors was analyzed by computing descriptive statistics, as shown in Fig. 6.2. It was observed that systems 1 and 3 showed the highest quality ratings, which can be expected as both corresponded to natural voices. However, the other two natural voice systems (2 and 4) were rated at medium quality levels. This was due to the fact that the speaker used for system 4 was specifically asked to speak with a neutral intonation and listeners reported voice 2 as sounding breathy, thus lower in quality than the other natural voices. Regarding the TTS systems, system 11 scored the least in terms of quality, valence and arousal. In general, the synthesized speech systems scored lower than natural systems. However, comparing the systems which used synthesized voices, system 5 scored the maximum in terms of quality and valence. In order to test the effects of the natural and synthesised speech systems in terms of perceived QoE, an ANOVA was used. A significant effect was found ($F(10, 913) = 143.32; p \leq 0.01$). Moreover, post-hoc pairwise t-test comparisons with Bonferroni correction showed QoE-MOS scores to significantly differ between natural voices and TTS system outputs.

Similar analysis as above was performed for the arousal and valence ratings. For valence, a statistical difference across eleven condition groups was found ($F(10, 913) = 96.28; p \leq 0.01$), as was the case with arousal ($F(10, 913) = 31.5; p \leq 0.01$). The stronger F-statistic seen for valence over arousal suggests that synthesized speech quality has a stronger influence on the perceived pleasantness of the experienced files. Post-hoc pairwise t-test comparisons with Bonferroni correction were also computed for the two emotional primitives. It was found that valence and arousal ratings significantly differed between the natural and synthesized voices. Moreover, to better understand the

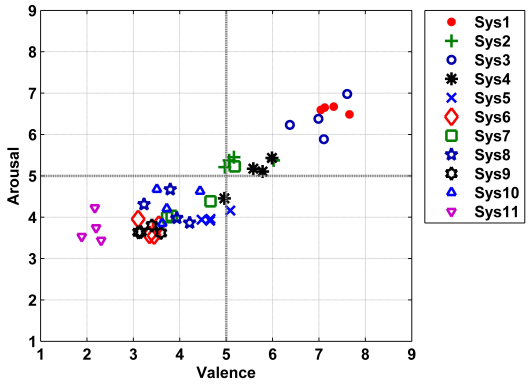


Figure 6.3 – Subjective Valence vs. Arousal emotional map across the 11 tested conditions.

Table 6.1 – 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	RMSE
1	S	$QoE_{Tech} = 0.36 - 0.56 * MFCC_2 + 0.44 * sF0''$	0.76	0.14
2	S, Sub	$QoE_{HIF} = 0.004 + 0.02 * MFCC_2 + 0.05 * sF0'' + 1.53 * Val - 0.52 * Ar$	0.96	0.05
3	S, E	$QoE_{EEG} = 0.30 - 0.87 * MFCC_2 + 0.67 * sF0'' + 0.51 * E_l + 0.80 * MBP$	0.87	0.11
4	S, F	$QoE_{fNIRS} = 0.87 - 0.56 * MFCC_2 + 0.34 * sF0'' - 0.33 * HbR - 0.44 * HbO$	0.84	0.13
5	S, H	$QoE_{HR} = 0.46 - 0.49 * MFCC_2 + 0.39 * sF0'' - 0.15 * HRV$	0.79	0.13
6	S, E, F, H	$QoE_{hybrid} = 0.05 - 0.74 * MFCC_2 + 0.59 * sF0'' + 0.40 * E_l + 0.70 * MBP - 0.19 * HbR - 0.10 * HbO - 0.12 * HRV$	0.90	0.10

impact of TTS system quality on users' emotional ratings, the 2-dimensional valence-arousal map was used, as depicted by Fig. 6.3. It can be seen that the natural voice cases were present mostly in the high valence and high arousal quadrant of the valence-arousal map, whereas all synthesized voices existed in the low valence and low arousal quadrant.

6.4.2 Objective Model Evaluation

As mentioned in Section 6.3.2, three QoE models were implemented in order to gauge the benefits of including HIFs, as well as BCI features into the equation. For this study, the QoE models reported in Table 6.1 were found.

The obtained goodness of fit (r^2) value for model 1 was 0.76 with an RMSE of 0.136. For model 2, in turn, the obtained r^2 value was 0.96 with an RMSE of 0.05, thus again highlighting the importance of HIFs in QoE perception modelling. For models based on BCI features derived from individual physiological modalities, the obtained r^2 values were 0.87, 0.84 and 0.79 for EEG, fNIRS

and heart rate, respectively, as obtained from models 3, 4 and 5. Lastly, for hybrid BCI model 6, the obtained r^2 value was 0.90 with an RMSE of 0.10. When comparing the output of the objective QoE model in 6 and model in 2, a Pearson correlation coefficient of 0.96 was obtained. Finally, the F-tests for model comparison revealed significant difference ($F > 2.88, p < 0.05$) between models 1, 2 and 6. Thus, indicating that a QoE model that incorporates information regarding TIFs and HIFs, using a hybrid BCI approach, resulted in significantly better performance compared to model based on only TIFs.

It is important to notice that results from linear modelling techniques are reported here. In fact, through Pearson correlation analysis between neurophysiological features and QoE we have shown linear relationship between them in previous chapters. However, mutual information, which quantifies non-linear dependencies, between neurophysiological features and QoE was always lower than Pearson correlation coefficients. Therefore, only linear models were chosen to model QoE using neurophysiological features.

6.5 Discussion

In this section, we discuss the experimental results obtained from the subjective and objective methodologies used.

6.5.1 Role of HIFs in QoE Modelling

Recently, HIFs and objective HIF characterization have gained increasing attention from QoE researchers [81, 82, 83]. Previously, researchers have investigated the effects of user expectation on QoE [211]. In a similar vein, we have evaluated the effects of users' affective states on overall QoE perception. We have found evidence from the subjective assessment tests that indeed the users' perceived affective states change with varying speech quality.

It is visible from the valence-arousal maps depicted by Fig. 6.3 that poor quality speech stimuli produced low arousal and low valence states, thus producing states ranging from 'sad' to 'miserable' in listeners. High quality stimuli, on the other hand, incited high arousal and high valence states, thus making users feel 'alert' or 'amused'. Moreover, as found in chapter 2, the measured

HIFs showed high (significant) correlation with QoE-MOS. When HIFs were combined with existing state-of-the-art technology-centric speech quality metrics, as in model 2 in Table 6.1, improvements in QoE measurement performance were observed and a relative gain of 26.3% was seen for TTS systems. These findings suggest that affective states can indeed directly influence a listener's perceived experience (or QoE) with speech synthesizers.

Nonetheless, despite the improvements seen when adding HIFs to objective quality models (i.e., model 2 in Table 6.1), there was still a gap to perfect goodness-of-fit, thus suggesting that the inclusion of alternate additional HIFs may be important. To this end, future studies should investigate the effects of e.g., attention, cognitive load, fatigue and/or user engagement.

6.5.2 Hybrid BCI Advantages and Limitations

The use of BCIs during subjective QoE assessment has two major advantages. First, BCIs may allow for monitoring of the listener's affective states in an objective manner, thus potentially reducing listener biases in subjective tests, particularly for TTS systems [212]. To this end, first, typical EEG-based metrics were used to quantify two emotional primitives: arousal and valence. More specifically, the local efficiency metric derived from high-beta band was used as a correlate of valence and the medial beta power (MBP) as a correlate of arousal. Next, fNIRS-derived measures of average $\Delta[HbR]$ computed from the right temporal region and average $\Delta[HbO]$ at the temporo-parietal region were used to quantify valence and arousal, respectively. Finally, a heart rate-derived measure of average HRV was used as a correlate for valence. However, some gaps were observed between models 2 and 3, 4 and 5 for TTS systems.

