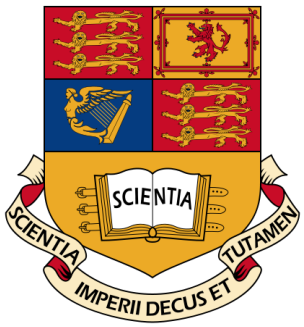
Imperial College London



Heterogeneous recognition of bioacoustic signals for human-machine interfaces

by

Michael Alec Vere Mace
Department of Mechanical Engineering

Supervised by
Dr. Ravi Vaidyanathan

A dissertation submitted to Imperial College London in accordance with the requirements for award of the degree of *Doctor of Philosophy* in the Faculty of Engineering

February 2013

Dedicated to the entire Mace clan, from oldest to youngest,
especially Mama and Pops.

Acknowledgements

First and foremost I would like to express my deepest appreciation to my supervisor Dr Ravi Vaidyanathan. I am grateful for the opportunity he gave me prior to carrying out this research and his continual guidance over the four years that I have known him. Without his constant enthusiasm, ideas and support this work would not have been possible. Special thanks are due to the rest of the research team, specifically, Dr Khondaker Abdullah-Al-Mamun and Dr Shouyan Wang who have provided invaluable discussions and co-operation in this project. Their technical work in terms of the interference rejection has been a key addition to the final real-time system. This research would not have been possible without the expertise of Dr Lalit Gupta and the provision of the generic earpieces from Think-A-Move Ltd, and for this I am extremely grateful.

I would like to acknowledge my friends and colleagues in the Mechanical Engineering department, both at the University of Bristol where I first started this research and at Imperial College London. They have provided an environment that has been both constructive and enjoyable. Thanks are also due to all the volunteers who participated in my testing, I hope that they have benefited from the added tongue dexterity they now own.

Lastly, I would like to offer my heartfelt gratitude to my whole family, the Mace clan, whose love, support and encouragement have made it all worthwhile.

Declaration of originality

I hereby certify that all material in this dissertation which is not my own work has been duly acknowledged. Selected work associated with this dissertation have been disseminated in scientific publications as detailed in Appendix D.

Michael AV Mace

26th February 2013

©

The copyright of this thesis rests with the author and is made available under a Creative Commons Attribution Non-Commercial No Derivatives licence. Researchers are free to copy, distribute or transmit the thesis on the condition that they attribute it, that they do not use it for commercial purposes and that they do not alter, transform or build upon it. For any reuse or redistribution, researchers must make clear to others the licence terms of this work

Abstract

Human-machine interfaces (HMI) provide a communication pathway between man and machine. Not only do they augment existing pathways, they can substitute or even bypass these pathways where functional motor loss prevents the use of standard interfaces. This is especially important for individuals who rely on assistive technology in their everyday life. By utilising bioacoustic activity, it can lead to an assistive HMI concept which is unobtrusive, minimally disruptive and cosmetically appealing to the user. However, due to the complexity of the signals it remains relatively underexplored in the HMI field.

This thesis investigates extracting and decoding volition from bioacoustic activity with the aim of generating real-time commands. The developed framework is a systemisation of various processing blocks enabling the mapping of continuous signals into M discrete classes. Class independent extraction efficiently detects and segments the continuous signals while class-specific extraction exemplifies each pattern set using a novel template creation process stable to permutations of the data set. These templates are utilised by a generalised single channel discrimination model, whereby each signal is template aligned prior to classification. The real-time decoding subsystem uses a multichannel heterogeneous ensemble architecture which fuses the output from a diverse set of these individual discrimination models. This enhances the classification performance by elevating both the sensitivity and specificity, with the increased specificity due to a natural rejection capacity based on a non-parametric majority vote. Such a strategy is useful when analysing signals which have diverse characteristics, false positives are prevalent and have strong consequences, and when there is limited training data available. The framework has been developed with generality in mind with wide applicability to a broad spectrum of biosignals.

The processing system has been demonstrated on real-time decoding of tongue-movement ear pressure signals using both single and dual channel setups. This has included in-depth evaluation of these methods in both offline and online scenarios. During online evaluation, a stimulus based test methodology was devised, while representative interference was used to contaminate the decoding process in a relevant and real fashion. The results of this research provide a strong case for the utility of such techniques in real world applications of human-machine communication using impulsive bioacoustic signals and biosignals in general.

Contents

Abstract **v**

Contents **vi**

List of Figures **x**

List of Tables **xiv**

Nomenclature **xiv**

1 Introduction **1**

1.1 Human-machine interfaces 1

1.1.1 Assistive communication 2

1.1.2 Volitional bioacoustic signals 3

1.1.3 Tongue-movement ear pressure concept 5

1.2 Motivation and focus of the thesis 6

1.3 Summary of contributions 7

1.4 List of Publications 8

1.5 Structure of the thesis 8

2 Overview of assistive Human-Machine Interfaces **10**

2.1 Approach to Literature survey 10

2.2 The importance of assistive technologies 11

2.3 Key drivers for comparison of HMI concepts 14

2.3.1 Feasibility and functionality 14

2.3.2 Usability and form 15

2.4 Description of interfacing technologies 16

2.4.1 Body input systems 16

2.4.2 Biofeedback devices 20

2.4.3 Gaze tracking systems 24

2.4.4 Brain-machine interfaces 29

2.4.5 Orally generated communication . . . 33
2.4.6 Tongue-operated systems . . . 38
2.5 HMI comparison space . . . 41
2.6 Bioacoustic signal capture . . . 49
2.7 Chapter summary . . . 53

3 Background to bioacoustic tongue-movement ear pressure signals 55
3.1 Physical and physiological description . . . 55
3.2 Previous work . . . 58
3.3 Bioacoustic characteristics . . . 60
3.4 System . . . 66
3.4.1 Hardware . . . 66
3.4.2 Earpieces . . . 67
3.4.3 Bioacoustic processing . . . 70
3.4.4 BIOacoustic HMI Software . . . 72
3.5 Off-line data sets . . . 73
3.5.1 Participants . . . 73
3.5.2 Experimental conditions . . . 73
3.5.3 Experiment and signal recording . . . 75
3.6 Chapter summary . . . 77

4 Bioacoustic signal extraction 78
4.1 Activity detection . . . 78
4.1.1 Background . . . 78
4.1.2 Short-term energy contour . . . 81
4.1.3 Thresholds . . . 82
4.2 Signal segmentation . . . 84
4.3 Real-time considerations . . . 84
4.3.1 Ring buffer . . . 84
4.3.2 Parsimonious computation of short-term energy . . . 86
4.4 Class-specific signatures . . . 87
4.4.1 Cross-correlation alignment . . . 87
4.4.2 Template creation . . . 88
4.4.3 Partial-tree implementation . . . 88
4.4.4 Robust template creation . . . 90
4.4.5 Evaluation of template creation procedures . . . 92
4.5 Signal standardisation . . . 94
4.6 Chapter summary . . . 95

5 Decoding bioacoustic activity **97**
5.1 Learning from examples . . . 97
5.2 Learning challenges . . . 99
5.3 Classifier evaluation . . . 100
5.3.1 Performance measures . . . 100
5.3.2 Off-line performance estimation . . . 102
5.4 Single channel classification model . . . 104
5.5 Feature extraction . . . 107
5.5.1 Autoregressive coefficients . . . 109
5.5.2 Discrete Fourier transform . . . 110
5.5.3 Discrete cosine transform . . . 113
5.5.4 Principle component analysis . . . 114
5.6 Feature selection . . . 119
5.6.1 Sequential feature selection . . . 122
5.6.2 Feature weighting filters . . . 123
5.6.3 Subset size selection . . . 125
5.7 Increasing stability during feature selection . . . 126
5.7.1 Weighting envelope . . . 126
5.7.2 Weighted sequential feature selection . . . 127
5.8 Class-specific classifier models . . . 129
5.8.1 Gaussian Bayes . . . 130
5.8.2 Decision fusion . . . 132
5.8.3 Matched filter . . . 134
5.9 Decision theory and expected loss . . . 135
5.10 Validation of proposed methods on TMEP data . . . 136
5.10.1 Bioacoustic discrimination strategies . . . 137
5.10.2 A note on statistical comparisons . . . 138
5.10.3 Effect of training set size and filtering on performance . . . 140
5.10.4 Comparison of discrimination strategies . . . 142
5.10.5 Comparison of feature selection methods . . . 146
5.10.6 Custom vs. generic earpieces . . . 150
5.10.7 Well vs. minimally trained subjects . . . 151
5.10.8 Discussion . . . 151
5.11 Ensemble framework . . . 154
5.11.1 Background to multi-expert committees . . . 154
5.11.2 Fusion by preference . . . 156
5.11.3 Heterogeneous implementation for multi-channel systems . . . 158
5.11.4 Diversity measures . . . 158

5.11.5 Study on the diversity of a dual-channel TMEP ensemble 162
5.11.6 TMEP fusion results 165
5.11.7 Discussion 171
5.12 Chapter summary 171

6 Real-time bioacoustic HMI based on TMEP signals 173
6.1 Interference rejection subsystem 174
6.2 Testing strategy 176
6.3 Results 178
6.3.1 Participants 178
6.3.2 Three-action classification 179
6.3.3 Four-action classification 181
6.3.4 Interference rejection testing 181
6.4 Discussion of results 189
6.5 Comparison to alternative interfaces 192
6.6 Chapter summary 194

7 Conclusions and future work 196
7.1 Thesis summary 196
7.2 Achievements 197
7.3 Conclusions 201
7.3.1 Summary of contributions 201
7.3.2 Suggested future directions 202
7.4 Final comments 204

A A note on the wavelet packet transform 205

B Comparison of discrimination strategies 208
B.1 AR results 208
B.2 DFT results 209
B.3 DCT results 211
B.4 PCA results 212

C Paper submitted to journal of neuroscience methods 215

D Publication list 238
D.1 Journal Papers (*Pending*) 238
D.2 Journal Papers (*Accepted*) 238
D.3 Conference Papers (*Accepted*) 239

Bibliography 241

List of Figures

1.1 The human-machine interface concept 2

2.1 Relevance map depicting content and dependencies of literature reviewed in this thesis 11
2.2 Examples of body input systems 17
2.3 Surface EMG electrode placement 21
2.4 Invasive EMG electrodes and their proximity to the nerves 23
2.5 Examples of optical gaze tracking systems 25
2.6 Surface EOG electrode placement 28
2.7 Relationship between spatial resolution, level of invasiveness and neural population involved for various neural signals 29
2.8 EEG electrode placement 30
2.9 A typical sip and puff device 34
2.10 Arrangement of the vocal organs inside the head and neck 35
2.11 Tongue-touch system 39
2.12 Tongue drive system (TDS) 41
2.13 HMI comparison chart: cost vs. information transfer rate 44
2.14 HMI comparison chart: health risks vs. robustness 45
2.15 HMI comparison chart: calibration vs. accessibility 46
2.16 HMI comparison chart: simplicity vs. perceptibility 47
2.17 HMI comparison chart: normal function vs. portability 48

3.1 Schematic of the earpiece inserted into the ear during TMEP capture 56
3.2 TMEP tongue actions 58
3.3 Example tongue signals 61
3.4 Example data sets of tongue action segment 62
3.5 Example data sets of aligned tongue action segment 63
3.6 Power spectral density (PSD) plots of the four actions 64
3.7 Average PSD plots of the four actions 64

3.8 Typical speech signal characteristics captured in the ear and in front of the mouth . . . 65
3.9 TMEP hardware block diagram . . . 66
3.10 Frequency response of microphone used during TMEP signal capture . . . 68
3.11 Custom earpiece design . . . 68
3.12 Generic earpiece design . . . 70
3.13 Generalised bioacoustic processing flowchart . . . 71
3.14 BIOacoustic HMI software frontend . . . 74
3.15 Raw TMEP signals . . . 76

4.1 Highlighting the activity detection process using a TMEP signal . . . 83
4.2 Representation of a circular buffer . . . 85
4.3 Efficient STE calculation . . . 86
4.4 Full tree decomposition during template creation . . . 89
4.5 Partial tree decomposition during template creation . . . 90
4.6 Robust template creation process . . . 91
4.7 Average jackknife standard error associated with the robust PPCCAT template creation process . . . 93
4.8 Average jackknife absolute bias associated with the robust PPCCAT template creation process . . . 94

5.1 Standard confusion matrix for multi-class classification . . . 101
5.2 Confusion matrix developed for a multi-class classification system with interference rejection capabilities . . . 102
5.3 The cross-validation procedure . . . 104
5.4 Stratified CV for arbitrary training subset size . . . 105
5.5 General classification approach incorporating template alignment . . . 106
5.6 Typical autoregressive coefficients associated with the TMEP data . . . 111
5.7 Typical Fourier magnitudes associated with the TMEP data . . . 113
5.8 Typical DCT coefficients associated with TMEP data . . . 115
5.9 The PCA transforms . . . 117
5.10 Typical PCA_{ucs} coefficients associated with TMEP data . . . 118
5.11 Typical PCA_{pcs} coefficients associated with TMEP data . . . 119
5.12 Typical PCA_{pci} coefficients associated with TMEP data . . . 120
5.13 Feature weighting envelope . . . 127
5.14 Example of a CD diagram . . . 140
5.15 eDC error surfaces: training size vs. filter cut-off . . . 141
5.16 Initial performance comparisons for various TMEP discrimination strategies 144

5.17 Critical difference diagrams for initial comparison of the various TMEP discrimination strategies . . . 145
5.18 Classification accuracy against number of features for various TMEP discrimination strategies . . . 146
5.19 Feature selection performance for the DFT_{real}, DCT, PCA_{pci} feature extraction strategies . . . 148
5.20 Critical difference diagrams for the various feature selection methods . . . 149
5.21 Final performance comparisons for various TMEP discrimination strategies . 152
5.22 Critical difference diagrams for the final comparison of the various TMEP discrimination strategies . . . 153
5.23 Example of learning a complex decision boundary using an ensemble of simple classifiers . . . 156
5.24 Heterogeneous multichannel ensemble framework based on preferential fusion . . . 158
5.25 Computing the diversity matrix . . . 161
5.26 The dual-channel TMEP interface concept . . . 162
5.27 Pairwise diversity measures between heterogeneous ensemble members . . 164
5.28 Average ensemble performance across different ensemble sizes using a single channel setup . . . 166
5.29 Average ensemble performance across different ensemble sizes using a dual channel setup . . . 167
5.30 Minimum error performance for different size ensembles . . . 169
5.31 Relative contributions of the heterogeneous ensemble members . . . 170

6.1 IR WPT power . . . 175
6.2 Real-time testing strategy . . . 177
6.3 Real-time confusion matrices for three class inter-action discrimination. Confusion matrices based on Fig. 5.2 and described in Section 5.3.1 . . . 180
6.4 Real-time confusion matrices for four class inter-action discrimination. Confusion matrices based on Fig. 5.2 and described in Section 5.3.1. . . . 182
6.5 Examples of segmented interference signals . . . 185
6.6 Interference testing results with three action training sets . . . 187
6.7 Interference testing results with four action training sets . . . 188
6.8 Real-time performance results for the individual subjects . . . 190
6.9 Average real-time performance results . . . 192
6.10 HMI comparison space showing TMEP against other HMI concepts . . . 193

A.1 WPT decomposition . . . 206

B.1 Initial performance comparisons for the AR discrimination strategies . . . 209

B.2 Critical difference diagrams for initial comparison of the AR discrimination strategies 210
B.3 Initial performance comparisons for the DFT discrimination strategies 210
B.4 Critical difference diagrams for initial comparison of the DFT_{real} discrimination strategies 211
B.5 Initial performance comparisons for the DCT discrimination strategies . . . 212
B.6 Critical difference diagrams for initial comparison of the DCT sUGC discrimination strategies 213
B.7 Initial performance comparisons for the PCA discrimination strategies 213
B.8 Critical difference diagrams for initial comparison of the PCA_{pci} discrimination strategies 214

List of Tables

2.1 Table summarising the pros and cons of the current state of art in assistive HMI systems 42
2.2 Table summarising the pros and cons of the current state of art in assistive HMI systems 43

3.1 Andrea electronics Pureaudio USB-SA soundcard specifications 67
3.2 Comparison of the custom and generic earpiece 70
3.3 Overview of all the participants utilised during off-line experiments 73

4.1 Description of fixed thresholds associated with volitional bioacoustic activity detection 82

5.1 Comparing the performance of filter and wrapper models 121
5.2 Discrimination strategies employed for decoding volitional bioacoustic signals 138
5.3 Custom vs. generic earpiece subjects based on statistical significant values . 150
5.4 Well vs. minimally trained subjects based on statistical significant values . . 151
5.5 The 2×2 diversity matrix 159

6.1 Description and content of the three levels of interference tested 183

Nomenclature

Abbreviations

ADC	Analogue-to-digital convertor
AGS	Acoustic gel sensor
AIC	Akaike information criterion
ALS	Amyotrophic lateral sclerosis
ANN	Artificial neural network
ANOVA	Analysis-of-variance
ANS	Autonomic nervous system
AP	Action potential
AR	Autoregressive
ASR	Automatic speech recognition
AT	Assistive technology
BCI	Brain computer interface
BCM	Bone conduction microphone
BMI	Brain-machine interface
CA	Classification accuracy
CD	Critical difference
CNS	Central nervous system
CV	Cross-validation
CWT	Continuous wavelet transform
DBS	Deep brain stimulation
DCT	Discrete cosine transform
DFC	Decision fusion classifier
DFT	Discrete Fourier transform
DoF	Degree of freedom
DST	Discrete sine transform

DTW	Dynamic time warping
DWT	Discrete Wavelet transform
ECG	Electrocardiography
ECoG	Electrocorticographic
eDC	Euclidean distance classifier
EEG	Electroencephalography
EMG	Electromyography
EOG	Electrooculography
ERD	Event-related desynchronisation
ERS	Event-related synchronisation
ET	Eustachian tube
FFT	Fast Fourier transform
FIR	Finite impulse response
GT3D	Gaze tracker 3D
HCI	Human-computer interaction
HMI	Human-machine interface
HMM	Hidden Markov model
ICA	Independent component analysis
iEMG	Intramuscular electromyography
IIR	Infinite impulse response
IMU	Inertial measurement unit
IQR	Interquartile range
IR	Interference rejection
ITR	Information transfer rate
kNN	K-nearest-neighbour
LED	Light-emitting diode
LFP	Local field potential
LIFE	Longitudinally implanted intrafascicular electrode
MA	Moving average
MAD	Median absolute deviation
MEA	Micro-electrode arrays
MEG	Magnetoencephalography

MFC	Matched filter classifier
MFCC	Mel-frequency cepstral coefficients
mFS	Multi-class F-score
MMG	Mechanomyography
mRmM	Mean rank mean magnitude
mRmRM	Mean rank mean rank magnitude
mRV	Mean rank variance
MS	Multiple sclerosis
MSE	Mean squared error
MUAP	Motor unit action potential
NI	Neural interface
OOP	Object-oriented programming
PCA	Principal component analysis
pci	Pooled class-independent
pcs	Pooled class-specific
PD	Parkinson's disease
pFS	Pairwise F-score
PNS	Peripheral nervous system
PoG	Point-of-gaze
PPCCAT	Partial pairwise cross-correlation averaging tree
PSD	Power spectral density
QoL	Quality of life
rMGC	Multivariate Gaussian classifier using robust statistics
RMSE	Root mean squared energy
ROC	Receiver operating characteristic
rUGC	Univariate Gaussian classifier using robust statistics
SCI	Spinal cord injury
SCP	Slow cortical potentials
SD	Standard deviation
sEMG	Surface electromyography
SFS	Sequential feature selection
sMGC	Multivariate Gaussian classifier using standard statistics