Towards improving the performance of models based on individual neurophysiological modalities, a "hybrid" BCI-based approach was leveraged in model 6, reported in Table 6.1. The model 6 performed better than the models based on BCIs developed using individual modalities; however, model 2 still performed better than model 6. The observed gap between QoE models found with subjective and with hybrid BCI features can be attributed to the low average correlations obtained between the neurophysiological features and subjective valence and arousal. Overall, it is expected that more powerful models can be obtained once improved BCI features are developed. Alternately, additional neuro-physiological signal modalities may be incorporated for human affective state mon-

itoring, such as galvanic skin response and eyetracking. The development of such "hybrid" BCIs is the aim of our ongoing research.

The second main advantage of using hybrid BCIs to objectively monitor listener affective states is that it allows for continuous real-time monitoring of listener affective states. In practice, it is not possible to have listeners attend to the quality of a presented stimuli continuously, as well as report the elicited affective states. Such cognitive load demands will result in unwanted effects in the obtained ratings, as recently reported by [213]. As such, the use of a hybrid BCI can allow the participants to focus on the QoE experiment fully, particularly if it involves time-varying distortions, such as voice over internet protocol (VoIP). While the present experiments did not involve time-varying distortions, the high correlations obtained between the objective and subjective ratings suggest that the proposed objective regressors could be used for such tasks. Overall, a gain of 18.4% in QoE measurement performance could be seen once hybrid BCI features were used, relative to using only technological factors for TTS systems.

Furthermore, it is important to validate the results of these models on a wider population, mainly older population, as majority of the participants in our study were young and the average age of participants was approximately 24 years. In general, cortical activations might differ between the two populations to some extent. However, it is expected that the neurophysiological features developed in this thesis will be largely useful in monitoring QoE, across different populations.

6.6 Conclusion

Speech Quality-of-Experience (QoE) perception is known to be influenced by internal human factors, as well as external technological and contextual factors. Existing objective QoE models, however, have focused mostly on the latter two and have omitted human QoE factors, such as affective states, from the equation. In this chapter, we have taken the first steps towards showing the importance of incorporating human affective states into speech QoE models using hybrid BCIs. Subjectively, we showed the impact of speech distortions on the listener's perceived valence and arousal states, and in turn, their effect on perceived QoE. Objectively, on the other hand, we have proposed the use of hybrid BCIs to measure the listener's valence and arousal levels. Through

regression analysis, we showed that features extracted from an hybrid BCI could improve QoE models performance by as much as 18.40% for TTS systems.

Chapter 7

Summary and Future Research Directions

In this chapter, a general discussion for this doctoral thesis is presented, moreover some suggestions for future research directions are also proposed.

7.1 Summary

This doctoral thesis investigated the application of hybrid BCIs based on neuroimaging techniques, such as EEG and fNIRS, to quantify HIFs and incorporate them into state-of-the-art objective QoE models. The objective was to develop, validate and fuse different EEG and fNIRS based features that can be used for long-term monitoring of HIFs, and ultimately incorporated into state-of-the-art objective QoE models based on TIFs and CIFs. In the following subsections, we will discuss the contributions of this doctoral thesis towards achieving the goal of development of a hybrid BCI system that can be incorporated into an objective QoE assessment model.

7.1.1 Development of a neurophysiological database

Towards developing EEG and fNIRS derived features for characterising HIFs, it is important to first acquire neurophysiological data. However, as acquisition of neurophysiological data requires

expert supervision and interdisciplinary research teams. To date, very few open-source databases exist, such as DEAP [28], DECAF [29], that can be used by the research community to develop features that quantify HIFs. Moreover, none of the above mentioned databases were developed with a focus on quantifying QoE. Therefore, we developed an open-source database for physiological evaluation of QoE, the so-called PhySyQX database. The PhySyQX database consists of EEG and fNIRS data acquired by probing various cortical regions, using 62 and 60 channels, respectively. The test stimuli used for the experiment consisted of 7 different state-of-the-art TTS systems and 4 natural voices. A total of 21 participants (8 females) volunteered to participate in the study.

Furthermore, for subjective assessment, the the PhySyQX database consists of scores for attitudinal, as well as affective dimensions. Previously developed databases for TTS QoE measurement, such as Blizzard Challenges [50], only consider effects of attitudinal HIFs on QoE. However, in this thesis we have established the importance of measuring affective HIFs using factor analysis, where the affective dimensions of valence and arousal showed significant factor loadings on two different perceptual dimensions of 'voice pleasantness' and 'prosody', respectively, along with several other attitudinal HIFs. This suggests that, both attitudinal and affective dimensions are equally important for quantifying the QoE.

7.1.2 EEG-based BCI system for HIFs characterisation

Previously proposed EEG-based BCI systems leveraged ERP based features to characterise QoE. However, such features can only characterise short duration multimedia signals corrupted by stationary noises, whereas most of the real-world multimedia signals are long duration signals and noise sources are time-varying. Therefore, in this thesis we proposed two different classes of features derived from EEG signals. The first class of features were derived using the power spectrum of EEG signals from different EEG electrodes. These consisted of event related desynchronisation (ERD) and asymmetry index based features. The second class of features, in turn, were derived from the cross-spectrum analysis between different EEG electrodes. These consisted of graph theoretical features, such as local and global efficiencies, clustering coefficient, characteristic pathlength and small-worldness.

In order to validate the features, we utilised the the PhySyQX and the DEAP database. Both classes of features were able to significantly characterise the HIFs and showed similar performance

for the TTS stimuli. However, comparatively, graph theoretical features showed significantly better performance than power spectrum based features, for music video stimuli. Thus, indicating the superiority of features that characterise information flow between different cortical regions for characterising the QoE of audio-vidual stimuli. As such, audio-visual stimuli processing recruits more cortical regions as compared to just audio stimuli, thus leading to more complex dynamics of cortical information flow.

Furthermore, EEG based features highlighted the activation of different cortical regions in response to the test stimuli. Specifically, it was observed that with better quality TTS stimuli, there was a higher right and left temporal activation. As indicated by previous research, activation of right and left temporal regions is associated with prosody and intelligibility perception, respectively. Therefore, these reuslts provide deeper neural insights into QoE formation processes.

7.1.3 fNIRS-based BCI system for HIFs characterisation

fNIRS has recently emerged as a neuroimaging technique that provides equivalent information regarding cortical activity, as compared to EEG. However, fNIRS provides better spatial resolution as compared to EEG. Hence, we leveraged fNIRS recordings to validate the neural insights obtained from EEG. Moreover, we explored various fNIRS based features that were used to characterise HIFs. The fNIRS-based features were derived from $\Delta[HbO]$ and $\Delta[HbR]$ concentration curves. It was observed that the fNIRS based features showed equivalent performance, as compared to EEG based features, across various subjective dimensions. Furthermore, we extracted heart-rate from the fNIRS signals, and used it to develop different features for characterising HIFs. The heart-rate based features, generally, showed lower performance as compared to EEG and fNIRS based features.

7.1.4 Hybrid BCI system for HIFs characterisation

The success of EEG and fNIRS based BCI systems inspired the development of hybrid BCI systems, that fuse information from different neurophysiological and physiological modalities. As such, we found that the developed fusion classifiers showed better performance across various subjective dimensions, compared to using BCIs based on single modalities. Therefore, future studies should explore various techniques to fuse information from different BCI and physiological modalities.

This might require more data to optimize the modality weights. Towards this end, the developed PhySyQX data can be appended with the other databases collected by researchers, for example, DEAP [28] and DECAF [29].