SN	Sensitivity
SNR	Signal-to-noise ratio
SNS	Somatic nervous system
SP	Specificity
ssVEP	Steady-state visually-evoked potentials
STE	Short-term energy
sUGC	Univariate Gaussian classifier using standard statistics
SVM	Support vector machine
sWPT	Stationary wavelet packet transform
TDS	Tongue drive system
TIME	Transverse intrafascicular multichannel electrode
TMEP	Tongue-movement ear pressure
TMR	Targeted muscle re-innervation
ucs	Unpooled class-specific
VEP	Visually-evoked potentials
WPT	Wavelet packet transform
WSFS	Weighted sequential feature selection
ZCR	Zero-crossing rate

Symbols

$\mathbf{x}$	Signal vector
$\mathbf{x_i}$	Signal instance i
$\mathbf{X}$	Matrix of column-wise signal vectors
$\mathbf{X_n}$	Column n of matrix $\mathbf{X}$
x_k	Scalar value of $\mathbf{x}$ at sample k
$\mathbf{s}$	Feature vector
E_n	STE at sample n
W_E	Length of STE calculation window
T_L	Lower energy detection threshold
T_U	Upper energy detection threshold
D_L	Lower duration detection threshold
D_U	Upper duration detection threshold

N	Number of samples
N_p	Number of signal patterns in data set
D	Number of features
$\hat{\mathbf{x}}$	Template aligned vector
$\bar{\mathbf{x}}_{1:N_p}$	Template created from pattern instances $[1, 2, \ldots, N_p]$
$\mathcal{C}^{(m)}$	Class-specific classifier model associated with class m
$\boldsymbol{\Theta}$	True statistic
$\tilde{\boldsymbol{\Theta}}$	Sample statistic
$\check{\mathbf{x}}$	Standardised signal vector
$\mu_{\mathbf{x}}$	Mean across vector $\mathbf{x}$
$\sigma_{\mathbf{x}}$	SD across vector $\mathbf{x}$
$\Sigma_{\mathbf{X}}$	Covariance matrix associated with data matrix $\mathbf{X}$
$\mathbf{w}^{\{j,p\}}$	Wavelet packet coefficients vector at level j and band p
m	Pattern class index
M	Total number of pattern classes
TN_m	True positive rate associated with class m
$\text{FN}^{-}_{m,c}$	False negative rate associated with an instance from class m classified as class $c \quad (c \neq m)$
FN^{+}_{m}	False negative rate associated with class m due to rejection
FP	False positive rate
SN^{-}	Sensitivity percentage based on FN^{-} only
$\text{SN}^{\pm}$	Sensitivity percentage based on FN^{-} and FN^{+}
SP	Specificity percentage
CA^{-}	Classification accuracy percentage based on SN^{-} and SP
$\text{CA}^{\pm}$	Classification accuracy percentage based on $\text{SN}^{\pm}$ and SP
K	Number of CV folds
$\mathcal{F}()$	Fourier transform
$\mathcal{D}()$	DCT transform
$\boldsymbol{\Phi}$	Basis matrix
A_k	k^{th} autoregressive coefficient
$\phi()$	General feature selection function
$\Omega()$	Feature weighting function

$\Psi()$	Feature subset size function
$\mathcal{R}$	Ranking order
ω_n	Weight associated with feature n
α_n	Classification accuracy associated with feature n
$\mathcal{F}_w()$	Feature weighting function
$\delta()$	Indicator function
Q	Number of WSFS CV iterations
ρ_d	WSFS contribution weighting associated with feature d
$d()$	Distance function
$\mathcal{L}$	Log-likelihood matrix
$\mathbf{L}$	Loss matrix
$D^{(u,v)}$	Pairwise diversity matrix between ensemble member u and v
$Q^{(u,v)}$	Pairwise Q-statistic metric between ensemble member u and v
$\rho^{(u,v)}$	Pairwise correlation metric between ensemble member u and v
$DA^{(u,v)}$	Pairwise disagreement metric between ensemble member u and v
$DF^{(u,v)}$	Pairwise double fault metric between ensemble member u and v

Chapter 1

Introduction

This thesis is an investigation into using volitional bioacoustic signals as command input for assistive human-machine interfaces. Motivating this investigation is the ability of bioacoustic signals to be sensed away from their generating source, thus leaving it free and uninhibited by hardware. This can lead to the creation of imperceptible, portable and completely non-invasive interfaces based on various input actions, such as tongue movements, vocalisations, mouth or teeth clicks and mechanomyography signals. Translation of these actions into useful machine commands requires effective algorithms to be developed, enabling accurate deciphering of the bioacoustic signals. As such, the predominant focus of this work has been the formulation and implementation of a real-time bioacoustic processing framework for extraction and decoding of this volitional activity. To highlight the feasibility of the framework, it has been demonstrated in real-time using tongue-movement ear pressure signals (TMEP) as volitional bioacoustic input.

1.1 Human-machine interfaces

A human-machine interface (HMI), as the name suggests, provides a link between human and machine. This allows for manipulation of peripheral devices in a manner intended by the user. The term HMI refers to the complete system, including mapping and decoding of user intention, allowing for a machine command to be generated. Figure 1.1 illustrates the HMI concept, highlighting the various ways the user can express their intention through either motor, physiological or nervous commands. Therefore, HMI can be thought of as mediating the interaction between human and machine [1]. This allows for effective control and manipulation of peripherals including, but not limited to, computers, domotics, robotic devices and wheelchairs. Many conventional HMI systems utilise human movement for direct machine manipulation. For example, the standard computer mouse, developed in the sixties, requires translational motions of the hand to direct a computer cursor [2]. Although its widespread and long-term use is testament to its efficacy, such

interfaces do not cater for the physically impaired.

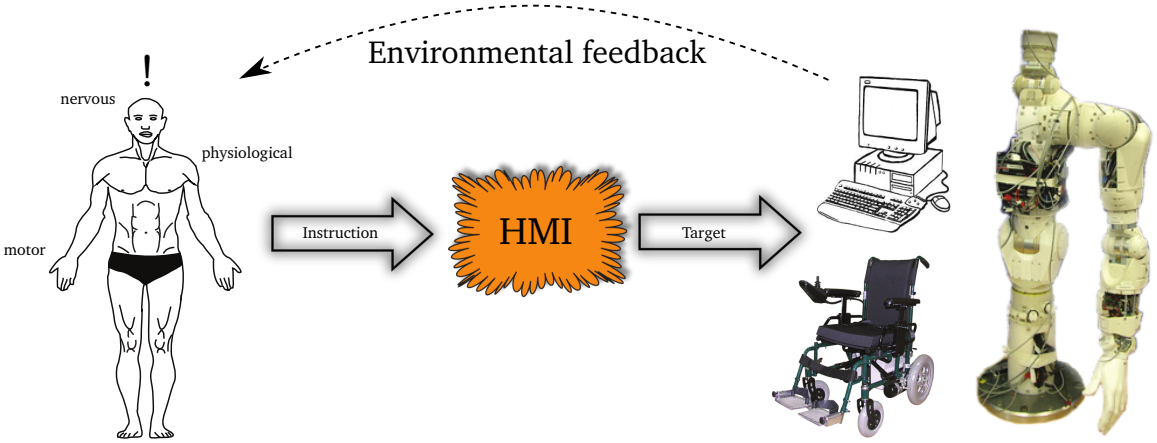


Figure 1.1: The human-machine interface concept

1.1.1 Assistive communication

Assistive HMI aims to restore the functional loss that can be incurred by physically challenged individuals, allowing them access to assistive technologies (ATs). It bridges the gap between their reduced capacity and the ability required to perform an activity, thus increasing their quality of life (QoL). This restoration can be achieved through either substitution or bypassing of the functional loss. While substitution aims at replacing the loss through an alternative form of interaction, bypassing the loss, aims at removing the reliance on the physical interaction by directly interfacing with the underlying physiological control mechanisms. Brain-machine interfaces (BMI) based on electroencephalographic (EEG) activity, for example, provide a basis for bypassing functional loss by directly decoding user intention from the neurological activity in the brain [3]. Within an assistive setting, a HMI can be considered a necessary subsystem for enabling technologies. For the purposes of this work, an assistive system will be characterised by (1) the device or AT being controlled, (2) the communication link through the user (human-sensor interaction) and (3) the HMI translating intention into machine commands. The interaction between human and AT is therefore at the core of technology aids [4]. The main focus in the design of these aids, is determining the method by which the user expresses their intention due to incapacitation or inhibition of normal communication routes. A predominant cause of this incapacitation is due to damage of the upper extremities or its underlying natural control system. This includes, but is not limited to, impairments such as stroke, spinal cord injury (SCI), paraplegia, locked-in syndrome, amyotrophic lateral sclerosis (ALS), muscular dystrophy, congenital limb deformities, cerebral palsy and arthritis [5–10]. Furthermore, the diversity of general disability has led to the development of the international classification

of functioning, disability and health (ICF) [11]. This categorises problems associated with human functioning into three interconnected areas, namely impairment, activity limitations and participation restrictions. As such, the predominant challenge for the designer is to replace standard communication pathways, for individuals with a diverse range of physical ability levels, in a manner that is both natural and effective.

Human-machine communication can be accomplished using a variety of human activities, including speech, vision, head movements or various human physiological and neurophysiological signals. The diverse nature of such activities leads to a wide variety of transducers that can be used to capture and transmit information, based on considerations of the human-sensor interaction. The various sensors can be grouped according to their operational mode and further categorised according to the way they function. For example, two predominant operational modes can be identified, namely switched-based control and proportional-type sensors [12]. Switch-based control expresses intention by switching between one of M polychotomous classes while proportional-type sensors generate an analogue output proportional to the expression of user intention. Depending upon the controlling activity, the manner by which an individual interacts with the HMI can also vary and includes direct manipulation or indirect gestural-based interaction. Direct manipulation of the interface requires the user to have physical contact with the sensor. For example, modifications have been made to the standard joystick design allowing it to be manipulated through physical movements of the head [13]. On the other hand, gestural-based interactions allow the human and sensor to be separated and are commonly captured using optical sensors [14]. These factors are only a few of the broad spectrum of design issues that contribute towards the useability and functionality of a HMI system. For a full discussion of these factors and the characteristics associated with the current state-of-art in assistive HMI, the reader is directed towards Chapter 2. Ultimately, this chapter demonstrates that, although there is a multitude of diverse alternative interfaces aimed at the physically challenged, no single device addresses all or even the majority of their needs, ability level and range of impairment. This highlights the importance of leveraging new methods for providing alternative communication pathways especially if their independence and QoL is to be maximised. Furthermore, new input methods can provide complimentary communication routes thus providing multi-modal control and extending their scope beyond that of just the disabled demographic.

1.1.2 Volitional bioacoustic signals

Volitional bioacoustic activity provides a feasible avenue for input into HMI and assistive communication systems in general. Bioacoustic signals themselves are a family of signals that are produced by the human body [15]. They are generated through localised pressure variations and bodily movements, creating sound waves in the surrounding tissue

and bone. The impedance characteristics between the source and surrounding anatomy will determine the acoustic conductivity associated with the resulting bioacoustic activity. These acoustic oscillations provide a completely noninvasive means of assessing human function, requiring only microphones or piezoelectric transducers positioned externally on an individual's body. This assessment has classically been used for diagnosis of medical conditions, relating for example, to the cardiopulmonary system or internal knee sounds [15, 16]. This is possible due to the audible output emitted from the human body which possesses valuable information regarding the physiology involved in the sound production mechanisms. The penetrative nature of the bioacoustic waves, implies that with suitable impedance matching, a signal with an effective signal-to-noise (SNR) ratio can be captured. Sensing of these signals is generally positioned some distance from the source of the sound, leading to the notion of disruption-free monitoring with regards to the physiology of interest.

Volitional bioacoustic signals are bodily sounds generated through intentional movements and can be considered a subset of bioacoustic signals themselves [17]. This activity is captured noninvasively on the surface of the body due to either internally or externally generated bodily movements. To date, the majority of bioacoustic signal analysis has generally involved the diagnosis of autonomically controlled physiological subsystems or nonvoluntary movement-related sounds. By capturing volitional activity and employing a signal translation architecture, it allows for deciphering of the individual's objective based on recognising different movements or intentions. This is the basis underlying HMI applications in general, with the potential of bioacoustic signals in this domain still very much under-explored. Volitional bioacoustic activity that has been explored by others include tooth clicks, MMG, finger tapping and vocalisations [17–19]. Although speech can be considered an intentional bioacoustic signal, in this work, it is regarded as a separate entity due to the production of external airborne sounds. For bioacoustic-based HMI concepts, the signals will generally be lower frequency when compared to airborne sounds. This is due to transmission across a bodily medium. To produce unique discernable signals, the morphology of the bioacoustic activity should be associated with a pattern signature. Impulsive bioacoustic signals are more likely to have a temporal-based pattern signature and have the additional benefit of a command set that would be quick to execute, thus allowing for efficient and direct control of peripheries. The scope of the volitional activity, can be produced through a wide variety of physiological interaction, for example, flicking, tapping or rubbing, between various anatomical locations.

The main benefit that bioacoustic signals provide is the ability for the sensor to be positioned away from the generating action. This can be especially important in the context of assitive HMI as, in many cases, the chosen access pathway is in a functionally important and/or sensitive location. For example, the oral and ocular systems provide

suitable anatomy for generating commands as they often remain functioning following disability. However, these areas are constantly required during daily living and therefore any sensory equipment located in or around them can interfere with their normal function. By utilising bioacoustic signals, volitional activity can still be generated from such areas without constraining the sensors to be placed in their vicinity. Instead, they can be positioned in a completely noninvasive, discreet and unobtrusive manner. This therefore leads to the potential for completely imperceptible assitive HMI concepts based on bioacoustic signals. Although conceptually simple, translation of bioacoustic signals presents a challenging task due to their intra-class variability, complex nonstationary dynamics and their subject-specific nature. There is also the risk of interfering signals imitating, merging and/or corrupting the volitional activity leading to potential confusion in the system. Such interference can be generated both internally from the user and externally from the environment, for example, from conversation or heart beats. Therefore, it is imperative to develop a robust and reliable processing framework when dealing with these types of signals. This will ensure that effective translation of volitional bioacoustic activity can be achieved in real-time.

1.1.3 Tongue-movement ear pressure concept

There has been much evidence about the robustness and continual functioning of the tongue when an individual experiences a severe motor impairment [20]. This is mainly due to the close proximity of the cranial nerves to the brainstem, with the hypoglossal nerve sending information to the tongues' motor units. Most injuries to the peripheral nervous system (PNS), either through physical damage or physiological degradation, generally occur between the spinal cord and extremities. The PNS consists of nerves and ganglia which extend out from the central nervous system (CNS), transferring information to and from the human sensory and motor units. A characteristic of the PNS (as opposed to the CNS) is that it is not protected by the skull, spine or the blood-brain barrier. This leaves it exposed to the possibility of mechanical damage and toxins. The twelve cranial nerves and their associated ganglia originate from the CNS with ten nerves extending from the brainstem and two from the cerebrum. These latter two are considered to be part of the PNS due to this extension, but are still well protected by the skull. Furthermore, the tongue is known to have rich sensory and motor cortex representation, rivaling that of fine finger motor control, such as that required during dextrous manipulation of small objects [21–23]. Based on these observations alone, employing the tongue as an alternative communication pathway is an attractive option.

A novel tongue-based HMI system based on bioacoustic signals has been previously introduced and conceptualised [24, 25]. The user expresses their intention by impulsive actions of the tongue, creating distinct signals within the ear canal. These have been

coined as tongue-movement ear pressure (TMEP) signals, due to the nature of their generation and evolution within the oral and auditory regions. The actions themselves involve placement of the tip of the tongue at the base of the central incisor, left or right first molar and flicking the tongue up (bottom, left or right action) and placing the tip of the tongue against the top of the palate and flicking down (top action). This instruction set was chosen as it directly relates to the fundamental control strategy of up, down, left and right whilst still retaining discernibility between the actions themselves. The uniqueness of the system, in terms of perceptibility and invasiveness, compared to alternative tongue-based concepts (see Section 2.4.6) is due to (1) the method and equipment utilised to capture the proposed tongue movements and (2) the location of this sensory hardware away from the mouth. By placing a microphone within the external acoustic meatus of an individual, the bioacoustic signal that is initiated by the tongue can be captured by the transducer due to the penetrative nature of the mechanical waves. These impulsive signals have distinctive waveform representations which are localised in both time and frequency.

1.2 Motivation and focus of the thesis

A broad range of assistive HMI devices have been designed and developed, thus helping to bridge the gap between the physically impaired and their environment. A multitude of interfaces exist in the research domain while others have made it to market. In spite of this significant progress, current techniques do not fully address the requirements of the user and better communication between individuals and assistive devices is still greatly expected. The diverse nature and range of physical impairments is indicative of the conflicting requirements expected from this demographic, with no single device satisfying all their needs. Many HMI solutions revolve around bodily movement for operation which may be difficult or even impossible for individuals with severe movement disabilities. Although important, the impairment itself is not the only driver in the design of assistive HMI, with many other factors influencing their usefulness. Other factors include bit rate, invasiveness, robustness, cosmetic appearance and perceptibility, to name a few. Many assistive HMI designs focus predominantly on impairment constraints and accuracy, without considering useability or form factors such as simplicity and perceptibility. Furthermore, these factors may actually be more important to the end user than functional drivers such as bit rate or even the objective efficacy [26]. It is therefore essential that innovative interfacing techniques are developed, with the aim of providing *practical* communication links between the physically challenged demographic and their environment.

Volitional bioacoustic signals can be used for control of assistive HMI systems. Utilising this type of activity addresses many of the issues associated with the useability and form of an assistive interface. Central to this is the concept of disruption-free signal capture

whereby, the sensors can be positioned freely away from the source of the activity. This is possible due to the penetrative nature of bioacoustic signals, allowing for a noninvasive and imperceptible solution to be developed. Furthermore, the concept is not constrained to a particular impairment model as the source of the activity can potentially be defined by the user.

This work focuses on the development of a real-time processing framework which will allow for effective translation of volitional bioacoustic signals into useful machine commands. This translation can be divided into extraction and decoding of the user's intention. Extraction deals with the detection and segmentation of volitional activity from the bioacoustic data and further exemplification of the signals associated with a particular intention expressed through motor actions. Bioacoustic decoding introduces various pattern recognition strategies for discrimination between the predefined actions and additionally provides rejection of potential interfering activity. This is achieved using concepts of feature spaces, feature selection, class-specific classification and heterogeneous ensemble methods.

Throughout the thesis, the proposed framework (and various aspects of it) are demonstrated on TMEP signals. The culmination is a real-time assistive HMI system based on the proposed bioacoustic processing framework with TMEP signals as command input. This highlights the potential for translation of volitional bioacoustic signals into useful machine commands using highly discreet and completely noninvasive hardware.

1.3 Summary of contributions

This thesis reports several technical contributions which are of particular interest to the area of biomedical signal processing. The following points summarise these contributions.

- Systematisation and development of an architecture for extraction and decoding of volitional bioacoustic activity with applications in the field of assistive HMI.
- Introduction of a new robust template creation process based on randomisation of a partial pairwise cross-correlation averaging tree (PPCCAT). This is a generalised extraction process used as a precursor to and during classification of volitional activity within HMI systems. It is specifically implemented for use in the bioacoustic TMEP application.
- Formulation and implementation of a multichannel ensemble framework based on heterogeneous discrimination models. This enhances classification performance while enabling natural interference rejection through preferential fusion strategies.
- Real-time implementation of the proposed bioacoustic architecture and evaluation on an imperceptible HMI system using completely noninvasive TMEP signals. This

included both three and four action discrimination and evaluation of the interference rejection capabilities under challenging conditions.