7.1.5 Incorporation of hybrid BCI into objective QoE assessment model

Finally, first steps were taken towards incorporating HIFs into state-of-the-art objective QoE assessment models. To achieve this, the physiological and neurophysiological features that showed highest significant correlation with subjective dimensions were used in the model. Furthermore, we compared the performance of different models based on individual modalities and hybrid BCIs, which indicated slightly better performance of hybrid BCI based models.

7.2 Future Research Directions

1. *Development of better neurophysiological features:* Although the hybrid BCI tools proposed in this thesis demonstrate excellent capability to characterise HIFs, there is still scope to improve their performance. This can be achieved using better neurophysiological features that encode more information regarding HIFs. First steps towards this have been taken in [214], where we leveraged the mutual information between the interhemispheric interactions in spectro-temporal patterns of EEG activity to characterise users' affective states. Moreover, EEG-based features that characterise accurate 'directional' flow of information between cortical regions, such as directed transfer function, can provide better neurophysiological insights [215]. However, recently, for EEG-based BCIs, wavelet transformation based features have shown great potential in characterising users' affective states [175]. Furthermore, for fNIRS-based BCIs, exploring laterality features [37], that combine information from the two hemispheres, or time-frequency based features [216] can also prove useful.
2. *Alternate neurophysiological modalities:* This thesis has focussed on using EEG and fNIRS as neuroimaging tools for characterising HIFs. However, various other neurophysiological modalities have been proven useful in characterising users' states, such as magnetoencephalography (MEG) [29] and functional magnetic resonance imaging (fMRI)[217]. Moreover, addition of different physiological signals, such as respiration, galvanic skin response and skin temperature, can also prove useful in the development of better performing hybrid BCI models [28, 29].

Recent advances in wearable technologies have made seamless acquisition of such biosignals possible. Therefore, future studies could aim to develop techniques to incorporate different wearable technologies, for continuous measurement of users' internal states, into objective QoE assessment models.

3. *Development of signal quality adaptive hybrid BCI systems:* The signals from different physiological and neurophysiological modalities are often contaminated with various sources of noise [218, 219], that significantly hinders the task of HIFs characterisation; this is particularly true when low-cost portable devices are used which are sensitive to movement artefacts. Therefore, it is necessary to develop techniques that adapt to biosignal quality. Towards this end, the first steps have already been taken in [127], where we proposed a multi-modal quality adaptive affect recognition system. The proposed system fuses information from various biomedical sources, such as wearable EEG, ECG, face tracking and GSR sensors, for affect characterisation, while adapting to the quality of the signals. The proposed system performed significantly better than the non-adaptive affect recognition system. However, there still exists a scope for improvement in developing better signal quality estimators and adaptive BCI systems, which should be explored in future studies.

4. *Combining hybrid BCIs for QoE assessment of different multimedia technologies:* The last chapter in this thesis explores the advantages of incorporating hybrid BCIs into objective QoE assessment model for TTS systems. However, as new multimedia technologies gain popularity, such as hands free communications, voice over internet protocol (VoIP) services etc., these new technologies would need to continuously monitor the delivered QoE for customer satisfaction. Therefore, future studies should explore the extension of proposed methodologies in Chapters 3-6 for objective QoE assessment of different multimedia technologies. A pilot study involving hands-free communications in reverberant rooms showed similar benefits as those reported in 6.5.2 [48].

Appendix - A

The sentence group A consisted of the following 4 sentences:

1. You need to make a right turn at the next intersection. I could not find any inexpensive Japanese restaurant, however there are 3 inexpensive Chinese restaurants and 1 inexpensive Vietnamese restaurant there. There is a good one a few blocks away. They may still have tables available. Do you want me to book a table? I'll do my best to find you a table.
2. You have meeting tomorrow. It has been scheduled for 10-'o'-clock. Is that okay? You will be going to go over last quarter's sales figures. Please make some suggestions on improving the bottom line. Can you make it? Flying to San Francisco is not possible right now so you won't be able to make it. Do you want to book another flight tomorrow?. Thank you.
3. Did you say you wanted to hear the forecast for New York? New York will be dealing with a storm of its own as it moves through the north-east. Initially the storm will be relatively warm leading to moderate to occasional heavy rainfall from the coast of North Carolina through New England. The south-west will se a range of temperatures from the 30s in the higher elevatiosn to the 60s closer to the sea level.
4. New York celebrated the Chinese year of the monkey on Monday the 25th of January 2007, with a program of events spanning Washington Square, Chinatown and the newly renovated Empire State building. The area was packed with around 40000 New Yorkers throughout the day.

The sentence group B consisted of the following 4 sentences:

1. Many of the early experimenters spent a lot of time in trying to improve tiny transparent crystals they had made. Crystals which they erroneously believed were only silica or other hard but non-diamond crystals.
2. Slowly life returned to something like normal in the straits and the bewildered survivors were able to bury their dead and salvage what they could have homes and villages that had been swept by Tsunamis and showered with ash.
3. Many city stock brokers, advertising agencies and banking houses are paying substantial bonuses to their high earning staff to beat the effects of a substantial increase in taxes in the wake of possible labour party victory of April 9.
4. To our utter dismay and astonishment he told us that our certificates meant nothing at all to him or BSAC and even if Jacques Cousteau would have come along with a PADI qualification no notice will be taken of it.

Bibliography

[1] S. Moller and A. Raake, eds., *Quality of Experience Advanced Concepts, Applications and Methods*. Springer, 2014.

[2] F. Sharbrough, G. Chatrian, R. Lesser, H. Lüders, M. Nuwer, and T. Picton, "American electroencephalographic society guidelines for standard electrode position nomenclature," *J. Clin. Neurophysiol*, vol. 8, no. 2, pp. 200–202, 1991.

[3] ITU-T, "ITU-T recommendation E.800: Definitions of terms related to quality of service.," tech. rep., International Telecommunication Union, Geneva, Switzerland, 2008.

[4] Qualinet, "Qualinet White Paper on Definitions of Quality of Experience Output from the fifth Qualinet meeting, Novi Sad, March 12, 2013, Version 1.2," tech. rep., Qualinet COST IC 1003, 2013.

[5] J. R. Wolpaw and E. W. Wolpaw, "Brain-computer interfaces: something new under the sun," *Brain-computer interfaces: Principles and practice*, pp. 3–12, 2012.

[6] H. Banville and T. Falk, "Recent advances and open challenges in hybrid brain-computer interfacing: a technological review of non-invasive human research," *Brain-Computer Interfaces*, vol. 3, no. 1, pp. 9–46, 2016.

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

[8] K. Kondo, "Subjective quality measurement of speech: its evaluation, estimation and applications," 2012.

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

[10] P.J.Lang, "The emotion probe: Studies of motivation and attention.," *American Psychologist*, vol. 50(5), pp. 372–385, 1995.

[11] 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.

[12] ITU-T, "Itu-t recommendation p.862: Perceptual Evaluation of Speech Quality (PESQ): an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," tech. rep., ITU-T. Rec., 2001.

[13] ITU-T, "ITU-T Recommendation P.863: Perceptual objective listening quality assessment," tech. rep., ITU-T Geneva, 2011.

[14] ITU-T, "ITU-T recommendation P.563: Single-ended method for objective speech quality assessment in narrow-band telephony applications," tech. rep., ITU-T, Geneva, Switzerland, 2004.

[15] T. H. Falk and S. Möller, "Towards signal-based instrumental quality diagnosis for text-to-speech systems," *Signal Processing Letters, IEEE*, vol. 15, pp. 781–784, 2008.

[16] T. O. Zander and C. Kothe, "Towards passive brain–computer interfaces: applying brain–computer interface technology to human–machine systems in general," *Journal of neural engineering*, vol. 8, no. 2, p. 025005, 2011.