1.4 List of Publications

A list of publications associated with and submitted during the duration of this research is available in Appendix D. This includes journal papers which have been accepted, submitted or are in preparation, alongside a list of accepted conference papers.

1.5 Structure of the thesis

This thesis has been organised into seven chapters with a synopsis of the remaining six chapters given as follows.

Chapter 2 gives an overview of the current state of art in the assistive HMI domain. The various technologies are grouped based on the source of the commanding signal, namely, body, biofeedback, gaze, brain, oral and tongue based interfaces have been categorised. The major assistive technologies are compared across the groupings with these comparisons visualised using the concept of *key drivers* and HMI comparison charts. Systems utilising volitional bioacoustic signals are also described, with the scope of the concepts discussed not only constrained to the assistive HMI domain.

Chapter 3 provides background to the TMEP concept. This is in terms of their generation within the oral cavity, capture within the ear canal and their general bioacoustic characteristics. The process of collecting off-line TMEP data sets is outlined, with these data-sets used throughout the thesis for the purposes of validation. General BIOacoustic HMI software, developed to support this research, is also briefly introduced.

Chapter 4 describes the proposed bioacoustic extraction methods. The main focus of this chapter is activity detection, signal segmentation and creating robust class-specific templates. Due to real-time considerations, the activity detection and signal segmentation have been developed with the idea of parsimonious computation. Alignment of the segmented signals to class-specific templates is considered an important precursor to classification. Thus, a robust method for creating these exemplars has been developed, based on a partial pairwise cross-correlation averaging tree (PPCCAT).

Chapter 5 describes the bioacoustic decoding strategies. A general classification model is introduced and incorporates template alignment, feature extraction, feature selec-

tion and class-specific classification. Techniques for performing these various sub-processes are described and compared using single channel TMEP data and statistical analysis. Finally, a heterogeneous ensemble framework is introduced, increasing the decoding performance in terms of classification accuracy whilst introducing additional interference rejection capabilities into the system. Final evaluation of the ensemble is carried out using a dual-channel TMEP setup, with fusion among the ensemble members achieved using various voting and rank-based consolidation methods.

Chapter 6 presents full implementation and real-time testing of a dual-channel bioacoustic HMI system based on TMEP signals. This includes the incorporation of a specific interference rejection (IR) subsystem which although not work specifically carried out by the author, was developed collaboratively with members of a larger research team [27]. The real-time testing strategy is described and results are presented for both three and four action discrimination. Full off-line interference testing is performed using three data sets of increasing difficulty. This highlights the potential of the heterogeneous ensemble for inter-action classification while effective IR is demonstrated by combining the capabilities of both the IR and heterogenous ensemble subsystems.

Chapter 7 gives a general discussion of the research carried out, alongside conclusions and possible future work.

Chapter 2

Overview of assistive Human-Machine Interfaces

As discussed in the introduction, an assistive HMI is an assistive or augmentative communication technology. It replaces or bypasses a disabled individuals motor impairment so that they can perform actions that might otherwise be difficult or impossible [1, 12, 28]. These technologies are predominantly aimed at people who have little or no voluntary control of their extremities and are therefore not able to use standard able-bodied interfaces. Worldwide prevalence and diversity among such impairments has led to a rich body of research in this area. This chapter discusses the various technologies that have been explored to date, some available commercially and others in earlier stages of research and development. These have been grouped according to the type of instruction required from the user, allowing generalisations to be made. By defining *key drivers*, considered pertinent to both the design of an interface and needs of the user, the HMI technologies (and associated groups) can be reviewed and compared in a quantitative and relativistic manner. These are visualised through the use of two-dimensional HMI comparison charts which highlight the differences among the various devices in terms of these factors in a quick and intuitive manner. Also included on these charts are conventional able-bodied interfaces, so that the full spectrum of HMI technologies can be compared. The chapter concludes by discussing bioacoustic signal capture methods based on volitional input.

2.1 Approach to Literature survey

The author's approach to the literature survey is illustrated using the relevance map in Fig. 2.1. In this chapter, various assistive HMI technologies are explored and grouped according to input method. Additional short reviews relating to extraction and decoding techniques are presented in Chapters 4 and 5. This includes methods associated with ac-

tivity detection, feature extraction, feature selection and pattern recognition and are given as and when such methods are introduced within the bioacoustic processing framework.

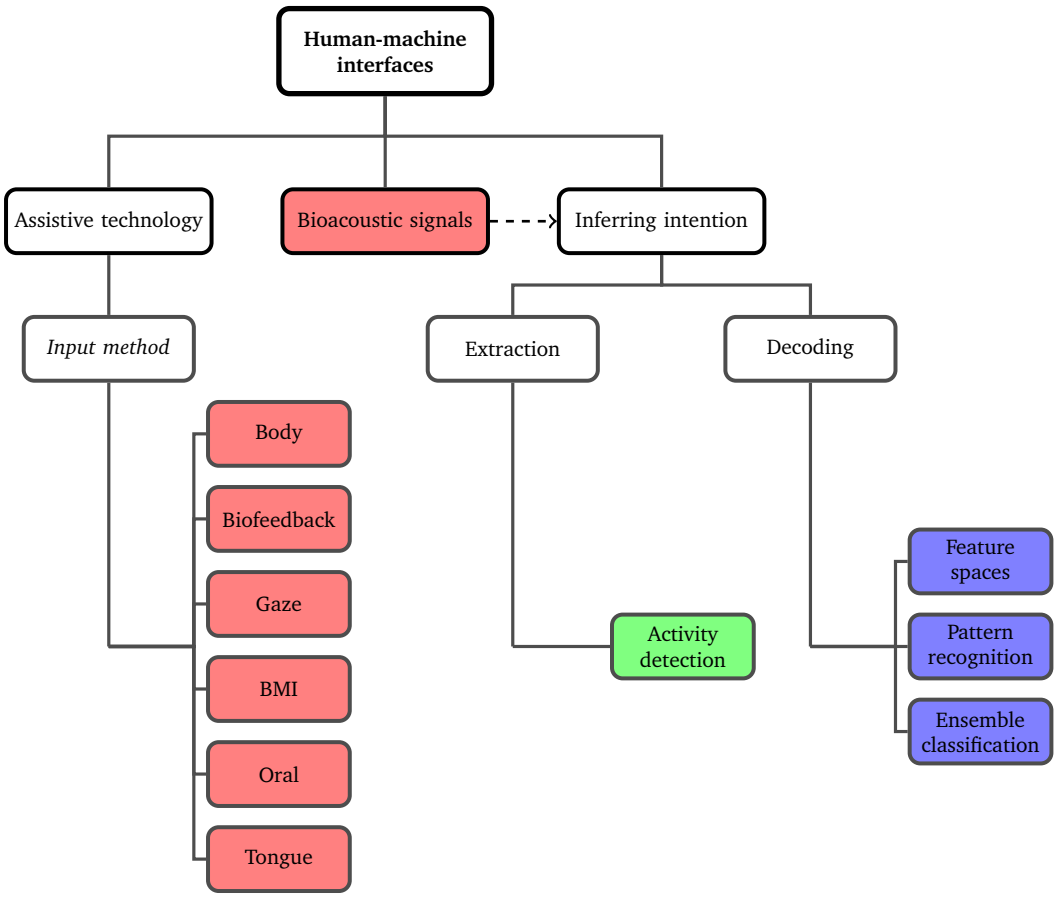


Figure 2.1: Relevance map depicting content and dependencies of literature reviewed in this thesis. The root of the literature survey is based around the topic of human-machine interfaces. This is predominantly done in the context of assisitve HMI and are grouped according to their input method. Bioacoustic signals are reviewed in the context of volitional activity and influences the type of extraction and decoding algorithms reviewed. The colours indicate which chapter the survey can be found: Chapter 2 (*this chapter*) - red, Chapter 4 - green and Chapter 5 - purple.

2.2 The importance of assistive technologies

According to the Assistive Technology Act and ISO9999, an AT device is 'any item, piece of equipment, or product system, whether acquired commercially, modified, or customised, that is used to increase, maintain, or improve functional capabilities of individuals with disabilities' [4, 29]. There are a plethora of disabilities that affect a multitude of people worldwide, with many specific diseases having a wide range of symptoms. Take for exam-

ple, Parkinson's disease (PD), a degenerative neurological condition which is idiopathic in nature. Although its symptoms can be put under the umbrella term of Hypokinesia, they include a spectrum of motor disorders ranging from tremor and rigidity to gait freezing, to name but a few. This is without even mentioning, additional non-motor defects which can occur coincidentally, for example, anxiety, autonomic dysfunction and cognitive impairment [30]. Central to the usability of assistive technologies is the interaction between human and AT. The HMI concept can be considered the medium through which communication flows between user and device [4]. Thus to fully leverage the support that ATs provide, the interfaces need to bypass or replace the dysfunction in a relevant and intuitive manner. The development of novel communication methods provide an alternative means by which the physically challenged can interact, when conventional able-bodied interfaces can no longer be used.

Conventional interfaces include the computer mouse, computer keyboard, standard joystick and RF controllers to name a few. They have been around for decades and as such, are highly optimised and provide an efficient means of communication [2]. However, they all have one thing in common, that is, require both movement and control of at least one upper extremity. As touched on earlier, there are a wide variety of diseases that affect upper body extremity control with a broad range of pathologies. For instance, some people are born with congenital defects, other disabilities can occur acutely due to either accident or sickness. Degenerative diseases, on the other hand, deteriorate the structure or function of the body progressively over time.

Congenital abnormalities affecting upper body motor control include failure of formation, fusion and duplication of parts of the extremity [5]. The most serious being transverse deficiencies which include all congenital amputations and result in the entire extremity not forming. Acute abnormalities affecting the upper extremities are caused by physical trauma, either damaging the limb itself and/or its underlying control physiological circuitry. In many cases, the damage to the limb will be severe enough that a surgical amputation will have to be performed, with either a wrist, elbow or shoulder disarticulation or transradial/transhumeral amputation being required [31]. Lesions of the PNS and/or CNS can cause paralysis and loss of muscular tone and function. Depending on the extent of damage, this can range from loss of arm control, for example, Erb's palsy[1] to complete loss of motor control below the neck. In this most serious case, it is generally caused by high-level SCI and is commonly known as quadriplegia. SCI accounts for 23% of paralysis in the US alone, while 29% is attributed to stroke and a further 17% to multiple sclerosis (MS) [6]. Stroke is another acute disease caused by rapid loss of neurological function due to disruption of the blood supply going to and from the brain [7]. It can be attributed to either clotting or bleeding within the brain. On the other hand, MS is a

[1]Erb's palsy is the paralysis of the arm caused by injury to its main nerves

degenerative and inflammatory disease where the myelin sheaths around the axons of the brain and/or spinal cord are damaged [8]. Other degenerative diseases which affect motor control include the aforementioned PD, arthritis and ALS, to name a few. Complete loss of motor function is termed locked-in syndrome and although the patient has full cognitive function, there is complete paralysis of all or nearly all voluntary muscles [9, 10].

Despite the significant progress made towards constructing mechanisms that will assist individuals with said motor deficiencies, the creation of enabling devices allowing them to properly integrate with and control ATs remains a significant issue. The diversity of the symptoms and degree of impairment implies that no single device will ever satisfy the needs of the entire demographic. Not only is it this diversity, but also the prevalence of the diseases themselves, which make it such an active and essential research field. In a survey of 59 countries, the world health survey reports that 11.8% and 18% of the global population can be considered to have some kind of disability within both higher and lower income countries [11]. Just looking at those considered significant, still produces a staggering estimate of 2.2% across all 59 countries. Considering just the UK, it is estimated that about 8.6 million adults in the years 1996-97 had a disability that, 'affected their ability to perform normal activities due to impairment of a structure or function of the body or mind' [32]. While in 2005, it was estimated that about 35 million people had a severe disability in the United States alone. Not only are these numbers already extraordinary, they are also on the increase. For instance, in the last 100 years the average human lifespan has been increasing at a rate of 2.2 years per decade. Thus by the year 2074, it is predicted that the population in the UK, over the age of 65, will have tripled [33]. Between 1991 and 2001, life expectancy in the UK has also risen by 2.2 years while healthy life expectancy has only increased by 0.6 years. This implies that people will have experienced ill health for an additional 1.6 years of their life, over the period of only one decade[33].

To clearly highlight the huge demand for AT, specific statistics associated with the aforementioned motor syndromes are now presented. In the UK there has been more than a 15% increase in the number of new strokes since 1999. Just considering England, over 110,000 new stroke cases are reported each year with a quarter of these occurring in people under the age of 65. Consequently, over 300,000 people in England are now living with moderate to severe disabilities as a result of stroke [7, 34]. The prevalence of limb loss in the United States equates to approximately 1.9 million people [35]. There is over 5 million Americans reporting some form of paralysis, with paralysis defined as, 'a CNS disorder resulting in difficulty or inability to move an upper or lower extremity' [6]. Of this number, it is indicated that 1,275,000 are specifically living with SCI. China has reported the highest incident of SCI cases, than anywhere else worldwide, nearing 60,000 new cases a year [36].

Fully motivating the need and requirement for assistive technologies and interfaces is beyond the scope of this thesis. However, by comprehension of these few statistics, it can be seen that there are a vast number of individuals living with severe physical disability and/or motor impairment. Therefore, until researchers develop medical treatments or procedures to remedy such conditions, there will remain a significant societal need for AT to improve the lives and independence of this demographic. By increasing their independence, it will relieve and reduce associated health and home-care costs, which is considered essential to the vast majority of these individuals and those around them.

2.3 Key drivers for comparison of HMI concepts

An assistive HMI is designed with the primary goal of restoring or augmenting communication between a disabled person and assistive devices. Although the main distinction between the various interfaces is the method by which the command is generated, there are many other pertinent factors that dictate whether a HMI system will be adopted. These input methods have been identified as body, biofeedback, gaze, brain, oral and tongue (see Fig. 2.1). When dealing with multiple factors, there is generally a trade-off or compromise required between them with, no single concept satisfying them all. By establishing these additional factors as *key drivers* within the assistive HMI design space and grouping them according to the input method provides a useful way for evaluating the various HMI concepts. From both the perspective of the designer and user, it provides an intuitive and quick method for comparing and quantifying the differences between the specified interfaces. Ten *key drivers* have been defined and are a compilation of factors that have been highlighted individually in other work [1, 26, 37, 38]. It should be noted that these drivers have been specifically defined by the author and are considered the most pertinent factors when matching an interface to a user. They have been grouped into two categories, namely, four drivers associated with the feasibility and functionality of a HMI concept and six relating to their usability and form. Furthermore, by rating each concept and associated HMI category in terms of these *key drivers*, they can be quantified and visualised within this relative comparison space.

2.3.1 Feasibility and functionality

These factors are related to the performance-cost trade-off of a given HMI device. They are generally considered independent of the end-user, as-well as being less subjective. The performance measures include information transfer rate (ITR) and robustness while the costs are associated with both financial and health factors.

Information transfer rate (ITR) This is defined as the number of bits of information (or

throughput) that the HMI can convey between the human and target device per second. It can be thought of as a measure of the channel capacity of the HMI communication link and is generally stated as a maximum achievable value.

Cost This is the financial cost of the overall HMI system and treatment, including the price of the interface itself and, if required, the cost of having the system clinically or surgically fitted.

Health risks This encapsulates many health factors including surgical risks associated with the clinical fitting of the HMI, if and when surgical intervention is a requirement. It also incorporates hygiene risks associated with its daily use and other general health risks that can potentially occur during the normal operation of the interface.

Robustness Persistence of the HMIs characteristic behaviour under perturbations or conditions of uncertainty. The ability of the HMI system to perform at or near its maximum performance level under changing environmental conditions and in a variety of potentially unforeseen scenarios.

2.3.2 Usability and form

These factors are related to operability-morphology-accessibility trade-offs for a given HMI device. Although these are more subjective measures and in part depend on the end-user, relative quantifications can be made and provide useful comparisons among the devices, in terms of these proposed drivers. These factors are considered as important as an interface's feasibility and functionality. It has been identified in the literature that, 'Individual patients may select future BCI systems for practical, cosmetic, or other reasons independent of bit rate or even objective efficacy' [26]. While BCI specifically relates to brain interfaces, the same idea readily extends to all types of HMI systems. The measures associated with the operability of a device include calibration, simplicity and normal functioning, while the morphology measures relate to perceptibility and portability factors.

Calibration The difficulty associated with system setup each time an individual wishes to use the device. This may also include the requirement of routine calibration steps to ensure proper functioning of the device.

Accessibility The proportion of the impaired populace that have the ability to use a particular HMI device. This is based solely on the individual's physical/neurological ability, regardless of other factors such as cost and access to treatment. It will generally be the case that the entire able-bodied demographic will be able to use any HMI concept while certain debilitating conditions will mean that the use of a particular instance or class of interfaces will be prohibited due to the physical limitations

of the individual. The accessibility of a specific interface can also change based on the flexibility of the instruction set, for example, which muscle group are employed within an EMG-based concept. The higher the flexibility of this instruction set the more accessible the interface will tend to be.

Simplicity The ease with which an individual can learn to use the interface. This includes factors such as the intuitive nature of the instruction set, its user-friendliness and the need for a care-holders intervention during device donning and operation.

Perceptibility How noticeable the HMI's hardware is to both the user and those in the vicinity of the user, once the device has been donned. This includes factors relating to its comfort, aesthetics and obtrusiveness. With respect to those around the user, this will be based on how much unwanted attention the device brings to the individual. A key factor in the comfort associated with a device, is whether bodily contact is required by the sensors.

Normal Functioning The inhibition of a user's daily functioning while the HMI is being operated. For example, does it interfere with an individual's ability to converse, see, eat or perform simple tasks. This includes allowing for shared conscious (or somatic) outputs from the user, implying that other tasks can be concurrently performed while the interface is being worn and operated.

Portability The ease with which a HMI technology can be transported from place-to-place and more importantly, the ability of it to be used whilst on the move.

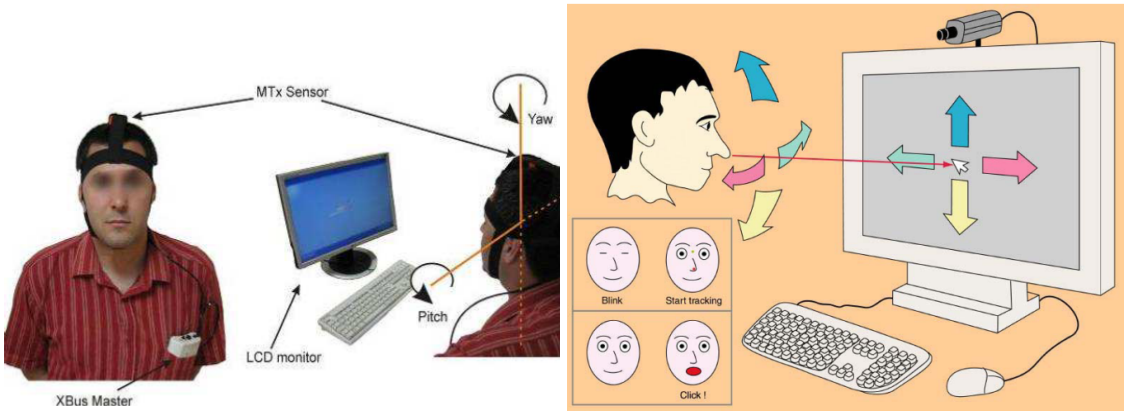
2.4 Description of interfacing technologies

In this section, various assistive communication technologies are introduced and described. They have been grouped into the following categories based on the method of user input, namely, body, biofeedback, gaze, brain, oral and tongue -based inputs. This allows for generalisations to be made as, in many cases, the input method dictates the device design. Due to the bio-acoustic nature of TMEP signals, an additional category discussing bio-acoustic signals and interfaces which utilise these signals has also been included.