[17] K. Polat and S. Güneş, "Classification of epileptiform EEG using a hybrid system based on decision tree classifier and fast Fourier transform," *Applied Mathematics and Computation*, vol. 187, no. 2, pp. 1017–1026, 2007.

[18] J. d. R. Millán, M. Franzé, J. Mouriño, F. Cincotti, and F. Babiloni, "Relevant EEG features for the classification of spontaneous motor-related tasks," *Biological cybernetics*, vol. 86, no. 2, pp. 89–95, 2002.

[19] C. A. M. Lima, A. L. V. Coelho, and S. Chagas, "Automatic EEG signal classification for epilepsy diagnosis with Relevance Vector Machines," *Expert Systems with Applications*, vol. 36, no. 6, pp. 10054–10059, 2009.

[20] D. Garrett, D. A. Peterson, C. W. Anderson, and M. H. Thaut, "Comparison of linear, nonlinear, and feature selection methods for EEG signal classification," *IEEE Transactions on neural systems and rehabilitation engineering*, vol. 11, no. 2, pp. 141–144, 2003.

[21] A. Subasi and M. I. Gursoy, "EEG signal classification using PCA, ICA, LDA and support vector machines," *Expert Systems with Applications*, vol. 37, no. 12, pp. 8659–8666, 2010.

[22] M. E. Tipping, "Sparse Bayesian learning and the relevance vector machine," *The journal of machine learning research*, vol. 1, pp. 211–244, 2001.

[23] M. Han and L. Sun, "EEG signal classification for epilepsy diagnosis based on AR model and RVM," pp. 134–139, 2010.

[24] V. Jurcak, D. Tsuzuki, and I. Dan, "10/20, 10/10, and 10/5 systems revisited: their validity as relative head-surface-based positioning systems," *Neuroimage*, vol. 34, no. 4, pp. 1600–1611, 2007.

[25] M. Coles and M. Rugg, *"Event-related brain potentials: An introduction."*. Oxford University Press, New York, 1995.

[26] J. Antons *et al.*, "Subjective listening tests and neural correlates of speech degradation in case of signal-correlated noise," *Audio Engineering Society Convention 129*, pp. 1–4, 2010.

[27] S. Arndt, J.-N. Antons, R. Schleicher, S. Moller, and G. Curio, "Using electroencephalography to measure perceived video quality," *IEEE J. of Selected Topics in Signal Processing,*, vol. 8, no. 3, pp. 366–376, 2014.

[28] S. Koelstra *et al.*, "DEAP: A database for emotion analysis; using physiological signals," *Affective Computing, IEEE Transactions on*, vol. 3, no. 1, pp. 18–31, 2012.

[29] M. Abadi, R. Subramanian, S. Kia, P. Avesani, I. Patras, and N. Sebe, "DECAF: MEG-Based Multimodal Database for Decoding Affective Physiological Responses," *IEEE Transactions on Affective Computing*, vol. 6, pp. 209–222, July 2015.

[30] S. Arndt, R. Schleicher, and J.-N. Antons, "Does low quality audiovisual content increase fatigue of viewers," *Proceedings perceptual quality of systems (PQS)*, pp. 1–4, 2013.

[31] J.-N. Antons, R. Schleicher, S. Arndt, S. Möller, and G. Curio, "Too tired for calling? A physiological measure of fatigue caused by bandwidth limitations," *Proceedings of Fourth International Workshop on Quality of Multimedia Experience (QoMEX), 2012*, pp. 63–67, 2012.

[32] 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.

[33] C. Lithari *et al.*, "How does the metric choice affect brain functional connectivity networks?," *Biomedical Signal Processing and Control*, vol. 7, no. 3, pp. 228–236, 2012.

[34] T. E. Ward, "Hybrid Optical–Electrical Brain Computer Interfaces, Practices and Possibilities," *Towards Practical Brain-Computer Interfaces*, pp. 17–40, 2012.

[35] T. Fekete *et al.*, "The NIRS analysis package: Noise reduction and statistical inference," *PloS one*, vol. 6, no. 9, p. e24322, 2011.

[36] E. Peck, D. Afergan, and R. Jacob, "Investigation of fNIRS brain sensing as input to information filtering systems," *Proceedings of the 4th Augmented Human International Conference*, pp. 142–149, 2013.

[37] S. Moghimi *et al.*, "Automatic detection of a prefrontal cortical response to emotionally rated music using multi-channel near-infrared spectroscopy," *Journal of Neural Engin.*, vol. 9, no. 2, p. 026022, 2012.

[38] M. R. O. Figner.B., *A handbook of process tracing methods for decision research: A critical review and user's guide.* Using skin conductance in judgment and decision making research., New York, NY: Psychology Press., 2012.

[39] T. Falk *et al.*, "Taking NIRS-BCIs outside the lab: Towards achieving robustness against environment noise," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 19, no. 2, pp. 136–146, 2011.

[40] T. Falk *et al.*, "On the use of peripheral autonomic signals for binary control of body–machine interfaces," *Physiological measurement*, vol. 31, no. 11, p. 1411, 2010.

[41] K. Wiens.S, Mezzacappa.ES, "Heartbeat detection and the experience of emotions," *Cogn Emotion*, vol. 14, pp. 417–427, 2001.

[42] B. Appelhans and L. Luecken, "Heart Rate Variability as an Index of Regulated Emotional Responding," *Review of General Psychology*, vol. 10, pp. 229–240, 2006.

[43] 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.

[44] R. Gupta and T. H. Falk, "Latent factor analysis for synthesized speech quality-of-experience assessment (under review)," *Quality and User Experience*, 2016.

[45] R. Gupta, K. ur Rehman Laghari, and T. H. Falk, "Relevance vector classifier decision fusion and EEG graph-theoretic features for automatic affective state characterization," *Neurocomputing*, vol. 174, Part B, pp. 875 – 884, 2016.

[46] R. Gupta, H. Banville, and T. H. Falk, "Multimodal physiological quality-of-experience assessment of text-to-speech systems (under review)," *IEEE Journal for Selected Topics in Signal Processing: Special Issue on Measuring Quality of Experience for Advanced Media Technologies and Services*.

[47] 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.

[48] 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.

[49] R. Gupta and T. H. Falk, "Physiological quality-of-experience assessment of text-to-speech systems," *IEEE Workshop on Multimedia Signal Processing*, 2016.

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

[51] J. Morris, "Observations: SAM:The self assessment manikin, An effecient cross-cultural measurement of emotional response.," *Journal of advertising research*, vol. 35, no. 6, pp. 63–68, 1995.

[52] D. Tedesco and T. Tullis, "A comparison of methods for eliciting post-task subjective ratings in usability testing," *Usability Professionals Association (UPA)*, pp. 1–9, 2006.

[53] F. Hinterleitner, G. Neitzel, S. Möller, and C. Norrenbrock, "An evaluation protocol for the subjective assessment of text-to-speech in audiobook reading tasks," *Proceedings of the Blizzard challenge workshop, Florence, Italy*, 2011.

[54] A. Tseng, R. Bansal, J. Liu, A. J. Gerber, S. Goh, J. Posner, T. Colibazzi, M. Algermissen, I.-C. Chiang, J. A. Russell, *et al.*, "Using the circumplex model of affect to study valence and arousal ratings of emotional faces by children and adults with autism spectrum disorders," *Journal of autism and developmental disorders*, vol. 44, no. 6, pp. 1332–1346, 2014.

[55] A. K. Syrdal and Y.-J. Kim, "Dialog speech acts and prosody: Considerations for TTS," *Proceedings of Speech Prosody*, pp. 661–665, 2008.