2.4.1 Body input systems

Body input systems can be thought of as a generalisation of the classical method of interfacing man and machine through physical interaction. In terms of assistive HMIs and interface manipulation specifically, rather than bypassing the loss of function or dexterity in the extremity, body based systems aim at substitution of their associated function by utilisation of an alternative and functioning part of the body. As such, the choice and

scope of body parts that can potentially be used is extremely wide-ranging. This generally implies that specific systems are aimed at individuals who still have complete (or near complete) control of the proposed body area. In this section, a body part is considered any remaining functioning appendage including the head and trunk of an individual. A body input system is therefore based on the detection of voluntary physical motion of a specific anatomical body part and can be further partitioned into two disjoint subgroups, namely, mechanical input devices and body tracking systems. These subgroups have been differentiated by whether or not physical contact is required between user and interface hardware. Mechanical input systems require physical contact and thus movement of the interface, with the transducer mechanism either grasped, held or attached to the subject, thus allowing for a resultant force vector to be exerted on the sensor. Body tracking systems in contrast, require no physical contact with the device and instead are based on indirectly tracking movement or gestural based commands, usually via some kind of optical sensor. Figure 2.2 shows an example from these two subgroups of body input systems; A mechanical input device based on inertial measurement of the head (Fig. 2.2a) and a body tracking system based on facial recognition using an optical sensor (Fig 2.2b)[39, 40].



(a) IMU head input concept [39] (b) Optical face tracking concept [40]

Figure 2.2: Examples of body input systems

Mechanical input systems

Considering HMI systems in general, conventional interfaces, such as, the computer mouse, computer keyboard, standard joysticks and various hand-operated controllers would actually fall under the umbrella of mechanical-input systems [2, 41]. However, as they are not actually substituting a loss of functional movement and for the majority are aimed at able-bodied users, they are considered as a separate subset, namely, conventional interfaces. Having said this, their wide spread use and longevity are a testament to their performance with many assistive mechanical-input based systems adapted from these conventional in-

terface designs. A reason for this adaptation is the fact that many of these interfaces are aimed at a specific target device, for example, controlling a computer using a mouse-like interface or navigating a powered wheelchair using joystick-like interface. One such modification is to the standard joystick design, allowing an individual to use their head or chin to move the input lever, thus providing users who have limited extremity movement to control target devices using this analogue input [13]. Although its widespread and commercial use is a clear testament to its durability and effectiveness, these types of devices require constant bodily movements. This can be tiring or uncomfortable to the user, as well as the potential for frequent skin contact to cause irritation. Furthermore, the single lever design limits the allowable command set that can be sent and additionally needs attachment to a fixed base thus reducing its portability.

Leveraging of head motion, is one of the most widely used methods of substituting loss of upper extremity control [39, 42–45]. The main reason for this, is that volitional control of the motion of the head is through the spinal accessory nerve, one of the twelve pairs of cranial nerves which emerges from the cerebrum and brainstem. As such, any damage to the spinal cord or remaining PNS, will generally not affect one's ability to move their head. There are two main ways in which mechanical-input systems, utilising motions of the head, are measured. This is either through inertial measurements or by tracking an emitter. Inertial measurements are made using an inertial measurement unit (IMU) which is attached to the head of the individual, with signals from accelerometers, gyroscopes and/or magnetometers fused to provide an estimation of dynamic head posture [39, 42]. By attaching a transmitter to the head of an individual and placing the corresponding receiver(s) some distance in front of them, an alternative estimation of the direction and pose of the head can be obtained. Infra-red sensors are the most commonly used transmitter-receiver pairings, however ultrasonic and light-emitting diodes (LED) have also been used in the past [43–46]. The IMU offers a significant advantage over the emitter-based system, in the fact that no receiver is required and thus the device can be used in an ad-hoc manner. This is at the expense of a reduction in positional accuracy due to issues associated with integral drift. Furthermore, both types of interface require their hardware to be firmly secured to the apex of the chosen appendage and in the case of the emitter-based device have clear line of sight with the receiver. This leads to systems which can be obtrusive and highly perceptible to both the user and those around them.

Although assistive exoskeletons were primarily envisioned for active rehabilitation, they can also be considered a form of assistive interface, albeit while integrated into a human-robotic system [47, 48]. Along with instrumented gloves and similar technologies, they track the movements of an appendage through an 'external skeleton', which maps its position and thus the position of the appendage in a proprioceptive-like manner [49, 50]. An exoskeleton is made up of a number of degrees of freedom (DoF), which loosely match

the kinematics and anatomy of the appendage. Each DoF is sensorised so that the position of the exoskeleton is known in regard to its joint space. This is generally achieved using potentiometers at each joint (or DoF), however flex sensors can also be utilised, making for an overall more compliant mechanism [50]. The majority of exoskeletons are targeted at either the upper or lower limb extremities, with a possible application being mirrored control of a prosthetic limb for unilateral amputees. Considering these systems in the context of just man-machine communication, the extremely bulky and obtrusive nature of such devices seems unwarranted, unless sophisticated and physiologically relevant movement of the assistive end-effector is required.

More recently, a device nick-named the 'Komekami' switch has been developed and revolves around a fixed transducer attached via a headband to an individual's temporal region. It utilises an optical distance sensor to track changes associated with the tensing of various facial muscles [51]. Although primarily aimed at the wearable computing market, its usability could easily be extrapolated to the assistive interface sector due to its hands-free operation. The main source of command input is based round a clenching of the jaw for a duration of one second. This causes a contraction of the temporal muscles which are subsequently monitored by the 'Komekami' switch, allowing it to be differentiated from routine and normally occurring movements. Although cheap, affordable and discrete, when considering it in the assistive technology market it bears some limitations, namely the relatively long execution time necessary to issue a command, the ability to produce only a single command and the aesthetics associated with donning and doffing of the headband.

Body tracking systems

The majority of body tracking systems use optical-based hardware and thus in many ways are related to gaze operated systems. The camera mouse is one such system and has specifically been designed as an assistive computer interface [14]. It can track various body parts defined by the user, for example, the tip of the nose, using only a cheap web-cam and associated computer for real-time processing. Initially, a template of the proposed feature is manually cropped from an image of the individual. It then tracks the position of this feature online, by finding the position of maximum correlation between the defined template and subsequent video frames. The selection of a template feature can reduce the useability associated with the tracking system and therefore alternative face tracking systems which don't necessarily require a template to be created a priori have been developed [40]. Instead they utilise common facial features, such as, distance between the eyes and nose, shape of the mouth and orientation of the head to estimate the facial pose of the user. Although optical systems can additionally be utilised to track different body parts, head tracking is considered to be the most pertinent in terms of accessibility. Alter-

native body tracking interfaces include 'an adaptive non-contact gesture-based system for augmentative communication' and is based on the detection of motion of an anatomical site through the analysis of skin speckle [52]. The skin speckle pattern is associated with the scattering of reflected laser light, due to its surface roughness. As the reflecting surface moves, the speckle pattern moves proportionally and thus motion of the surface can be estimated by tracking the reflected speckle pattern. Although body tracking systems alleviate the requirement for the hardware to be placed in contact with the skin and is instead positioned at a fixed datum away from the user. This implies that a clear line of sight between user and sensor is required, with further consequences being a system which has portability issues. This means that without considerably set-up effort and potential re-calibration of such systems, they can only really be used in a single situation, for instance, interaction with a single computer.

2.4.2 Biofeedback devices

Biofeedback is the process of monitoring and quantifying various physiological functions, through the use of specialist electrodes and instrumentation. In many cases, the physiological activity being measured originates from the peripheral nervous system, generated both somatically or autonomically and is measured with a diverse spectrum of sensors [53]. The autonomic nervous system (ANS) is the part of the PNS which deals with control of bodily functions generally occurring under the level of consciousness. On the other hand, the somatic nervous system (SNS) deals with control of voluntary and intentional movements usually via stimulated muscle contractions. One of the most clinically relevant biofeedback systems is the electrocardiogram (ECG), which non-invasively measures heart rate via its electrical activity and is controlled through the ANS [54]. Other processes that can be monitored include muscle tone, skin conductance and blood pressure, to name but a few. Biofeedback systems are generally utilised for medical diagnosis or analytical purposes, usually within the clinical environment. With respect to assistive HMI, the biofeedback systems of interest are those quantifying somatic behaviour, as a HMI command is expressed through volitional and intentional interactions. Of these the most widely researched is electromyography (EMG), which measures the electrical signal associated with muscular activity [12, 55, 56]. The origin of such commands is from the CNS which sends a muscle contraction signal, propagated through the PNS, to motor neurons which control a group of muscle fibers. The recorded EMG signal is the spatial summation of motor unit action potentials (MUAP), which fire repeatedly to maintain muscle tone during a contraction. The EMG signal is recorded via an electrode placed in close proximity to the muscle whose activity is of interest. This can be done either non-invasively on the surface of the skin or intramuscularly, with the trade-off being spatial resolution of the bio-signal versus the financial and health costs associated with the surgical procedure

which include risks of post-operative infection [57].

Surface electromyography (sEMG) electrodes are by far, the most common method of recording EMG signals. This is due to their main advantages of being non-invasive and easily adaptable. However, this is counteracted by several disadvantages including, the need for daily placement, frequent calibration and the low reproducibility and quality of the acquired signals [57]. The choice and number of muscle sites to be used as control input is dependent upon the designer, with the decoding algorithm based on either specific or arbitrary locations. Further to this, the control input can be with or without actual bodily movement, through either concentric or isotonic contractions. As such, only the nerves and muscle fibers need to be intact for an EMG signal to be recorded. Real-time processing of the EMG signal include the hardware sub-processes of amplification, filtering and sampling, and software sub-processes of segmentation, feature extraction and classification [56]. Due to their comparative simplicity, time domain features are the most commonly used in myoelectric classification. These are generally based on time-varying amplitudes of the signal, which are considered proportional to the number of active motor units and rate of their activation. These types of feature give an overall indication of signal energy, activation level, duration and force associated with a particular muscular contraction. One of the most common applications of sEMG (and EMG in general) interfaces is for command and control of powered wheelchairs [58–60]. This has included utilising specific muscle groups such as the wrist extensors and flexors, as well as the use of arbitrary muscle groups as control input [58, 60].

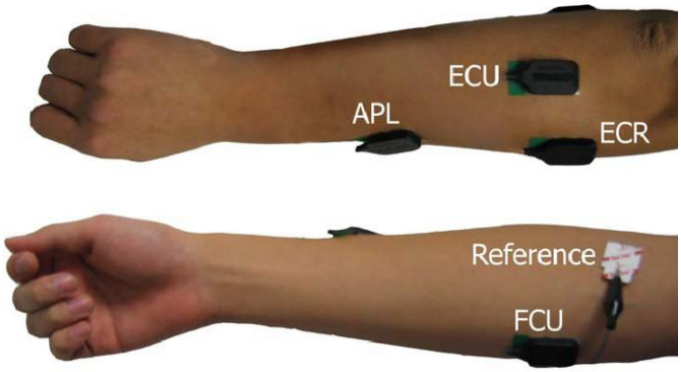


Figure 2.3: Examples of surface EMG electrode placement on a healthy arm [61]

More recently, intramuscular electromyography (iEMG) electrodes have been investigated in the context of assistive and augmentative communication [62, 63]. This has for the most part, been in the context of prosthesis control which due to the large number of degrees of freedom (DoF), require higher spatial resolution and reproducibility associated with the control signal. Prosthesis control using such invasive neural interfaces is particularly appealing, as specific nerves which were previously connected to the amputated ex-

tremity, can be directly targeted and reconnected to the prosthesis in a physiologically relevant manner [63]. There is a multitude of electrode types that can be used to invasively record from the PNS, including cuff electrodes, intrafascicular electrodes, micro-electrode grid arrays and implantable sensors [57, 63–65]. Cuff electrodes are a type of extra-neural interface composed of an insulating tubular sheath that completely encircles a nerve and has two or more contacts exposed at the interface between nerve and sheath [57]. An advantage of cuff electrodes is that the neural activity being recorded is confined to the inner space of the electrode, and in particular they display enhanced mechanical and signal stability, especially during chronic recordings. Longitudinally implanted intrafascicular electrode (LIFE) and transverse intrafascicular multichannel electrode (TIME) recordings use a type of intra-neural electrode which is placed in direct contact with the nerve fascicular they are recording from, thus increasing both the selectivity and SNR [57, 63, 66]. Micro-electrode arrays (MEA) generally consist of a square grid of micro-electrode contacts in a 1D, 2D or 3D configuration. They are inserted percutaneously into fascicles of peripheral nerves and as such, require a less invasive surgical procedure than alternative intra-neural recordings. The size (which is on the millimetre scale) and morphology of the MEA allows for action potentials from a small group of axons to be recorded. It should be noted however, that many of these electrodes will be redundant during recording, with only about 10 - 20% of the electrodes exhibiting a response [57, 64]. Although the recorded activity shows high spatial resolution, this is at the expense of mechanical stability due to its highly rigid construction. In the last few years, implantable sensors have been developed which completely negate the need for percutaneous wires [65]. Instead, bio-compatible capsules complete with custom designed integrated circuits are located extra-neurally within the muscles. Communication and powering of the sensors is achieved through a bidirectional telemetry controller, positioned externally around the muscle group. Although it has a reduced selectivity when compared with some of the other iEMG techniques, it shows great promise in terms of chronic long-term recording. Although iEMG interfaces show great promise for applications where such signals are physiologically relevant, for example, prosthesis control, they suffer from the obvious disadvantage of being highly invasive and require a surgical procedure during interface fitting.

In the case of high-level upper-limb amputees, there are generally very few control sites for command of externally powered prostheses [67]. As such, either many of the DoFs associated with the arm will be lost or alternative control sites need to be used. These sites tend to be one of the remaining pairs of agonist-antagonist muscles in the stump of the amputated limb [68]. Based on this restricted subset of sites, it usually only allows for sequential control and can have an unnatural feel due to shoulder, bicep and/or triceps muscles being used to direct wrist or hand motions. Although the muscles which control wrist and hand movement have been lost, the residual nerves which transported

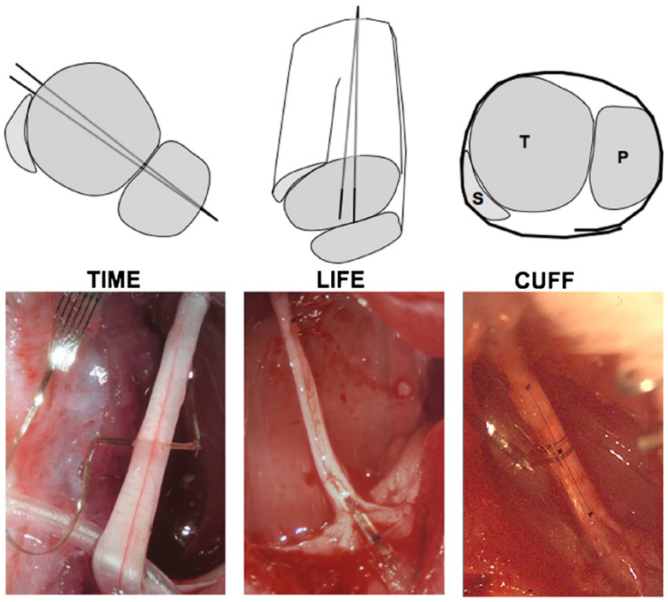


Figure 2.4: Examples of invasive EMG electrodes and their proximity to the nerves [66]

the electrical signal to them generally remains intact. To increase the number of sEMG control sites, a surgical procedure called targeted muscle re-innervation (TMR) has been developed which enables these functioning nerves to be utilised [69]. It involves transferring the residual nerves from the shoulder onto an alternate denervated muscle group, which are no longer 'biomechanically functional', since they are are no longer attached to the missing arm. The target muscle group is generally located in the chest or back areas, for example the pectoralis muscle, as these provide a large surface for post-operative electrode placement. Apart from increasing the number of sites that an amputee with high level disarticulation may need, there is a major benefit in terms of a much more natural and physiologically appropriate control signal being generated. Typically, the input sites are arranged in an agonist-antagonist configuration allowing the amputee to just think about contracting and relaxing the associated arm muscle, with the prosthesis mirroring this action. This negates the time required for the user to learn the control set. Once the initial surgery is completed, no other surgery is required. However, the recovery time post surgery is approximately 4-6 months. To highlight the increased level of control in terms of the number of DoFs that can be controlled, five patients with shoulder disarticulation or transhumeral amputations underwent TMR and were able to perform ten different elbow, wrist and hand motions, using twelve sEMG electrodes [69]. Although very suitable for its intended purpose, upper limb prosthesis control, in other AT applications such a procedure is generally completely unwarranted.

A completely different and unique biofeedback system has been developed based on infra-red thermal imaging (thermography) of the voluntary opening and closing of the

mouth [70, 71]. Using an IR camera, it monitors the oral cavity using the principle that the expired air during opening is generally warmer than the surrounding oral tissue. These temperature changes are monitored using video and image processing techniques performed on the associated thermographic data. It can therefore be used as a binary switch-based input, requiring only a very small cognitive load during operation. The opening of the mouth needs to be maintained for a duration of one second and thus the associated throughput is very low. However, for severely disabled individuals who, for example, show both motor and cognitive dysfunction, it has the potential of providing a simple access pathway to assistive devices. It should be noted that due to it being an optical based approach it has similar disadvantages, in terms of portability and set-up, to body tracking systems.

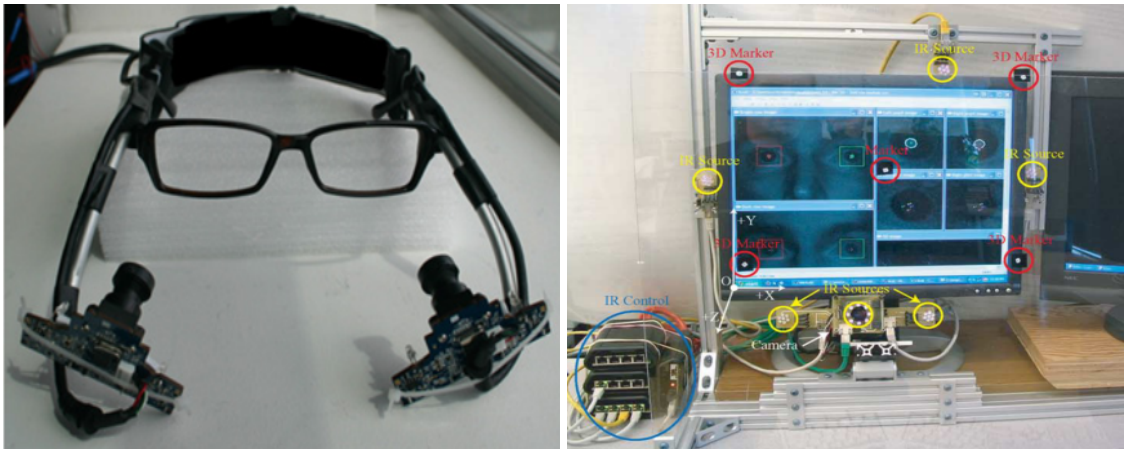
A less common interface, similar in setup to EMG-based systems, utilises bio-impedance sensors to measure the change in bio-electrical impedance due to the movement of muscles [72]. As the muscle contracts, it changes its sectional area and volume which gives a proportional change in its bio-impedance. By measuring this change using three electrodes positioned across the left and right trapezii muscles, a two channel system based on left and right contractions was designed. Additionally, a command strategy was developed for directing a wheelchair, which involved a left or right contraction and simulated a left or right motion in the wheelchair. A simultaneous contraction of both sides toggled a run/stop command. Adequate volume between the electrode pairs is required to obtain a suitable control signal and thus the selectivity as compared with sEMG can be considered much lower. Another communication method involving measurements of muscular activity is based on Mechanomyography (MMG) signals [73, 74]. Rather than directly measuring the electrical activity associated with muscle contractions, instead the muscle is quantified based on the associated mechanical and acoustic activity. It is therefore also known as acoustic myography and has a significant advantage over EMG electrodes as there is no requirement for electrical conductive gel or direct contact between sensor and skin [75]. Due to the nature of the MMG signals, the reader is directed towards the bio-acoustic subsection (Section 2.6) for a full discussion.

2.4.3 Gaze tracking systems

Tracking of eye motion provides an effective means for communicating with peripheral systems. Movements of the human eye are ultimately controlled by a paired structure in the midbrain known as the superior colliculus. It is attached, via contralateral connections to each eye through the optic and oculomotor nerves. The oculomotor nerve transmits the majority of information relating to eye motion to six extraocular muscles which actuate the eye within the orbit. Visual movements of the eyes is through discrete, rapid motion called saccades, which act to align the central fovea with a fixation point [37].

This high frequency saccadic movement provides insights into one's motor intention, as visual feedback plays an important role in motor planning. Thus by monitoring an individual's point-of-gaze (PoG) through the position of their fovea, generates an analogue signal directly relevant for neuroprosthetic control.

Two major methods exist for tracking an individual's gaze direction and are differentiated by the type of signal measured and sensing equipment required. The most popular method uses optical camera-based hardware and directly tracks the light that is reflected off the cornea [37, 76–78]. It can be further subdivided into two groups, namely, contact and non-contact based hardware solutions. Figure 2.5 shows typical hardware setups for the contact and non-contact based solutions. Contact-based solutions involve mounting the sensory equipment on the head and therefore contact between the face/head and sensory housing is required. In the non-contact case, the camera(s) are positioned some distance in front of the individual and are therefore physically uncoupled from them. As such, non-contact gaze interfaces require some estimation of head motion to avoid significant degradation in the estimated eye movement due to inherent coupling between head and eye motion [77]. Electrooculography (EOG) is the alternative method for monitoring gaze direction and is based on measuring an electrical signal from the eye. In this regard, it is similar to the previously mentioned EMG signal [79–81]. The EOG signal is generated due to the eye acting as an electric dipole, formed between the positive cornea and negative retina pole [81]. The associated change in the corneo-retinal potential can be measured using surface EOG electrodes and can be considered proportional to the movement of the eyes.



(a) Contact-based solution [37] (b) Remote solution setup [77]

Figure 2.5: Examples of optical gaze tracking systems

Optical systems utilise computer vision and image processing techniques to actively monitor the position of the fovea in relation to some fixed datum [82–84]. This is based on corneal reflection (or eye glint) which is the reflection of a light source on the outer

surface of the cornea [85]. In most cases, the eye is illuminated with an additional light or set of light sources, which although usually in the IR spectrum can also be in the visual spectrum. Their main role is to provide a ray trace, that can be used to estimate the position of the PoG, although in some cases it is to prevent variation in reflections due to changing background lighting conditions. There are two principal computational techniques for estimating the PoG from optical images, the first is based solely on the pupil-corneal reflection recorded monocularly, while the second is based on models of the eye and system in conjunction with a stereo vision setup [77].

The first of these techniques is based on a single estimated ray, traced between the light source, a single eye and the camera. The distance between the head and camera cannot be estimated and therefore unless found by some other means or kept fixed, will severely degrade the PoG estimate. This sort of setup is therefore only suitable for head-mounted (contact) solutions where the distance between the head and camera remains constant. This allows for an estimate of the PoG at its intersection with a 2D surface, such as a computer screen, to be determined. The model-based method, utilises both eyes and as such, implicitly estimates and compensates for head motion. The ray paths from both eyes can be used to determine the position of the eyes in 3D space, the direction of gaze and the PoG at the intersection of this axis and the surface of interest [77]. Parameters associated with the system and user model would need to be determined and is generally done using a calibration step involving predetermined fixation points. This system setup is suitable for non-contact situations and enables 3D PoG estimation. Additionally, systems based on a single camera and two or more light sources which provide the multiple ray paths required, can be utilised in a non-contact setup, although a calibration procedure would also be required [86].

More recently, a low cost contact-based solution has been developed for 3D gaze estimation [37]. It utilises off-the-shelf cameras which are adapted and attached, in a binocular configuration, to a pair of standard reading glasses (see Fig. 2.5a). Rather than using eye glints and cornea reflections to locate the center of the fovea, each iris is illuminated with an IR light source thus allowing its outline to be extracted, based on the contrast between itself and the rest of the pupil. After noise and shape filtering, an ellipse is drawn round the iris and is assumed concentric to the fovea. By performing a preliminary calibration, the PoG of the user is able to be tracked in 3D space and as such, the system has been named gaze tracker 3D (GT3D). The resolution of the system is comparable to commercial systems, with an error of approximately 1.1 cm ($\pm$0.7), 1.2 cm ($\pm$1.1) and 5.1 cm ($\pm$5.0) in the x, y and z directions, respectively, following calibration at a distance of 50cm. Using off-the-shelf components, the associated hardware costs in development of the GT3D system is about $30, with an estimated achievable ITR of 43 bits/s based on this 3D setup.

The alternative method of tracking eye motion is to use an interface based on EOG signals [79, 87]. The system setup comprises a set of wearable electrodes, similar to both EEG and EMG electrodes, positioned at various points around the eye. As briefly mentioned earlier, EOG measures the electrical signal from a steady electric potential field originating from the eye. This is assuming a stable corneo-retinal potential difference occurring due to the dipole between the positive pole at the cornea and negative pole at the retina, with a range of about 2 - 20 mV (or 5 - 20 $\mu V/^{\circ}$) [80, 88]. Figure 2.6 shows the standard EOG electrode layout with three positioned around one eye (top, bottom and far side) and one on the far side of the remaining eye. An additional reference (r) electrode is occasionally placed on the forehead or above the remaining eye of the individual. This topology therefore allows movements in the coronal plane to be tracked because as the eyes move in the vertical (v) or horizontal (h) axis, there is an associated change in the resting potential of the corresponding electrode pairings. A required processing step is the removal of baseline drift which although unrelated to eye movement is intrinsic to the EOG recording. This can be done either via analogue filters or post-processed using, for example, a discrete wavelet filter. To enhance the useability of EOG interfaces, some research groups have recently integrated the electrodes into wearable goggles, similar in appearance to the contact-based optical solutions [87, 88]. The electrodes can then be connected through a wired connection or using a wireless link, such as a transmitting Zigbee protocol. It should be noted that the processing associated with EOG signals only involves two 1D channels of data whereas optical systems require 2D image processing techniques and thus have a much higher computational demand. This reduction in information alongside the issues associated with baseline drift lead to reduced resolution and reproducibility associated with the signal and as such, are generally used in gestural rather than analogue control modes. Furthermore, the electrodes require contact and placement with a prominent facial area leading to highly noticeable and cosmetically insensitive hardware.

Intrinsic to any volitional interface which utilises continuous eye movements as command input is what is termed the 'Midas-Touch' problem [37, 78, 87]. The 'Midas-Touch' problem can be described as differentiating normal behavioural eye motion from intentional eye commands [37]. As the eyes are the most important source of feedback during any motor-related task, the eyes are constantly surveying the task space and are never inactive. This means that an additional 'clutch-like' input is needed which can engage or disengage the eye-tracking control. The other option is to use specific discrete eye gestures as the instruction set. However, this naturally implies that the ability for analogue control ability will be lost [78]. In the continuous mode, options that have been explored include additional switches, eye winks or dwell time. As eye blinks occur in both eyes during normal activity, single eye winks can only be detected with a stereo system, thus allowing them to be differentiated from normally occurring eye blinks [37]. Regardless of

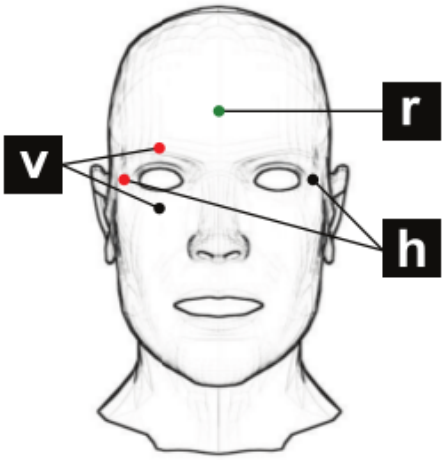


Figure 2.6: Placement of the EOG electrodes around the eyes [88]

such mitigating factors, the fact that the eyes can only perform a single task at any point in time implies that any type of gaze tracking system will inhibit normal functioning of the visual system during command and control of peripheral devices. This can be especially hindering if an individual wishes to switch between AT control and normal daily activities, further increasing the cognitive load associated with them. Although non-contact optical-based solutions are highly imperceptible to the user, like optical body tracking systems, they generally tether their usage to a single location. On the other hand, contact-based solutions significantly increase the portability associated with gaze tracking systems, at the expense of highly obtrusive hardware which is perceptible to both the user and those they interact with.

In the past, work has been done on an interface that uses eye winks solely for command input [89]. Rather than using cameras, infra-red sensors (emitters and detectors) are mounted on an ordinary pair of spectacles and focus on the same point at the lateral corner of both eyes. The premise being that the eyelid is more absorbent and reflects less of the spectrum than the sclera does. As mentioned previously, intentional movements can be separated from normal eye blinking by monitoring each eye separately. This does, however, only allow for two different commands to be issued (one from each eye) with augmentation of this instruction set through compound combinations of these actions. Although considerably less bulky than the contact-based optical interfaces, they still require mounting in front of the eyes and are therefore deemed obtrusive. Additionally, executing only eye winks for control purposes means issues relating to the 'midas-touch' problem are mitigated but issuing of commands can still somewhat interfere with the execution of normal visual tasks.

2.4.4 Brain-machine interfaces

At the forefront of current research trends in HMI technologies are brain machine interfaces (BMI), also known as brain computer interfaces (BCI) or simply neural interfaces (NI) [1, 90, 91]. The main reason for this interest, is that, rather than substituting one motor function with another, they bypass functional loss by connecting directly to the source of the motor command, within the brain. As such, it has the potential for providing a means of communication for even the most severely disabled, that is, those individuals who have lost all motor function entirely, for example, due to locked-in-syndrome. Neural interfaces can be separated into two main categories, namely, extracortical or intracortical and are differentiated by the proximity of the electrodes to the cortical structures [91]. Extracortical devices measure brain waves either invasively or noninvasively. Noninvasive techniques include electroencephalographic (EEG) and magnetoencephalographic (MEG) systems and are positioned externally on the scalp of an individual. Conversely, invasive electrocorticographic (ECoG) activity is measured, under the skull, on the surface of the cortex. Intracortical systems on the other hand, utilise either single unit action potentials (AP) recorded from individual spiking neurons or local field potentials (LFP) recorded within deep brain or cortical structures. The trade-off between the various techniques is similar in context to EMG signals and is specifically related to the depth that the electrode is placed within the cortex as opposed to muscular structures. As the spatial resolution and quality of the bio-signal increases, the associated costs, in terms of the surgical procedure and associated risk of infection also increase. Figure 2.7 highlights the relationship between the spatial resolution, level of invasiveness and neural population involved for the various neural signals described.

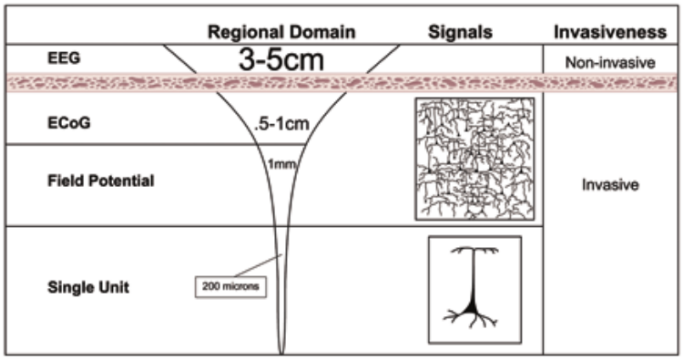


Figure 2.7: Relationship between spatial resolution, level of invasiveness and neural population involved for various neural signals [91]

To date, research in the field of BMI technology has predominantly focused on the analysis of single and multi-channel EEG signals for decoding of high-level volitional commands [3, 92–95]. This is mainly due to the noninvasive nature of the signals which alle-

viates the need for surgical intervention therefore significantly reducing the health risks. They are also comparatively cheap and in many cases have increased portability. As EEG signals are recorded through a set of electrodes positioned uniformly across the scalp of an individual, the reduction in spatial resolution is attributed to two main factors, namely, mixing of activity from disparate neural populations and attenuation of the signal as it transmits through the skull [26]. Although the spatial resolution and fidelity of the electrical signal is compromised, the brain wave can be recorded anywhere on the scalp, thus allowing activity associated with different functional areas of the cortex to be monitored. Usually signals are recorded from multiple electrodes with the layout of these electrodes following specific topologies. The internationally recognised method for recording scalp EEG signals is via the 10-20 system, shown in Fig. 2.8 below [96]. A more detailed setup can be specified and utilises additional points interpolated between these standard electrode positions. A smaller setup can also be used and generally employs a specific subset of the 10-20 topology.

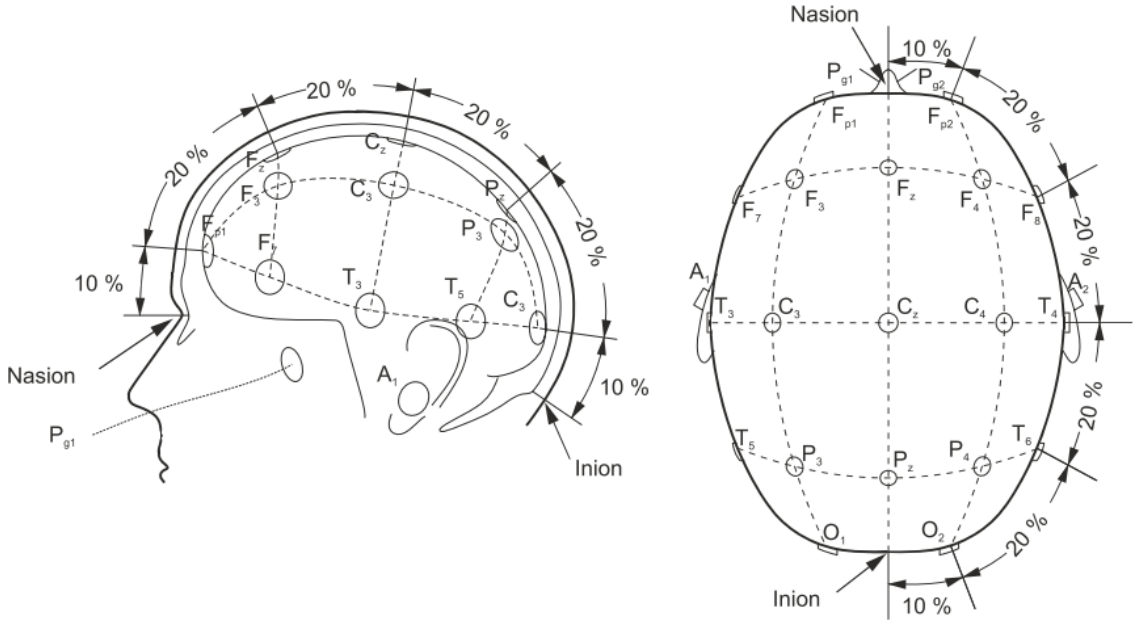


Figure 2.8: EEG electrode placement using the 10-20 topology [96]

As the brain is highly connected, electrical signals originating from it relate to a multitude of cognitive tasks and concurrently occurring phenomena. For the most part, these are highly incomprehensible and their specific origins unknown [96]. However, based on certain tasks, individuals can learn to modulate the brain signal in such a way, as to allow it to be decoded and thus a volitional command issued. There are four main methods which can be used to instigate these control modulations, namely, visually-evoked potentials (VEP), slow cortical potentials (SCP), P300 evoked potentials and sensorimotor

rhythms [93, 96]. When an individual receives a visual stimulus there is a relatively large modulation of brain activity within the visual cortex. The amplitude of the measured VEP is much greater when the stimulus is in the centre of the visual field. Steady-state visually-evoked potentials (ssVEP) are elicited when a stimulus is flashed at a frequency greater than 6 Hz. The typical VEP-based BCI involves a set of flashing stimuli, which when fixated on, induces a unique ssVEP response. The distinction between responses is achieved by flashing the stimuli at distinct times, frequencies or stochastically [96]. As VEP-based communication systems depend on the user's ability to control gaze direction, it is described as a dependent BCI, due to the generated control signal being in part dependent on their oculomotor function[1].

The remaining methods for controlling brain modulation can be considered motor independent, as they do not rely on the user retaining any motor function whatsoever. SCPs are among the lowest frequency features of the EEG signal, occurring at frequencies less than 1 Hz and lasting for several seconds [93]. They are associated with a change in cortical activation, with a negative SCP indicating an increase, while a positive SCP is indicative of a decrease in activity. Individuals can learn to modulate the SCP through a 'thought translation device' which ultimately gives the user feedback on their current level of SCP activity, thus allowing them to regulate it. Alternatively, a P300 evoked potential manifests as a positive peak in the EEG signal, with a lag of about 300ms which generally occurs over the parietal cortex. It is evoked due to infrequent or particularly significant auditory, visual or somatosensory stimuli [93]. The classical way of utilising the P300, is to present a 6-by-6 grid of symbols to the user. By alternately flashing the rows and columns every 125 ms, twice each and in a random order, the subjects intention can be decoded. The subject is instructed to choose a symbol of interest and then increment a count every time this symbol is illuminated. The average response for each row and column is computed in the parietal cortex and then for each symbol choice, the overall amplitude of the P300 is calculated. The symbol which shows the most significant P300 response can then be selected. Another independent BCI control strategy, utilises sensorimotor rhythms, which are oscillations in the mu (7 - 13 Hz) and beta (13 - 30 Hz) frequency bands and thought to be produced within thalamocortical circuits [97]. Movement and its associated preparation is typically accompanied by an event-related desynchronisation (ERD), equivalent to a decrease in the mu and beta rhythms, which is generally contralateral to the movement. Event-related synchronisation (ERS), on the other hand occurs after the movement and is associated with relaxation. As these events can also occur with just movement imagery (without the need for actual physical motion), by monitoring the level of synchronisation, an independent BCI command can be issued with the user only required to think about moving, for instance, their hand.

[1]The level of functional oculomotor control is much less than that required by gaze tracking systems

Another noninvasive extracortical technique, MEG, measures brain activity by detecting weak magnetic fields caused by current flows in the cortex [98]. These magnetic fields are measured using a multichannel super-conducting quantum interference device, within a shielded environment. Considering the proximity of the sensors to the cortex, this system offers excellent time resolution while also providing good spatial resolution which can be as fine as 2 - 3mm. Although mainly used clinically, they have been explored as a possible input for BCI. However, so far no-one has been able to demonstrate significant advantages over EEG as well as the system being inhibitively expensive [37]. By placing electrodes invasively on the surface of the cortex, specifically the dura, ECoG activity can be recorded. They are similar to EEG signals, in the sense that they both are indicative of synchronous, volume-averaged, low-pass filtered neural activity. In the case of ECoG signals, they show increased spatial resolution and improved SNR across a wider range of frequencies [26]. For this reason and due to ECoG recording already being an established treatment in focal epilepsy, researchers have begun investigating ECoG activity as a possible control signal for BCI applications [99]. As such, motor imagery based control of a 1D cursor, as well as open-loop classification of 2D tasks have been demonstrated [26, 99]. However, the ability to chronically record is still yet to be ascertained, while it is still unknown if the placement of these electrodes poses more or less surgical health risks than alternative intracortical recordings.