[56] G. Pfurtscheller and F. L. Da Silva, "Event-related EEG/MEG synchronization and desynchronization: basic principles," *Clinical neurophysiology*, vol. 110, no. 11, pp. 1842–1857, 1999.

[57] G. Pfurtscheller and W. Klimesch, "Functional topography during a visuoverbal judgment task studied with event-related desynchronization mapping.," *Journal of Clinical Neurophysiology*, vol. 9, no. 1, pp. 120–131, 1992.

[58] G. Pfurtscheller, "Event-related synchronization (ERS): an electrophysiological correlate of cortical areas at rest," *Electroencephalography and clinical neurophysiology*, vol. 83, no. 1, pp. 62–69, 1992.

[59] J. Kalcher and G. Pfurtscheller, "Discrimination between phase-locked and non-phase-locked event-related EEG activity," *Electroencephalography and clinical neurophysiology*, vol. 94, no. 5, pp. 381–384, 1995.

[60] P. D. Welch, "The use of fast Fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms," *IEEE Transactions on audio and electroacoustics*, vol. 15, no. 2, pp. 70–73, 1967.

[61] G. Sageder, M. Zaharieva, and C. Breiteneder, "Group Feature Selection for Audio-Based Video Genre Classification," *MultiMedia Modeling*, pp. 29–41, 2016.

[62] S. R. Baum and M. D. Pell, "The neural bases of prosody: Insights from lesion studies and neuroimaging," *Aphasiology*, vol. 13, no. 8, pp. 581–608, 1999.

[63] S. K. Scott, C. C. Blank, S. Rosen, and R. J. Wise, "Identification of a pathway for intelligible speech in the left temporal lobe," *Brain*, vol. 123, no. 12, pp. 2400–2406, 2000.

[64] R. J. Zatorre, P. Belin, and V. B. Penhune, "Structure and function of auditory cortex: music and speech," *Trends in cognitive sciences*, vol. 6, no. 1, pp. 37–46, 2002.

[65] C. Mayo, R. A. Clark, and S. King, "Multidimensional scaling of listener responses to synthetic speech," *Proceedings of the 6th Annual Conference of the International Speech Communication Association (Interspeech)*, pp. 1725–1728, 2005.

[66] M. Schröder, "Emotional speech synthesis: a review.," *INTERSPEECH*, pp. 561–564, 2001.

[67] D. B. Headley and D. Paré, "In sync: gamma oscillations and emotional memory," *Frontiers in behavioral neuroscience*, vol. 7, 2013.

[68] L. Pessoa, "On the relationship between emotion and cognition," *Nature Reviews Neuroscience*, vol. 9, no. 2, pp. 148–158, 2008.

[69] E. 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.

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

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

[72] A. Bechara, H. Damasio, A. Damasio, and G. 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.

[73] J. Kim and E. André, "Emotion recognition based on physiological changes in music listening," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 12, pp. 2067–2083, 2008.

[74] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE transactions on pattern analysis and machine intelligence*, vol. 31, no. 1, pp. 39–58, 2009.

[75] R. W. Picard and R. Picard, *Affective computing*, vol. 252. MIT press Cambridge, 1997.

[76] S. Koelstra and I. Patras, "Fusion of facial expressions and EEG for implicit affective tagging," *Image and Vision Computing*, vol. 31, no. 2, pp. 164–174, 2013.

[77] F. Pereira, "Panel on Quality of Experience in Applications, Standardization and Certification," in *in Proc. Quality of Multimedia Experience Workshop, Belgium*, Sept. 2011.

[78] C. Mühl *et al.*, "A survey of affective brain computer interfaces: principles, state-of-the-art, and challenges," *Brain-Computer Interfaces*, vol. 1, no. 2, pp. 66–84, 2014.

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

[80] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," *Proceedings of the International Conference on Multimedia*, pp. 1459–1462, 2010.

[81] D. Geerts *et al.*, "Linking an integrated framework with appropriate methods for measuring QoE," *Second International Workshop on Quality of Multimedia Experience (QoMEX)*, pp. 158–163, 2010.

[82] I. Wechsung *et al.*, "Measuring the Quality of Service and Quality of Experience of multimodal human–machine interaction," *Journal on Multimodal User Interfaces*, vol. 6, no. 1-2, pp. 73–85, 2012.

[83] K. Laghari, K. Connelly, and N. Crespi, "Toward total Quality of Experience: A QoE model in a communication ecosystem," *IEEE Communications Magazine*, vol. 50, pp. 58–65, April 2012.

[84] J. Antons, R. Schleicher, S. Arndt, S. Moller, A. Porbadnigk, and G. Curio, "Analyzing Speech Quality Perception Using Electroencephalography," *IEEE J. Select. Topics Signal Proc*, vol. 6(6), pp. 721–731, 2012.

[85] S. Scholler, S. Bosse, M. S. Treder, B. Blankertz, G. Curio, K.-R. Müller, and T. Wiegand, "Toward a direct measure of video quality perception using EEG," *IEEE Transactions on Image Processing*, vol. 21, no. 5, pp. 2619–2629, 2012.

[86] ITU-T, "ITU-T Recommendation P.800 Methods for subjective determination of transmission quality," tech. rep., International Telecommunication Union, Geneva, Switzerland, 1996.

[87] V. Kraft and T. Portele, "Quality evaluation of 5 german speech synthesis systems," *Acta acustica*, vol. 3, no. 4, pp. 351–365, 1995.

[88] F. Hinterleitner, S. Möller, C. Norrenbrock, and U. Heute, "Perceptual quality dimensions of text-to-speech systems," *Proceedings of the Twelfth Annual Conference of the International Speech Communication Association (Interspeech)*, pp. 2177–2180, 2011.

[89] F. Hinterleitner, C. Norrenbrock, S. Moller, and U. Heute, "What makes this voice sound so bad? A multidimensional analysis of state-of-the-art text-to-speech systems," *IEEE Spoken Language Technology Workshop (SLT)*, pp. 240–245, Dec 2012.

[90] L. R. Tucker and R. C. MacCallum, "Exploratory factor analysis," *Unpublished manuscript, Ohio State University, Columbus*, 1997.

[91] M. Viswanathan and M. Viswanathan, "Measuring speech quality for text-to-speech systems: development and assessment of a modified mean opinion score (MOS) scale," *Computer Speech & Language*, vol. 19, no. 1, pp. 55–83, 2005.

[92] I. Wechsung, M. Schulz, K.-P. Engelbrecht, J. Niemann, and S. Möller, "All users are (not) equal-the influence of user characteristics on perceived quality, modality choice and performance," *Proceedings of the Paralinguistic information and its integration in spoken dialogue systems workshop*, pp. 175–186, 2011.

[93] B. Rainer, M. Waltl, E. Cheng, M. Shujau, C. Timmerer, S. Davis, I. Burnett, C. Ritz, and H. Hellwagner, "Investigating the impact of sensory effects on the quality of experience and emotional response in web videos," *Fourth International Workshop on Quality of Multimedia Experience (QoMEX), 2012*, pp. 278–283, 2012.

[94] U. Reiter and K. De Moor, "Content categorization based on implicit and explicit user feedback: combining self-reports with EEG emotional state analysis," *Proceedings of Fourth international workshop on Quality of multimedia experience (QoMEX), 2012*, pp. 266–271, 2012.

[95] M. A. Hogg and D. Abrams, "Social cognition and attitudes," *Psychology*, pp. 684–721, 2007.

[96] A. Mehrabian, "Basic Dimensions for a General Psychological Theory Implications for Personality, Social, Environmental, and Developmental Studies," 1980.

[97] G. Wolf, "Measuring mood -current research and new ideas," 2009.

[98] D. Bos, "EEG-based emotion recognition: The influence of visual and auditory stimuli," 2006.