As the name suggests, intracortical BCI systems record brain activity by placing electrodes internally within the cortex, beneath the dura. Of these the most successful has been the BrainGate neural interface system, developed by Cyberkinetics Neurotechnology Systems and has been implanted into individuals suffering with paralysis, such as tetraplegia [26, 100, 101]. The system uses an MEA to record neural spikes extracellularly, by placing the microelectrodes less than a few hundred micrometres from the nerve cells soma. To obtain an effective and reliable control signal, the spiking patterns from a localised population of neurons is required, with the chosen MEA similar to the ones used for recording percutaneously from fascicles of the PNS. The electrodes are placed into the primary motor cortex, as it is believed to be the most closely coupled to the production of movement and as such, ideal for the generation of control signals based on motor imagery. The translation from neural to control signal requires the ability to decode the user's intention based on the spiking patterns. As hand kinematics, force and motion are well represented by neural firing in the arm area of the primary motor cortex, imagined movements of the hand were used to express user intention [100]. The system has been demonstrated while controlling the trajectory of a mouse cursor in 2D, as well as 3D control of the end effector and grasp action associated with an assistive robotic arm. Furthermore, the group have validated the ability of the system to chronically capture APs over an extended period of five years [101]. More recently, there has been exploration of

LFP activity as an alternative or complementary intracortical neural signal for controlling BCI systems [102, 103]. The LFP is a measure of the summed synaptic current flowing through a volume of tissue and as such, is not dominated by any individual neuronal firing. Although mainly explored in animal models, functional neurosurgery especially deep brain stimulation (DBS), has presented the opportunity of exploring LFP-based BMI in human subjects [97, 104].

2.4.5 Orally generated communication

Orally generated communication interfaces are defined as those which utilise commands involving parts of the vocal tract, including various signals emanating from within the oral cavity, vocally and/or nasal passage airways. The physiology and control underlying these systems is diverse, complex and usually involves many interacting parts. As such, the interfaces which capture such signals are also wide-ranging and involve many different sensing technologies. It should be noted that excluded from this particular discussion are tongue operated systems, which are described subsequently in Section 2.4.6, due to the similarity in terms of the generating commands to the TMEP interface described in this work. Currently one of the oldest and most widely used assistive HMIs is what is termed as a 'sip and puff' device [20, 105]. It comprises a tube connecting the users mouth to a pressure sensor allowing a positive or negative pressure to be exerted. This pressure differential is generated by changes in airflow caused by either blowing (puffing) or sucking (sipping) down the tube. The device therefore operates as a simple binary switch, with the user able to increase the command set by using multiple tubes. In the early sixties, such a device was integrated into a typewriter controller and is considered one of the earliest electric communication devices [105]. Today, its widespread commercialisation can be attributed to its simplicity which is also testament to its functionality, useability and reliability. Having said this, it does require individuals to have full control of their diaphragm and subsequent oral airflow, as well as the tube constantly moving between mouth and environment which can lead to associated hygiene issues. Adaptations to the sip and puff technology include making it into an adjustable multi-level switch, with different pressure bands mapped to polynary outputs [106]. Figure 2.9 shows a typical sip and puff device.

More recently, a similar device in terms of the hardware involved, has been developed to mitigate some of the aforementioned issues associated with the sip and puff technology. It utilises the sniffing reflex associated with controlling air flow through the nasal passage [107]. Due to sniffing requiring fine control of the soft palate or velum, which is innervated by multiple cranial nerves, it is hypothesised that sniffing control may remain preserved even after severe injury, such as those leading to locked-in-syndrome. The sniff controller consists of a nasal cannula that carries changes in nasal pressure from the nostrils to a pressure sensor, with inhaling and exhaling through the nasal cavity generating



Figure 2.9: A typical sip and puff device [106]

commands. To allow individuals who require artificial respiration to use the device, a passive controller was also developed and includes a small pump to generate a constant low-flow stream of air within the nasal passage. Thus by just opening or closing the soft palate, a change in pressure occurs and a command can be issued. The research team have demonstrated the technology on 96 healthy individuals, two out of three locked-in individuals and ten quadriplegics for various text entry, computer browsing and wheelchair navigation tasks using discrete combinations of sniffs. Although all healthy patients could use the original sniff controller, it was found that about 25% of healthy volunteers did not have volitional control of their soft palate at first screening and therefore would not have been able to use the passive sniff controller. Of the locked-in-patients, one could sniff straight away, one required training via a simple feedback program over the course of a few weeks, while one patient was unable to learn soft palate control at all. Although exhibiting many advantages over existing technologies, one key shortcoming with this technology, is it will not work if the nasal passage is blocked, for example, due to a common cold.

As speech is the natural medium by which humans communicate with one another it makes sense to extrapolate this idea to communication between man and machine. Automatic speech recognition (ASR) is a promising avenue for control of HMI-based systems and has led to several innovative solutions for tele-operation [108–110]. Human speech production involves a complex system of physiological apparatus, including the lungs, glottis (vocal folds) and vocal tract [111]. For reference, Fig. 2.10 shows the main features of the vocal organs and oral anatomy in general, alongside their arrangement inside the neck and head. Speech production can be modelled by the so-called source-filter model, with the air-stream generated by the lungs (source) and the vocal tract (filter). It is due to

the complex nature of these vocal organs, that each speaker possesses a unique 'timbre'. The vocal folds are a pair of band-like membranous tissues that sit across the larynx thus constricting the airflow and vibrating in a speaker characteristic way during pronunciation and phoneme forming. As briefly mentioned earlier, the vocal tract comprises three coupled cavities, namely, the pharyngeal, oral passage and nasal passage. Based on the position of the jaw, tongue, lips and soft palate, they resonate at different frequencies, thus filtering the signal as it leaves the speakers mouth. Based on all these complex physiological interactions, acoustic speech signals are considered both highly nonlinear and nonstationary [111].

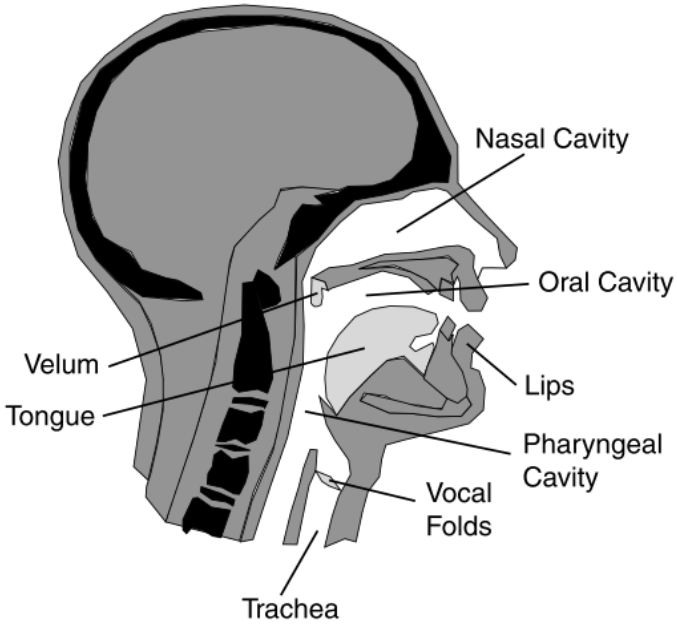


Figure 2.10: Arrangement of the vocal organs inside the head and neck [112]

Speech processing has been an active area of research for the last five decades [113]. Key contributions to the field were made in the seventies and eighties, including the use of statistic models such as HMM for recognition and MFCC for feature extraction [113, 114]. Over thirty years later, these effective and simple techniques still dominate. The cepstrum is described as the Fourier transform of the logarithm of the estimated spectrum of the speech signal and can be seen as information relating to the rate of change in different frequency bands. As speech signals are highly variable, not to mention the huge diversity of the human language, speech recognition requires a extremely large corpora, for speaker independent systems to achieve proficiency [114]. This coupled with the time-complexity associated with fitting the HMM parameters, for instance, when using the Baum-Welch algorithm, means that speech processing is very computationally intensive. Most ASR front-ends continually analyse short sequential signal frames, typically 30 ms in length, with speech assumed quasi-stationary over this period. Other approaches used

for speech recognition include neural networks and dynamic time warping (DTW) [110]. Whereas the HMM approach is widely used for speaker independent and large vocabulary situations, the DTW approach is more suited for speaker dependent and small vocabulary speech recognition [110]. This, combined with its simpler implementation, may make the DTW approach more suitable for accurate control of assistive technologies using discrete spoken commands.

In an attempt to permit analogue control of the environment, researchers have developed a vocal system which outputs a continuous control signal proportional to certain vocal parameters or prosodic features, for example, vowel quality, loudness and/or pitch [18, 108]. As such, it differs from traditional ASR systems which generally recognise a sequence of discrete sounds. As it subsumes the role of a conventional joystick, utilising vocal input rather than hand motions, it has been coined the vocal joystick. The vowel quality is classified using a multi-layer perceptron adapted to each user, which directly maps to a 2D vowel space. There are currently three control strategies, namely, 4-way mode where the vowel quality moves the control signal in the horizontal and vertical cardinal directions only, 8-way mode which adds the four additional diagonal directions, as well as a continuous adaptive filtering mode, which smooths the trajectory using a Kalman filter. In all modes, the associated speed of the control signal is mapped to the loudness of the produced sound. The vocal joystick system has been demonstrated for computer cursor control as well as control of a 5-DoF robotic arm [18, 108].

Due to communication being ubiquitous and intrinsic to humans and daily living, speech recognition systems can suffer from problems in terms of distinguishing normal conversation and speech from intended commands. This is similar to the 'Midas-Touch' problem (Section 2.4.3) that is experienced by gaze-tracking systems. Furthermore, the signals associated with non-intended commands, both internal and external to the user, can severely cross-contaminate, degrade and mask the quality of the intended signal. To mitigate such issues, the ASR would either require ON-OFF states whereby the user uses an additional input (or command) to toggle between them. Alternatively, non-verbal vocalisations within the command set could be employed to prevent confusion with normally occurring speech, as in the case of the vocal joystick. The first solution implies that during an ASR ON state, while the user is performing volitional speech control of an AT, the user is potentially unable to communicate with others. If conversation does occur during the ON state then there is a heightened risk of false positives and therefore unintended control actions to be issued. Furthermore, the use of an additional input or command to toggle states adds complexity and an increased risk of confusion in the overall system. By issuing non-verbal vocalisations as commands, a user can potentially use an ASR system during normal and continual conversation. However due to the expansive nature of language, the command set has to be designed carefully and will generally involve

longer non-articulated sounds, such as extended vowels, leading to a less intuitive and thus slower interface. Even though the sounds don't normally occur during conversation, there is still an increased possibility of misinterpretation by the system leading to false positive classifications due to the chaotic and expansive nature of speech. As mentioned previously, additional vocal characteristics such as pitch, power and vowel quality can be incorporated into the system enabling analogue control of the environment [18, 108]. Speech recognition also has the benefit of an extremely large instruction set from which the designer can choose, with the trade-off being speed and complexity to the user.

Standard ASR systems sense airborne acoustic signals usually captured via a microphone positioned in front of the users mouth. This makes an issued command perceptible to those around you and can thus bring additional unwanted attention to the user. In noisy environments, speech recognition may be completely infeasible and is generally also implausible in cases where an individual exhibits an additional speech disorder. Due to the complex motor synergies required during vocalisation, some neurological speech impairments such as Apraxia, Aphasia or Dysarthria can lead to reduced or complete loss of speech [115]. In some cases, only a localised motor area, associated with a small component of the speech production system will be affected or at a higher level, co-ordination between the various muscle synergies required during speech is affected. In some cases, there is still the ability to fully control localised motor areas associated with speech, such as intentional lip and/or tongue motion. To aide individuals exhibiting speech impairments, researchers have started looking toward silent-speech interfaces, whereby speech is performed without vocal fold vibration and thus audible sound. These have the additional benefit that the issue associated with an interface being susceptible to external audible noise is mitigated, as the airborne acoustic signals are no longer being tracked. There are a variety of ways that silent speech can be interpreted and include but are not limited to, body-conducted vocal-tract resonance, lip reading using optical sensing and/or electromyography and ultrasonic imaging of the tongue and lips during speech [116, 117]. Although the majority of work in this field so far has generally been related to signal enhancement rather than recognition, silent speech interfaces are a promising avenue for assistive HMI research. However, due to the wide-ranging spectrum of silent speech systems it is beyond the scope of this work to go into further detail about them.

The final oral interface worth introducing is based on tooth clicks and involves a wireless communication device worn comfortably around an individual's ear [19, 118]. The device has a sensor arm containing a triaxial accelerometer, which is positioned against the anterior tragus of the individual. The accelerometer continually monitors the vibrations occurring at this anatomical location and is able to differentiate the strong and unique vibrations associated with tooth-clicks from those occurring otherwise. It does this by ensuring that the spike energy associated with a tooth-click occurs within a certain time

window and is surrounded by a period of relative quiet [118]. The interface was developed to act as a hands-free remote switch and has been integrated into multi-modal systems, assuming the role of a computer mouse click. Initially it was paired with a IR head tracker (see Section 2.4.1) and compared against the sip and puff device and simple dwell-time as a means of performing cursor selections [118]. In various computer-based tasks, for both healthy and SCI subjects, sip and puff and tooth-clicks had a comparable and faster reaction time than dwell-time, while dwell-time and tooth-click performed with fewer mistakes than sip and puff did. Further to this, various combinations of optical head tracking or IMU-based head tracking for cursor control, with either tooth-clicks or speech recognition for cursor selection were evaluated [19]. In this research, a click would bring up a menu, which is followed by an associated movement of the head to select the required click functionality. Using this interface, tooth-clicks were found to be faster than speech (of the word 'click'), while tooth-clicks in combination with the optical head tracker was generally the fastest multi-modal interface out of those tested. The simplicity and subject independent processing associated with the system are indicative of its robustness, however, it operates only as a binary switch and thus has reduced ITR as well as high potential for false negatives due to missed clicks.

2.4.6 Tongue-operated systems

In many motor impairments, full function and control of the tongue generally remains, including high-level SCI, stroke and amputees. As touched on earlier, this can be mainly attributed to the hypoglossal and vagus efferent connections emerging above the spinal cord and are therefore generally unaffected when damage occurs to the PNS. Furthermore, tongue motion is known to have a large and distributed neural substrate especially when performing dextrous actions such as, horizontal and vertical motions, isometric contractions and tongue pointing [21–23, 119, 120]. These substrates include brain areas as diverse as large sections of the ventral primary motor cortex, the postcentral gyrus, the supplementary motor area and the putamen. Therefore, there is increased conservation of tongue function during neural dysfunction, as it is highly unlikely that damage to all these areas will occur. Not only is this neural network widely distributed, it also shares similarities with fine finger control and language production, further indicating its suitability for replacing conventional human communication methods [23, 121].

Over the years, there has been a broad spectrum of research into various methods of utilising the tongue as control input for an assistive HMI system. Of these, one of the most familiar types will be an intra-oral 'tongue-touch' system whereby a series of discrete switches are positioned within the oral cavity [122–126]. Figure 2.11 shows a typical tongue-touch device embedded within a dental retainer. The tongue is used in a manner whereby it assumes the pointing reflex associated with fine finger motor control, allowing

the user to generate a discrete control signal whenever the tongue comes into contact with one of the switches. A dental baseplate is normally employed as a mounting for the switches and is generally located above the tongue against the palate of the mouth and is fixed in place using the teeth. There is a variety of different sensors and switch topologies employed. These are designed to allow for ease of switching and manipulation whilst concurrently attempting to maximise the ITR. Inductive-based solutions have been designed based on Faraday's law of electromotive induction, either by pressing a spring mounted ferrite slug into a coil or using a magnetic activator attached to the tip of the tongue and touching this against a Hall effect sensor [122, 124]. The latter solution allows for a more compact topology due to a smaller switch profile. Other transducers used include simple membrane covered pressure and/or proximity sensors [123, 125]. There are various switch configurations, with the topology, in part, dependent on the required application. The standard configuration, designed for directional control, is a layout of five switches relating to the up, down, left and right functions, with a central fifth transducer providing a selection input. Bigger topologies have been utilised, namely, a nine switch configuration, thus increasing the instruction set and allowing the user to type based on compound sequences of these nine switches [127, 128]. Due to the required back plate necessary to mount the transducers to and the fixings needed to hold this within the oral cavity, 'tongue-touch' systems tend to be relatively bulky and inhibit the user's ability to masticate and verbally communicate without first removing it. Having said this, they can give efficient, robust and accurate control over the environment, although having a device continually removed and inserted from the mouth has associated health risks and issues concerning user comfort.

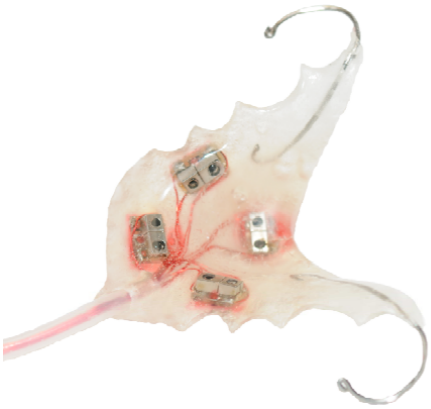


Figure 2.11: Tongue-touch system [126]

To allow for an analogue control signal to be generated, devices operated by the tongue but performing in a similar fashion to a joystick or computer mouse have been developed. One such system utilises a miniature joystick mounted on a mouthpiece allowing it to be

positioned at the root of the user's top central incisors and manipulated by the tip of the tongue [129]. Alternatively, the tongue-mouse makes use of a discrete grid of piezoelectric strips positioned in an x-y plane which produce a small charge proportional to the force of contact. This allows the tip of the tongue to be tracked over the nodes of the grid while it is in contact with the surface [130]. Both of these devices employ a modular bite switch for button selection. Even though these devices give the user direct control over analogue systems they require more complex mechanisms than the 'tongue-touch' systems and as such have an even bulkier profile within the oral cavity. This leads to increased inhibition of normal oral functions and the obvious hygiene risks mentioned previously.

More recently, a system based on tracking of volitional tongue motion by detecting the position of a small permanent magnetic tracer attached to the tip of the tongue has been developed [131–133]. This is known as the tongue drive system (TDS) with Fig. 2.12 highlighting the various hardware components associated with it. The tracer is attached to the tongue using either tissue adhesive, a piercing or implantation. A pair of triaxial magnetic sensor modules, fitted externally and bilaterally around the oral cavity, measures the change in magnetic field due to the movement of the tongue and sends the six channels of information wirelessly to a laptop for further processing. Six recommended tongue commands were chosen and included touching the roots of the lower left teeth, lower right teeth, upper right teeth, upper left teeth as well as the right and left cheeks [132]. During training ten of each of these actions is collected and a sensor signal processing algorithm is used to train the system. The algorithm uses principal component analysis (PCA) to extract a three feature vector from the six channel concatenated waveforms for each command. Based on the training set, a mapping to this feature space is found, with new data points continually mapped to this subspace and discriminated using a K-nearest-neighbour (kNN) classifier. Furthermore, a neutral command is additionally defined corresponding to the tongue resting in its neutral position and represents a zero state output. Generally, users were dissuaded from performing commands in the sagittal plane as it is in this midline that the tongue moves during normal function. The system has been evaluated on both healthy and high-level SCI subjects, under various environmental control paradigms including computer access and navigation simulations. It has also been demonstrated for real-time control of a power-assisted wheelchair. This evaluation has been in terms of classification accuracy (CA), ITR and response time. Although, showing promise as a tongue-based interface, it still requires the tracer to be placed within the mouth. Furthermore, the sensor modules need to positioned in relatively close proximity to this tracer either inside the oral cavity or just outside in front of the mouth (see Fig. 2.12). This can limit its appropriation due to drawbacks associated with hygiene, perceptibility and obtrusiveness.



Figure 2.12: Tongue drive system (TDS) [133]

2.5 HMI comparison space

To highlight and summarise the relative attributes of the previously described HMI concepts, Tables 2.1 and 2.2 provide a summary of the main pros and cons of these different communication and assistive interfacing technologies. The interfaces are grouped as previously, that is body, biofeedback, gaze, BMI, oral and tongue generated communication, with the main concepts from each group compared. Additionally, a conventional grouping is included, composed of the standard interface technologies including the computer keyboard, computer mouse, standard joystick and games/RF controller. Although these conventional technologies cannot be used by a wide sector of the disabled demographic, they provide a useful benchmark when comparing the assistive HMI concepts.

Using the proposed *key drivers*, described previously in Section 2.3, the HMI concepts can be visually compared within a HMI comparison space. The visualisations have been developed based on material selection charts and show two dimensional pairwise comparisons between the ten *key drivers*. Therefore, a HMI concept is represented by a localised space in this ten dimensional space. Each of the specific HMI concepts listed in Tables 2.1 and 2.2 are compared and indicated by either a circle or ellipse. Each HMI grouping is differentiated using a different colour (conventional - red, body - yellow, biofeedback - turqoise, gaze - purple, BMI - light green, oral - blue and tongue - dark green) with an additional boundary around the HMI instances associated with a specific grouping. Figures 2.13 and 2.14 show comparisons between the various HMI technologies for the *key drivers* associated with the feasibility and functionality of the concepts. Figures 2.15, 2.16 and 2.17 are associated with the useability and form comparisons between the same interfaces. The cost and ITR are based on quantitative data compiled from the following references: [1, 19, 37, 39–41, 45, 51, 69, 70, 74, 107, 118, 123, 133–138]. In some cases,

Interface	Pros	Cons
Conventional		
keyboard	high ITR, optimised for computer interaction, cheap	requires hand movement, tethered
mouse	high ITR, analogue control, optimised for computer interaction, cheap, intuitive	requires hand movement, tethered
joystick	high ITR, analogue control, optimised for directional control, intuitive	requires hand movement, tethered
controller	high ITR, analogue and discrete control, suitable for tele-operation	requires hand movement
Body		
(chin) *joystick*	doesn't require hand movement	constant skin contact, requires large movements, tethered, requires bodily movement
IMU	portable	needs attachment to extremity, requires bodily movement, drift
IR (emitter)	cheap	clear line of sight needed, tethered, requires bodily movement
optical (tracking)	analogue control, no sensors attached to body, relatively cheap	clear line of sight needed, tethered, requires bodily movement
exoskeleton	multiple DoFs	bulky, highly perceptible, requires bodily movement
Biofeedback		
sEMG	physiological relevant prosthesis control	requires muscular contraction, daily sensor placement, signal reproducibility, conductive gel required
iEMG	physiological relevant prosthesis control, better resolution and signal reproducibility	highly invasive, surgery required, percutaneous wires, requires muscular contraction
trEMG	physiological relevant prosthesis control, more sEMG sites	surgery required, requires muscular contraction
MMG	no conductive gel required, no crosstalk	requires muscular contraction, daily sensor placement, signal reproducibility, motion artifacts
thermography	highly accessible	clear line of sight needed, small instruction set
Gaze		
contact	analogue control	'midas-touch' problem, expensive, bulky, cosmetic issues
non-contact	analogue control, portable	'midas-touch' problem, expensive, clear line of sight needed, tethered
GT3D	cheap, analogue 3D control, high ITR, portable	'midas-touch' problem, bulky, highly perceptible, calibration

Table 2.1: Table summarising the pros and cons of the current state of art in assistive HMI systems

Interface	Pros	Cons
Gaze ***cont'd***		
EoG	portable, analogue control	'midas-touch' problem, baseline drift, sensor placement
BMI		
EEG	highly accessible, noninvasive, cheaper	slow, low spatial resolution (2-3mm), high cognitive load, daily sensor placement, signal reproducibility, conductive gel required
MEG	highly accessible, noninvasive, better spatial resolution then EEG	extremely expensive, unportable, high cognitive load
ECoG	highly accessible, increased spatial resolution, improved SNR	highly invasive, surgery required
LFP	highly accessible, increased spatial resolution, improved SNR	highly invasive, surgery required
MEA	highly accessible, analogue control, increased spatial resolution, improved SNR	highly invasive, surgery required
Oral		
sip and puff	simple, subject independent	hygiene issues, requires diaphragm control
sniff	simple, subject independent, highly accessible, robust, discrete and analogue control	doesn't work if user has blocked nasal passage, training required, can be unintuitive as requires soft-palate control
ASR	diverse instruction set, natural method of communication, accurate classification models	sound perceptible to others, noise-free environment required, relatively slow, interferes with communication
silent	soundless, accessible to individuals who cannot vocalise	less robust, potentially complicated hardware
(vocal) *joystick*	analogue control	sound perceptible to others, noise-free environment required, interferes with communication
tooth (clicks)	simple, subject independent, cheap	small instruction set, susceptible to missing clicks
Tongue		
touch	simple, cheap, robust, good ITR	hygiene issues, interferes with oral function
joystick	analogue control	hygiene issues, interferes with oral function, bulky, requires dextrous manipulations
TDS	minimal hardware inside mouth, portable, good ITR	hygiene issues, perceptible and relatively bulky hardware, interferes with oral function

Table 2.2: Table summarising the pros and cons of the current state of art in assistive HMI systems

quantities have had to be estimated or calculated from published performance measures while some costing has been extrapolated based on similar devices. Although no absolute scale is given, the ITR values vary logarithmically (base 2) between approximately 0.02 - 43 bits/second while the cost values also vary logarithmically (base 10) between approximately £10 - 1,000,000. All plots exhibit better performance in terms of the *key drivers* toward the top right of the chart and worst performance toward the bottom left. For example, the cheapest interfaces are positioned on the right side of Fig. 2.13, while the most expensive are positioned on the left. The positioning of the concepts in terms of the remaining drivers is based on the author's discretion using relativistic weights between 0 - 10.

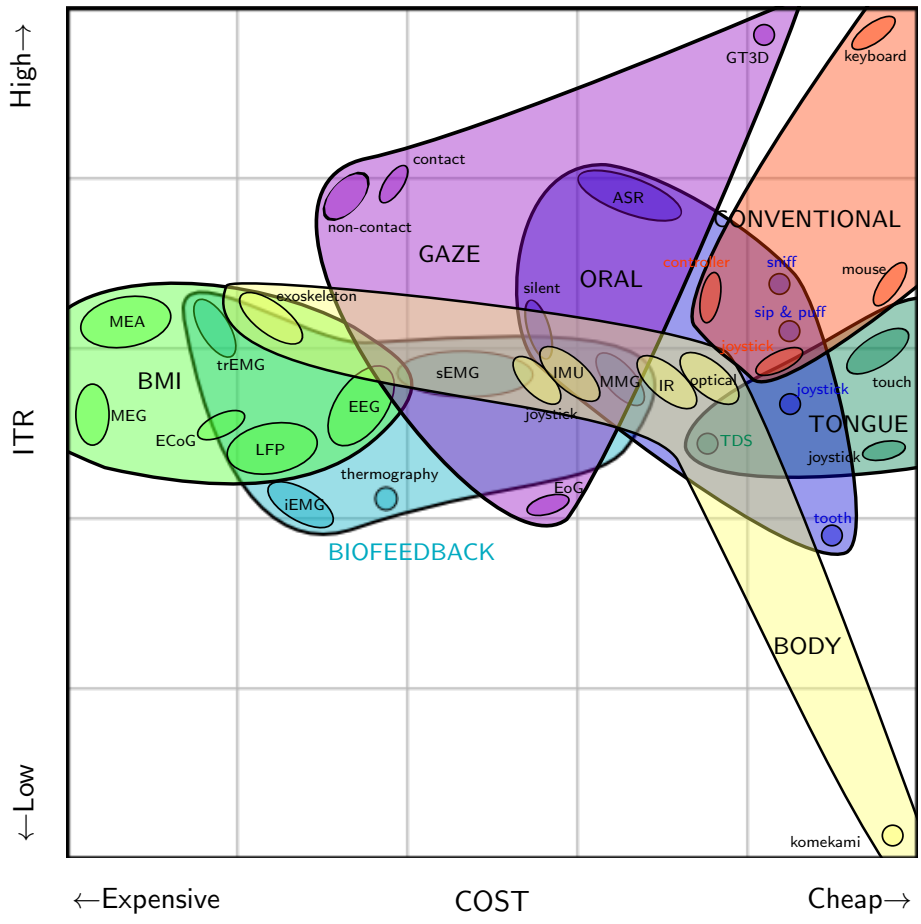


Figure 2.13: Feasibility and functionality: Cost vs. information transfer rate

These visualisations show that there is no single device that performs the best in terms of the specified *key drivers*. If there was, it would be located consistently in the top right of all the visualisations. It should be noted that some of the HMI concepts do perform comparatively better, across many of the *key drivers*, for example, the sniff based concept. While on the other hand, some of the systems perform comparatively worst, for instance,

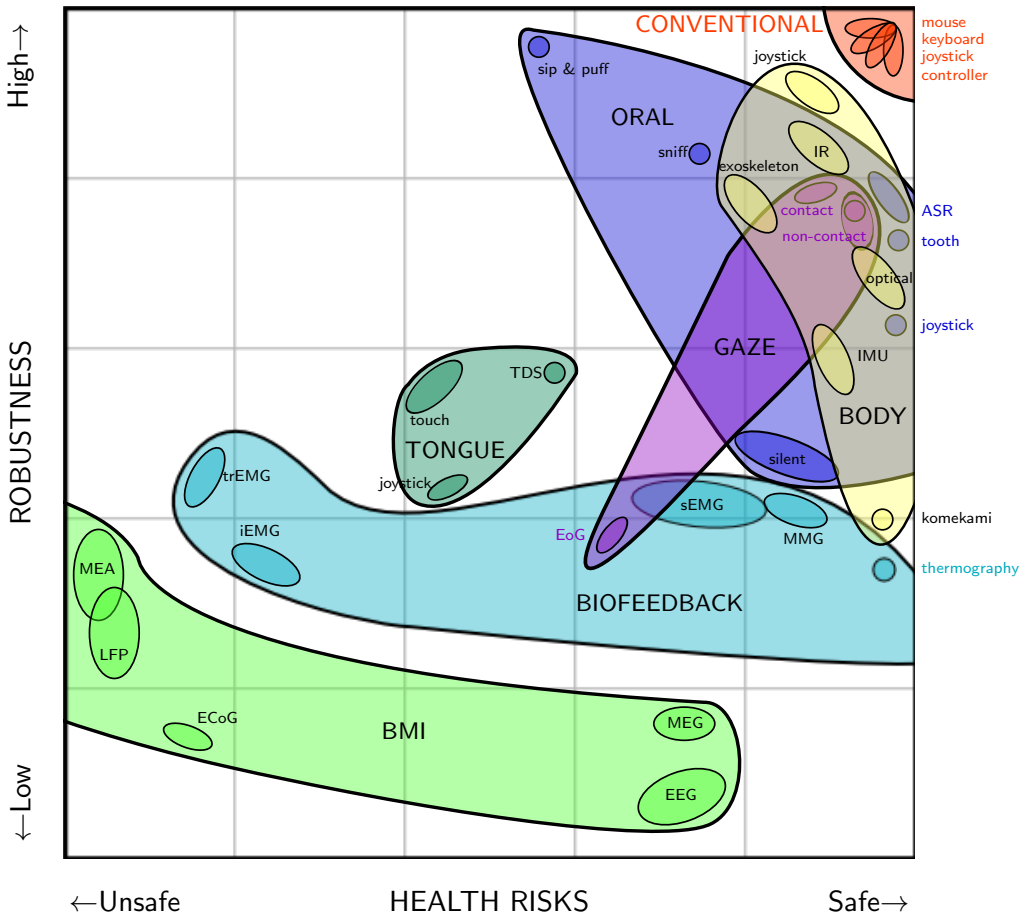


Figure 2.14: Feasibility and functionality: Health risks vs. robustness

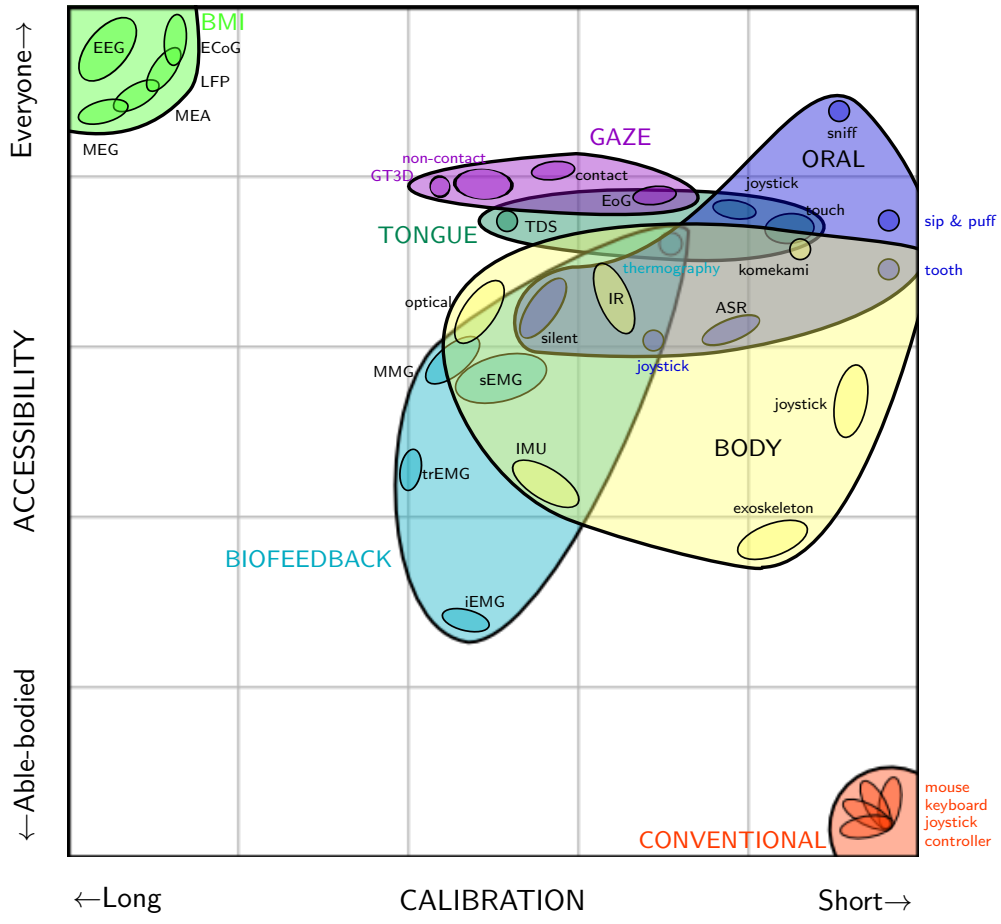


Figure 2.15: Useability and form: Calibration vs. accessibility

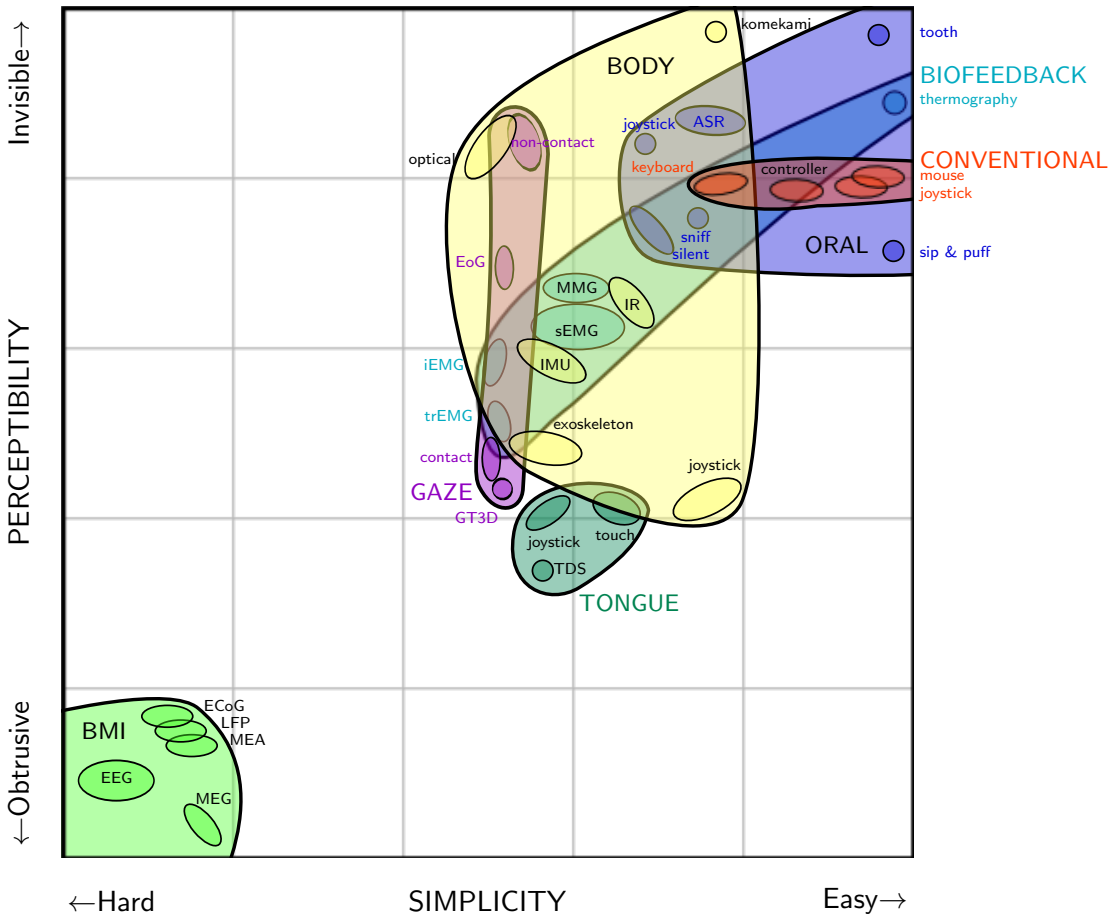


Figure 2.16: Useability and form: Simplicity vs. perceptibility

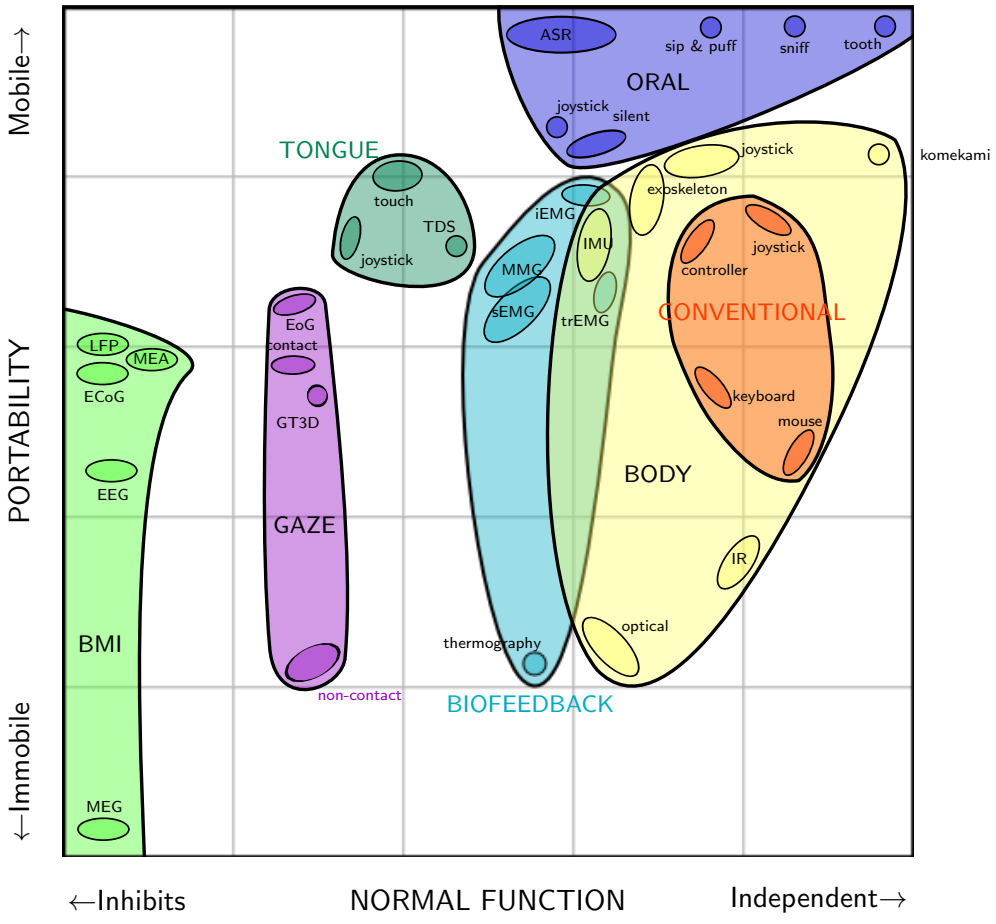


Figure 2.17: Useability and form: Normal function vs. portability

the 'komekami' switch, in regard to many of these factors. These relative positioning not only highlight the differences between the current HMI concepts but also indicate that there is still scope for new technologies based on these input methods and those leveraging alternative and novel pathways.