[99] S. Kai *et al.*, "An improved valence-arousal emotion space for video affective content representation and recognition," in *IEEE International Conference on Multimedia and Expo (ICME), 2009.*, pp. 566–569, 2009.

[100] G. Krausz, R. Scherer, G. Korisek, and G. Pfurtscheller, "Critical decision-speed and information transfer in the "Graz Brain–Computer Interface"," *Applied psychophysiology and biofeedback*, vol. 28, no. 3, pp. 233–240, 2003.

[101] B. Blankertz, G. Dornhege, M. Krauledat, M. Schroder, J. Williamson, R. Murray-Smith, and K.-R. Muller, "The Berlin Brain-Computer Interface presents the novel mental typewriter Hex-o-Spell.," *Proceedings of the 3rd International Brain Computer Interface Workshop and Training Course*, pp. 108–109, 2006.

[102] E. Donchin, K. M. Spencer, and R. Wijesinghe, "The mental prosthesis: assessing the speed of a P300-based brain-computer interface," *IEEE transactions on rehabilitation engineering*, vol. 8, no. 2, pp. 174–179, 2000.

[103] G. Muller-Putz, R. Scherer, C. Neuper, and G. Pfurtscheller, "Steady-state somatosensory evoked potentials: suitable brain signals for brain-computer interfaces?," *IEEE transactions on neural systems and rehabilitation engineering*, vol. 14, no. 1, pp. 30–37, 2006.

[104] G. Borghini, L. Astolfi, G. Vecchiato, D. Mattia, and F. Babiloni, "Measuring neurophysiological signals in aircraft pilots and car drivers for the assessment of mental workload, fatigue and drowsiness," *Neuroscience & Biobehavioral Reviews*, vol. 44, pp. 58–75, 2014.

[105] S. Arndt *et al.*, "Perception of low-quality Videos analyzed by means of Electroencephalography," *Fourth International Workshop on Quality of Multimedia Experience (QoMEX), AUS-Yarra Valley*, 2012.

[106] K. Laghari, R. Gupta, S. Arndt, J. Antons, R. Schleicher, S. Moller, and T. Falk, "Neurophysiological experimental facility for Quality of Experience (QoE) assessment," *Proceedings of International Conference on Quality of Experience Centric Management (QCMan)*, pp. 1300–1305, 2013.

[107] M. Steriade, *Electroencephalography: Basic Principles, Clinical Applications, And Related Fields - Chapter : Cellular substrates of brain rhythms.* Lippincott Williams and Wilkins, 5 ed., 2005.

[108] H. Lodish, A. Berk, S. L. Zipursky, P. Matsudaira, D. Baltimore, and J. Darnell, "Overview of neuron structure and function," *Molecular Cell Biology*, 2000.

[109] J.-N. Antons, "Neural correlates of quality perception for complex speech signals," 2015.

[110] D. A. Pizzagalli *et al.*, "Electroencephalography and high-density electrophysiological source localization," *Handbook of psychophysiology*, vol. 3, pp. 56–84, 2007.

[111] T.-W. Lee, "Independent component analysis," pp. 27–66, 1998.

[112] A. Mognon *et al.*, "ADJUST: An automatic EEG artifact detector based on the joint use of spatial and temporal features," *Psychophysiology*, vol. 48, no. 2, pp. 229–240, 2011.

[113] N. Mammone, F. La Foresta, and F. C. Morabito, "Automatic artifact rejection from multichannel scalp EEG by wavelet ICA," *IEEE Sensors Journal*, vol. 12, no. 3, pp. 533–542, 2012.

[114] J. Polich, "Updating P300: an integrative theory of P3a and P3b," *Clinical neurophysiology*, vol. 118, no. 10, pp. 2128–2148, 2007.

[115] S. Lloyd-Fox, A. Blasi, and C. E. Elwell, ""Illuminating the developing brain: the past, present and future of functional near infrared spectroscopy"," *Neuroscience & Biobehavioral Reviews*, vol. 34, no. 3, pp. 269–284, 2010.

[116] M. Ferrari and V. Quaresima, "A brief review on the history of human functional near-infrared spectroscopy (fNIRS) development and fields of application," *Neuroimage*, 2012.

[117] R. Sitaram, A. Caria, and N. Birbaumer, "Hemodynamic brain–computer interfaces for communication and rehabilitation," *Neural networks*, vol. 22, no. 9, pp. 1320–1328, 2009.

[118] D. Malonek and A. Grinvald, "Interactions between electrical activity and cortical microcirculation revealed by imaging spectroscopy: implications for functional brain mapping," *Science*, vol. 272, no. 5261, p. 551, 1996.

[119] S. A. Sheth, M. Nemoto, M. Guiou, M. Walker, N. Pouratian, and A. W. Toga, "Linear and nonlinear relationships between neuronal activity, oxygen metabolism, and hemodynamic responses," *Neuron*, vol. 42, no. 2, pp. 347–355, 2004.

[120] J. Steinbrink, A. Villringer, F. Kempf, D. Haux, S. Boden, and H. Obrig, "Illuminating the BOLD signal: combined fMRI–fNIRS studies," *Magnetic resonance imaging*, vol. 24, no. 4, pp. 495–505, 2006.

[121] F. Scholkmann, S. Kleiser, A. J. Metz, R. Zimmermann, J. M. Pavia, U. Wolf, and M. Wolf, "A review on continuous wave functional near-infrared spectroscopy and imaging instrumentation and methodology," *Neuroimage*, vol. 85, pp. 6–27, 2014.

[122] G. Strangman, J. P. Culver, J. H. Thompson, and D. A. Boas, "A quantitative comparison of simultaneous BOLD fMRI and NIRS recordings during functional brain activation," *Neuroimage*, vol. 17, no. 2, pp. 719–731, 2002.

[123] I. Wechsung and K. De Moor, *Quality of Experience Versus User Experience*. Springer, 2014.

[124] M. Strait and M. Scheutz, "What we can and cannot (yet) do with functional near infrared spectroscopy," *Frontiers in Neuroscience*, vol. 8, 2014.

[125] R. Stern, W. Ray, and K. Quigley, *Psychophysiological Recording*. Oxford University Press, New York, 2 ed., 2001.

[126] M. Soleymani *et al.*, "A multimodal database for affect recognition and implicit tagging," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 42–55, 2012.

[127] R. Gupta, M. Khomami Abadi, J. A. Cárdenes Cabré, F. Morreale, T. H. Falk, and N. Sebs, "A quality adaptive multimodal affect recognition system for user-centric multimedia indexing," *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*, pp. 317–320, 2016.

[128] ITU-T, "Itu-t recommendation p.910 subjective video quality assessment methods for multimedia applications," tech. rep., International Telecommunication Union, Geneva, Switzerland, 2008.

[129] ITU-T, "Itu-t recommendation p. 911 subjective audiovisual quality assessment methods for multimedia applications," 1998.

[130] D. Kim and A. Tarraf, "ANIQUE+: A New American National Standard for Non-intrusive Estimation of Narrowband Speech Quality: Research Articles," *Bell Lab. Tech. J.*, vol. 12, pp. 221–236, May 2007.

[131] T. H. Falk and W. Chan, "A non-intrusive quality measure of dereverberated speech," in *Proc. Int. Workshop Acoust. Echo Noise Control,*, 2008.

[132] J.-N. Antons *et al.*, "Brain Activity Correlates of Quality of Experience," *Quality of Experience*, pp. 109–119, 2014.

[133] J.-N. Antons, "EEG Frequency Band Power Changes Evoked by Listening to Audiobooks at Different Quality Levels," *Neural Correlates of Quality Perception for Complex Speech Signals*, pp. 63–72, 2015.

[134] S. Arndt *et al.*, "The effects of text-to-speech system quality on emotional states and frontal alpha band power," *6th International IEEE-EMBS Conference on Neural Engineering (NER)*, pp. 489–492, 2013.

[135] L. R. Fabrigar, D. T. Wegener, R. C. MacCallum, and E. J. Strahan, "Evaluating the use of exploratory factor analysis in psychological research.," *Psychological methods*, vol. 4, no. 3, p. 272, 1999.

[136] R. B. Kline, "Exploratory and confirmatory factor analysis," *In Y. Petscher & C. Schatschneider (Eds.), Applied quantitative analysis in the social sciences*, pp. 171–207, 2013.

[137] H. F. Kaiser, "A second generation little jiffy," *Psychometrika*, vol. 35, no. 4, pp. 401–415, 1970.

[138] M. S. Bartlett, "Tests of significance in factor analysis," *British Journal of Statistical Psychology*, vol. 3, no. 2, pp. 77–85, 1950.

[139] A. B. Costello and J. W. Osborne, "Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis," *Practical Assessment, Research & Evaluation*, vol. 10, pp. 173–178, 2005.

[140] H. F. Kaiser, "The application of electronic computers to factor analysis.," *Educational and psychological measurement*, pp. 141–151, 1960.

[141] S. A. Mulaik, *The foundations of factor analysis*. McGraw Hill, 1972.

[142] B. M. Byrne, *Structural equation modeling with Mplus: Basic concepts, applications, and programming*. Routledge, 2013.

[143] B. M. Byrne, *Structural equation modeling with AMOS: Basic concepts, applications, and programming*. Lawrence Erlbaum Associates, Mahwah, NJ, 2001.

[144] Y. Rosseel, "lavaan: An R package for structural equation modeling," *Journal of Statistical Software*, vol. 48, no. 2, pp. 1–36, 2012.

[145] D. L. Jackson, J. A. Gillaspy Jr, and R. Purc-Stephenson, "Reporting practices in confirmatory factor analysis: an overview and some recommendations.," *Psychological methods*, vol. 14, no. 1, p. 6, 2009.

[146] G. W. Cheung and R. B. Rensvold, "Evaluating goodness-of-fit indexes for testing measurement invariance," *Structural equation modeling*, vol. 9, no. 2, pp. 233–255, 2002.

[147] R. P. Bagozzi and Y. Yi, "On the evaluation of structural equation models," *Journal of the academy of marketing science*, vol. 16, no. 1, pp. 74–94, 1988.

[148] P. M. Bentler and D. G. Bonett, "Significance tests and goodness of fit in the analysis of covariance structures.," *Psychological bulletin*, vol. 88, no. 3, p. 588, 1980.

[149] R. J. Vandenberg and C. E. Lance, "A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research," *Organizational research methods*, vol. 3, no. 1, pp. 4–70, 2000.

[150] J.-O. Kim and C. W. Mueller, *Factor analysis: Statistical methods and practical issues*, vol. 14. Sage, 1978.

[151] J. F. Hair, W. C. Black, B. J. Babin, and R. E. Anderson, *Multivariate data analysis*, vol. 7. Pearson Prentice Hall Upper Saddle River, NJ, 2009.

[152] Y. Inoue and A. Ikeda, *Event-related Potentials in Patients with Epilepsy: from Current State to Future Prospects*. John Libbey Eurotext, 2008.

[153] Y. Ku, B. Hong, X. Gao, and S. Gao, "Spectra-temporal patterns underlying mental addition: an ERP and ERD/ERS study," *Neuroscience letters*, vol. 472, no. 1, pp. 5–10, 2010.

[154] K. Tavabi, D. Embick, and T. P. Roberts, "Spectral–temporal analysis of cortical oscillations during lexical processing," *Neuroreport*, vol. 22, no. 10, pp. 474–478, 2011.

[155] R. Gupta and T. H. Falk, "PHYSYQX." Online, Feb 2016.

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

[157] DEAP, "A dataset for emotion analysis using EEG, physiological and video signals." `http://www.eecs.qmul.ac.uk/mmv/datasets/deap/`. [Online; accessed 10-August-2016].

[158] F. Boiten, J. Sergeant, and R. Geuze, "Event-related desynchronization: the effects of energetic and computational demands," *Electroencephalography and clinical neurophysiology*, vol. 82, no. 4, pp. 302–309, 1992.

[159] W. Klimesch, P. Sauseng, and S. Hanslmayr, "EEG alpha oscillations: the inhibition–timing hypothesis," *Brain research reviews*, vol. 53, no. 1, pp. 63–88, 2007.

[160] W. Klimesch, M. Doppelmayr, H. Russegger, and T. Pachinger, "Theta band power in the human scalp EEG and the encoding of new information.," *Neuroreport*, vol. 7, no. 7, pp. 1235–1240, 1996.

[161] W. Singer, "Synchronization of cortical activity and its putative role in information processing and learning," *Annual review of physiology*, vol. 55, no. 1, pp. 349–374, 1993.

[162] W. Singer and C. M. Gray, "Visual feature integration and the temporal correlation hypothesis," *Annual review of neuroscience*, vol. 18, no. 1, pp. 555–586, 1995.

[163] R. Eckhorn *et al.*, "Coherent oscillations: A mechanism of feature linking in the visual cortex?," *Biological cybernetics*, vol. 60, no. 2, pp. 121–130, 1988.

[164] V. N. Murthy and E. E. Fetz, "Coherent 25-to 35-Hz oscillations in the sensorimotor cortex of awake behaving monkeys," *Proceedings of the National Academy of Sciences*, vol. 89, no. 12, pp. 5670–5674, 1992.

[165] S. L. Bressler and J. Kelso, "Cortical coordination dynamics and cognition," *Trends in Cognitive Sciences*, vol. 5, no. 1, pp. 26–36, 2001.

[166] F. Varela, J.-P. Lachaux, E. Rodriguez, and J. Martinerie, "The brainweb: phase synchronization and large-scale integration," *Nature reviews neuroscience*, vol. 2, no. 4, pp. 229–239, 2001.

[167] J. Dauwels, F. Vialatte, T. Musha, and A. Cichocki, "A comparative study of synchrony measures for the early diagnosis of Alzheimer's disease based on EEG," *NeuroImage*, vol. 49, no. 1, pp. 668–693, 2010.

[168] I. Daly *et al.*, "Neural correlates of emotional responses to music: An EEG study," *Neuroscience letters*, vol. 573, pp. 52–57, 2014.

[169] S. Weiss and H. M. Mueller, "The contribution of EEG coherence to the investigation of language," *Brain and language*, vol. 85, no. 2, pp. 325–343, 2003.

[170] E. Bullmore and O. Sporns, "The economy of brain network organization," *Nature Reviews Neuroscience*, vol. 13, no. 5, pp. 336–349, 2012.

[171] M. Rubinov and O. Sporns, "Complex network measures of brain connectivity: uses and interpretations," *Neuroimage*, vol. 52, no. 3, pp. 1059–1069, 2010.

[172] V. Latora and M. Marchiori, "Efficient behavior of small-world networks," *Physical review letters*, vol. 87, no. 19, p. 198701, 2001.

[173] B. J. Baars, "Global workspace theory of consciousness: toward a cognitive neuroscience of human experience," *Progress in Brain Research*, vol. 150, pp. 45–53, 2005.

[174] T. M. Loughin, "A systematic comparison of methods for combining p-values from independent tests," *Computational statistics & data analysis*, vol. 47, no. 3, pp. 467–485, 2004.

[175] S. Daimi and G. Saha, "Classification of emotions induced by music videos and correlation with participants' rating," *Expert Systems with Applications*, vol. 41, no. 13, pp. 6057–6065, 2014.

[176] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher,

M. Perrot, and E. Duchesnay, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[177] K. Morton and P. Torrione, "Pattern Recognition Toolbox for MATLAB." `http://newfolder.github.io/`, 2013. [Online; accessed 8-August-2016].

[178] M. J. Brookes, A. M. Gibson, S. D. Hall, P. L. Furlong, G. R. Barnes, A. Hillebrand, K. D. Singh, I. E. Holliday, S. T. Francis, and P. G. Morris, "GLM-beamformer method demonstrates stationary field, alpha ERD and gamma ERS co-localisation with fMRI BOLD response in visual cortex," *Neuroimage*, vol. 26, no. 1, pp. 302–308, 2005.

[179] C. M. Krause, B. Pörn, A. H. Lang, and M. Laine, "Relative alpha desynchronization and synchronization during speech perception," *Cognitive brain research*, vol. 5, no. 4, pp. 295–299, 1997.

[180] J. Obleser and N. Weisz, "Suppressed alpha oscillations predict intelligibility of speech and its acoustic details," *Cerebral cortex*, vol. 22, no. 11, pp. 2466–2477, 2012.

[181] N. E. Crone, D. Boatman, B. Gordon, and L. Hao, "Induced electrocorticographic gamma activity during auditory perception," *Clinical Neurophysiology*, vol. 112, no. 4, pp. 565–582, 2001.

[182] S. K. Scott and I. S. Johnsrude, "The neuroanatomical and functional organization of speech perception," *Trends in neurosciences*, vol. 26, no. 2, pp. 100–107, 2003.

[183] E. Macaluso, N. George, R. Dolan, C. Spence, and J. Driver, "Spatial and temporal factors during processing of audiovisual speech: a PET study," *Neuroimage*, vol. 21, no. 2, pp. 725–732, 2004.

[184] W. Heller and J. Levy, "Perception and expression of emotion in right-handers and left-handers.," *Neuropsychologia.*, vol. 19(2), pp. 263–72., 1981.

[185] D. RJ., *Hemispheric specialization for cognition and affect.* Academic Press London, 1983.

[186] B. Reuderink, C. Mühl, and M. Poel, "Valence, arousal and dominance in the EEG during game play," *International journal of autonomous and adaptive communications systems*, vol. 6, no. 1, pp. 45–62, 2013.

[187] C. D. Gilbert and M. Sigman, "Brain states: top-down influences in sensory processing," *Neuron*, vol. 54, no. 5, pp. 677–696, 2007.

[188] V. Miskovic and L. A. Schmidt, "Cross-regional cortical synchronization during affective image viewing," *Brain Research*, vol. 1362, pp. 102–111, 2010.

[189] A. K. Engel and P. Fries, "Beta-band oscillation-signalling the status quo?," *Current opinion in neurobiology*, vol. 20, no. 2, pp. 156–165, 2010.

[190] D. Tucker, D. Roth, and T. Bair, "Functional connections among cortical regions: topography of EEG coherence," *Electroencephalography and clinical neurophysiology*, vol. 63, no. 3, pp. 242–250, 1986.

[191] J. B. Crabbe, J. C. Smith, and R. K. Dishman, "Emotional & electroencephalographic responses during affective picture viewing after exercise," *Physiology & behavior*, vol. 90, no. 2, pp. 394–404, 2007.

[192] 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.

[193] A. Blood and R. 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.

[194] 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.

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

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

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

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

[199] 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.

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

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

[202] Y. Xu, H. L. Graber, and R. L. Barbour, "nirsLAB: A Computing Environment for fNIRS Neuroimaging Data Analysis," *Biomedical Optics*, pp. BM3A–1, 2014.

[203] A. Villringer and B. Chance, "Non-invasive optical spectroscopy and imaging of human brain function," *Trends in neurosciences*, vol. 20, no. 10, pp. 435–442, 1997.

[204] K. L. Perdue, A. Westerlund, S. A. McCormick, and C. A. Nelson, "Extraction of heart rate from functional near-infrared spectroscopy in infants," *Journal of biomedical optics*, vol. 19, no. 6, pp. 067010–067010, 2014.

[205] C. Herff, O. Fortmann, C.-Y. Tse, X. Cheng, F. Putze, D. Heger, and T. Schultz, "Hybrid fNIRS-EEG based discrimination of 5 levels of memory load," *7th International IEEE/EMBS Conference on Neural Engineering (NER), 2015*, pp. 5–8, 2015.

[206] J. S. Richman and J. R. Moorman, "Physiological time-series analysis using approximate entropy and sample entropy," *American Journal of Physiology-Heart and Circulatory Physiology*, vol. 278, no. 6, pp. H2039–H2049, 2000.

[207] J.-S. Lee and C. H. Park, "Robust audio-visual speech recognition based on late integration," *IEEE Transactions on Multimedia*, vol. 10, no. 5, pp. 767–779, 2008.

[208] P. K. Atrey, M. A. Hossain, A. El Saddik, and M. S. Kankanhalli, "Multimodal fusion for multimedia analysis: a survey," *Multimedia systems*, vol. 16, no. 6, pp. 345–379, 2010.

[209] V. Miskovic *et al.*, "Changes in EEG cross-frequency coupling during cognitive behavioral therapy for social anxiety disorder," *Psychological Science*, pp. 507–516, 2011.

[210] M. Blackwell, "Multiple hypothesis testing: The f-test," *url: http://www.mattblackwell.org.s3-website-us-east-1.amazonaws.com/files/teaching/ftests.pdf [Online Accessed on 11th October, 2016]*, 2008.

[211] A. Sackl *et al.*, "Wireless vs. wireline shootout: How user expectations influence quality of experience," *Fourth International Workshop on Quality of Multimedia Experience (QoMEX)*, pp. 148–149, 2012.

[212] D. B. Pisoni, "Perception of synthetic speech," *Progress in speech synthesis*, pp. 541–560, 1997.

[213] M. Wester, C. Valentini-Botinhao, and G. E. Henter, "Are we using enough listeners? No! An empirically-supported critique of Interspeech 2014 TTS evaluations," in *Proc. Interspeech*, pp. 3476–3480, 2015.

[214] A. Clerico, R. Gupta, and T. H. Falk, "Mutual information between inter-hemispheric EEG spectro-temporal patterns: A new feature for automated affect recognition," *Proceedings of 7th International IEEE/EMBS Conference on Neural Engineering (NER)*, pp. 914–917, 2015.

[215] M. Kamiński, M. Ding, W. A. Truccolo, and S. L. Bressler, "Evaluating causal relations in neural systems: Granger causality, directed transfer function and statistical assessment of significance," *Biological cybernetics*, vol. 85, no. 2, pp. 145–157, 2001.

[216] D. Heger, R. Mutter, C. Herff, F. Putze, and T. Schultz, "Continuous recognition of affective states by functional near infrared spectroscopy signals," pp. 832–837, 2013.

[217] S. S. Pillay, S. A. Gruber, J. Rogowska, N. Simpson, and D. A. Yurgelun-Todd, "fMRI of fearful facial affect recognition in panic disorder: the cingulate gyrus–amygdala connection," *Journal of affective disorders*, vol. 94, no. 1, pp. 173–181, 2006.

[218] T. Penzel, B. Kemp, G. Klosch, A. Schlogl, J. Hasan, A. Varri, and I. Korhonen, "Acquisition of biomedical signals databases," *IEEE Eng Med Biol Mag*, vol. 20, no. 3, pp. 25–32, 2001.

[219] D. Tobon Vallejo, T. Falk, and M. Maier, "MS-QI: A Modulation Spectrum-Based ECG Quality Index for Telehealth Applications," *IEEE Transactions on Biomedical Engineering*, vol. PP, no. 99, pp. 1–1, 2015.