Another useful observation is that some of the boundaries associated with a particular input method grouping can be very localised over the majority of the *key drivers* (e.g. tongue, gaze) while others cover a much wider area (e.g. body, biofeedback). Although this is obviously linked to the availability of sensor sites and physiological constraints intrinsic to the groupings, it is also indicative that there is the need for more creative solutions associated with a particular input method. This observation motivates the development of concepts, such as the TMEP system and bioacoustic activity in general, as they utilise novel physiological mechanisms for capturing volitional information. Furthermore, through careful design considerations many of the disadvantages associated with a particular grouping can potentially and to some extent be alleviated.

2.6 Bioacoustic signal capture

For decades, bioacoustic signals have been used as a highly informative means of assessing visceral function in patients. The ability of low frequency mechanical waves to permeate through the body without significant attenuation allows for completely noninvasive diagnosis, especially for conditions associated with pulmonary function. A major part of its appeal is its cost effective hardware, for example the stethoscope, which is owned and used daily by all medical doctors. The continued usage of bioacoustic signals for medical diagnosis, even in today's modern era of high tech medical instrumentation, indicate the efficacy of such signals for relaying information across the human body. The Wiley encyclopedia of biomedical engineering defines bioacoustic signals, while differentiating them from the study of bioacoutics as, 'a family of signals that includes the sounds produced by the human body, whereas, in general, the study of sounds produced by living organisms is referred to as bioacoustics' [15].

Bioacoustic signal processing applications generally involve quantitative analysis of these types of signal, usually recorded using microphone transducers positioned externally to the body. These sensors are either positioned directly on or close to the skin surface, with the main purpose being completely noninvasive diagnosis of the functioning of a particular part of the body, especially the visceral systems. The majority of research carried out in this area is based around sounds produced from the ANS, for example, the beating of the heart or lung sounds (cardiopulmonary system) and how these sounds change under varying conditions, for example, due to illnesses or stress levels [139]. Having said this, there has been a substantial amount of research that has also been carried

out into the acoustics associated with the SNS. The majority of bioacoustic signal processing research associated with activation of various SNS peripheries, generally assesses muscle activity [75, 140]. Whether initiated through autonomic or somatic functioning, bioacoustic signals share the common property of being generated internally within the body and translating through an elastic nonlinear media such as skin, flesh and/or bone via a combination of longitudinal and transverse vibrations.

As mentioned earlier, HMI concepts involve intentional activation of the target device through either nervous, physiological or motor actions. Bioacoustic signals which bear relevance to this field will be initiated due to deliberate body movements generated through motor actions. Under this premise, bioacoustic signals considered in this section will be those associated with the monitoring of the SNS and will exclude any systems solely intended for the monitoring of the ANS, for example lung sounds.

Apart from conventional microphones, one of the oldest bioacoustic methods of assessing somatic body functions is MMG [75, 140]. It is specifically utilised in the assessment of muscle sounds and the majority of research is based around how the amplitude of the recorded bioacoustic signal relates to muscle force. In the past it has also been known under alternative aliases including acoustic myography, sound myography, phonomyography and vibromyography [75]. MMG signals can be captured using either microphones, accelerometers or a combination of the two [141]. A coupled accelerometer is sometimes employed to mitigate the effects of motion artifacts due to limb motion for example. The transducers are fixed within a stiff case consisting of a airtight silicon chamber with a flexible membrane sitting between skin and air chamber. The enslosed chamber act as the mechanical coupling between the muscle and transducer. As the muscle contracts this membrane is displaced, thus changing the pressure inside the chamber, which is subsequently sensed by the microphone. The characteristic frequencies associated with MMG signals during a typical muscle contraction is in the 5-50Hz range with the nominal frequency at 25 ± 2.5Hz [74, 75]. MMG was devised as an alternative to EMG and exhibits significant advantages over it. These include requiring no conductive paste between skin and electrode contact as well as minimal crosstalk from disparate muscle groups. However, as MMG sensors measure the mechanical response rather than the electrical response, the MMG signal will generally lag behind it. Recently, MMG-based assistive HMI systems have been developed for controlling hand prosthesis and as a switch utilising small eyebrow movements as commands [74, 142]. Using a two site MMG socket, which locates the sensors against the forearm of the individual, two below elbow amputees were able to voluntarily open and close the grasp function of a hand prosthesis using physiologically relevant contractions [74]. Alternatively, by locating an MMG sensor on the forehead of an individual and utilising a continuous wavelet transform (CWT) to process the signal, eye blinks could be detected with extremely high sensitivity and specificity. These movements

can therefore be used as switch control for extremely disabled individuals [142].

Over a decade ago, the US army research laboratory developed the acoustic gel sensor (AGS) which was intended for physiological monitoring of first responders, soldiers and fire-fighters [139, 143]. The distinctive feature of the sensor is a gel bladder which sits on top of a piezoelectric element, thus acting as a medium between an individuals skin and the piezoelectric transducer. The gel bladder is designed to have similar acoustic impedance properties as that of human skin and the underlying flesh. This allows transmission of body sounds when the gel-coupled sensor is contacted with a specific anatomical site. The acoustic impedance is a measure of how readily the particles of the conducting medium can be displaced by the sound waves. An increase in the disparity between two mediums' impedance implies that more of the sound wave will be reflected at the interface between them [144]. This impedance match between skin and bladder also is indicative of the impedance mismatch between the air and sensor, implying a significant level of airborne acoustic noise rejection. The original purpose of the sensor was to track both autonomically and somatically regulated activities for health and performance monitoring. This included the monitoring of physiology including heartbeats, breathing and blood pressure (*autonomically regulated processes*), while speech and motion, including footfalls and crawling, could be tracked during prescribed missions (*somatically regulated activity*). It can be seen that the signals of interest can be generated and translated from a broad spectrum of anatomical areas. The ability of the human body to transport these mechanical oscillations allow the transducers to be positioned away from the source of the bioacoustic sound in discrete and unobtrusive locations. The positioning of the gel-coupled sensors included placement against the neck, chin, wrist and temple of the individual and was often integrated into the user's equipment, for instance, the chin strap of their helmet.

Bone conduction microphones (BCM) provide an alternative but related avenue for recording and monitoring of bioacoustic signals. They are differentiated from standard bioacoustic microphone recordings through placement on bone rather than muscle or flesh. Investigations into bone conduction and the principles guiding it have been carried out for more than a century but only recently have technologies been developed based on these principles [145, 146]. Simply put, bone conduction is the translation of sound to the inner ear through the bones of the skull. This is possible due to the ability of humans to hear and perceive sound in two ways, namely, through airborne and bone-conducted vibrations. Airborne vibrations provide the majority of sound which allow mammals to hear. Normal human hearing is based on sound waves travelling through the outer and inner ear where they vibrate the tympanic membrane. This in turn oscillates the connected ossicular chain (comprising the malleus, incus and stapes) thus behaves as a lever-like system. This in turn moves the flexible membrane of the cochlear containing the fluid of the inner ear. Within this fluid filled chamber, resides a structure called the organ of Corti,

which contains over fifteen thousand hairs. These hairs react to and encode the relative motion of the fluid, converting it to an electrical nervous signal. In contrast to this, bone conducted sound bypasses the outer and inner ear, travelling along the skull, directly setting up waves inside the cochlear. The reason airborne sounds takes precedence is due to the acoustic impedance mismatch between the skull and air. The main role of the middle ear (tympanic membrane and ossicular chain) is to provide this impedance matching while sound which bypasses this has significant attenuation due to the skull and cochlear interface. This means that sound energy is not transmitted efficiently from skull to cochlear and therefore contributes less to an individuals perception of it. Furthermore, this bone conduction system acts as a low pass filter and effectively attenuates higher frequency airborne noise both internally and externally.

BCM sensors were originally designed to capture speech signals, in high noise environments, in a manner that is both discrete and highly portable. They could potentially be used to capture a whole range of bioacoustic signals, generated internally within an individual, through this impedance matching between the microphone and skull. This allows them to efficiently capture bone conducted bioacoustic signals with this mechanism lacking at the skull-cochlear interface. A significant advantage over conventional microphone systems, for example boom microphones, is their effective use in noisy and crowded environments while being both unobtrusive in nature and can additionally be utilised while the speaker is on the move [147]. A related technology are throat microphones which provide another way of capturing speech directly from the vocal chords. The throat microphone is attached to the neck directly on top of the skin covering the vocal chords and as such is not as susceptible to ambient noise as standard air microphones [148]. Although throat microphones can also be used when the talker is in motion, they are considered inferior to air microphones in terms of their transmission capabilities and can additionally produce user discomfort.

More recently, a new technology based on finger tapping on the surface of a users forearm has been developed [17]. This system, named Skinput, was intended as a new means for interactive human-computer interaction (HCI) in a manner that is both naturally portable and always available. The technology revolves around an array of uniquely tuned acoustic sensors based on cantilevered piezoelectric films. They are designed to have different resonant frequencies by adding small weights to the end of each cantilever. The sensors are positioned axially around the circumference of the arm through integration within a prototype armband. This is located either above or below the elbow. The sensors are naturally responsive to motion perpendicular to the skin surface implying enhanced capture of the transverse and longitudinal waves associated with finger taps emanating from the surface of the forearm. Motions parallel to the skin surface for example shearing motions associated with skin stretch, are rejected due to its design. Different

configurations of tap locations were evaluated and were classified using 186 features and a support vector machine (SVM) classifier. The features included amongst others, average amplitude, standard deviation and total energy associated with the waveform of each channel. This technology highlights the potential for volitional bioacoustic activity generated through impulsive taps of the forearm, to be used as an efficient form of human-computer communication. Furthermore, the impulsive nature of the commands themselves allow for a responsive interface that could not be achieved using speech recognition, for instance. The positioning of the armband above the elbow allows for the whole forearm to used for command input and implies that the sensing hardware does not interfere with command generation. This is possible due to the ability of the arm to transport mechanical vibrations in the form of bioacoustic signals.

2.7 Chapter summary

The predominant focus of this chapter is to give the reader an overview of the current state-of-art in assistive human-machine interfaces. These interfaces provide a means of communication for the physically challenged, with the development of these systems, not only motivated by the broad spectrum of impairments, but also the sheer size of this demographic. For comparison among the various systems, ten *key drivers* have been proposed by the author and relate, not only to the feasibility and functioning associated with the devices, but also their useability and form. Such considerations are important from both the perspective of the designer and end user.

The various interfaces have been grouped according to their input method including, body, biofeedback, gaze, brain, oral and tongue based systems. They have been separated in this manner as this is considered the most pertinent consideration when designing or selecting an assistive communication device. Within each of these categories, various concepts have been highlighted and are generally differentiated by the hardware involved, location of the transducers relative to the individual and the physiological command generated. The described interfaces have been taken from both the commercial and research domain. It should be noted that some are highly specialised for a particular type of impairment while others attempt to appeal to a broad range of physical disability. For each device, the main pros and cons have been highlighted and are summarised. In terms of the *key drivers*, the interfaces have been compared relative to one another and visualised through the use of two-dimensional HMI comparison charts. They have been grouped according to their input method, with an additional conventional interface group, comprising computer mouse, computer keyboard, joystick and controller, included as a benchmark. These visualisations allow both the user and HMI designer to compare the current state-of-art in a quick and intuitive manner.

As this work is investigating using volitional bioacoustic signals for command of assistive devices, various methods for capturing bioacoustic signals have also been discussed. This has predominantly been in the context of intentional or somatically created activity. By bringing such activity to the attention of the reader, it helps to elucidate what exactly bioacoustic activity is and the multitude of ways in which it can be created and captured.

Chapter 3

Background to bioacoustic tongue-movement ear pressure signals

The basis of this chapter is to provide the reader with a description of tongue-movement ear pressure signals, in terms of their generation, evolution and sensory detection within the human oral-aural system. The initiating actions will be described and illustrated, enabling the reader to envisage the start point and trajectory of the tongue during the predefined motions. The location and placement of the microphone and associated housing within the ear canal will be detailed, emphasising the notion of a disruption-free oral cavity. Due to the highly complex and sensitive nature of the vibration mechanics during production of the TMEP signals, it is unfeasible for any useful modelling of the physiological processes to be performed. Instead, a bioacoustic translation framework is proposed based on a black-box approach to discriminate between the subject-specific signals. To this end, bio-acoustic characteristics have been detailed, specifically time and frequency domain information, associated with the microphone signals during execution of the specific tongue movements. These help in highlighting the feasibility of inter-action discrimination. The hardware associated with the TMEP concept is presented with a comprehensive description of both a generic and custom earpiece design. Finally, the collection procedure associated with the off-line data sets is also detailed.

3.1 Physical and physiological description

The potential of tongue-movement ear pressure (TMEP) signals for control of assistive or robotic devices has been highlighted previously [24, 25]. The concept itself centres on a microphone transducer being placed within the ear canal of an individual, specifically

their external acoustic meatus. The earpiece is located and retained using an associated housing, with this housing either being of generic or custom stock with both described fully in Section 3.4.2. The position of the earpiece within the ear canal and its relation to the physiology of the ear is shown in Fig. 3.1 [24]. This schematic shows the earpiece with an external microphone, envisaged for use as feedback during adaptive noise cancellation of external interference. However, the passive attenuation properties of the earpiece were sufficient to make this realisation unwarranted and therefore no earpieces used during this research have an additional external microphone.

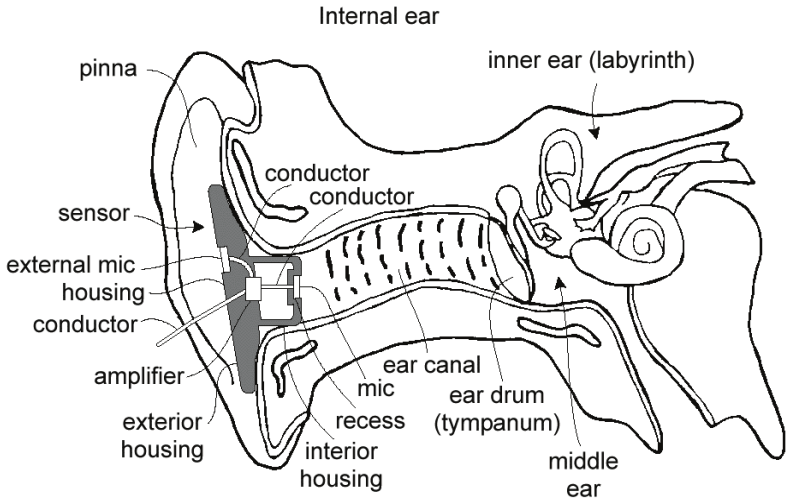


Figure 3.1: Annotated schematic of the ear highlighting the important anatomical features and showing the position of the earpiece in relation to it (taken from [24])

The main function of the earpiece is to locate the microphone within the ear canal so that (1) it sits in a position so as to maximise the TMEP information captured, including all frequencies associated with the TMEP bioacoustic signatures, (2) for passive external interference rejection of ambient noise (thus helping to increase the SNR) and (3) so that the device feels comfortable for the user to wear for extended periods of time. Although the earpiece, to some extent, obscures the user's hearing, this attenuation is a similar level to that of a standard (unplugged) in-ear headphone.

It has been been established that by performing prescribed motions of the tongue, which are best described as 'acoustic flicks', unique and repeatable bioacoustic signals are created within the ear canal [24]. The waveforms associated with these signals are impulsive in nature, allowing them to be detected and distinguished from normally occurring acoustic signals. Although comprehensive knowledge of the physical mechanisms underlying this phenomenon is unknown to the general research community, the following description is based on the author's conjecture[1]. The first mechanism is that the movement

[1]The phenomenon described being the initiation and translation of acoustic signals within the ear canal

of the tongue causes airflow and thus acoustic pressure changes within the oral cavity, which then migrate along the Eustachian tube (ET) and up into the ear canal. This in turn exerts a transient pressure differential directly onto the diaphragm of the microphone. The second and probably the most dominant mechanism, is that due to the physiological connections of the tongue to the rest of the head, that is, through the hypoglossal muscle and mandible, the movement of the tip of the tongue against the lower gum causes mechanical vibrations to be generated and travel along these flesh and bone mediums. These vibrations propagate along the maxillo-facial and temporal areas of the head, causing the microphone housing to vibrate relative to the microphone diaphragm. Furthermore, the function of the ET is to keep the middle ear clear of pathogens and fluids whilst also keeping it ventilated, it generally exists in a closed configuration and only opens periodically[1] [149]. For this reason, the secondary mechanism is thought to be the most important contributor and thus heavily influences the morphology and information content of the sensed bioacoustic signal.

To utilise this phenomenon effectively within a HMI setting, it is desirable to expand the instruction set thus making it easier for the user to communicate with peripheries. The impulsive nature of the signals allow for a rapid response due to all the signal information being contained in a short duration of time. The system acts as an asynchronous or self-paced HMI, initiating a machine response anytime an impulsive TMEP signal is detected. If a variety of machine responses can be initiated, then the communication time between human and machine can be greatly reduced. However, there is generally a trade-off between the number of permissible instructions and the accuracy (number of correctly classified instructions) of the system. Consequently, a basic instruction set of up to four distinct actions has been chosen during TMEP recognition as this maintains disciminability between the specified actions while intuitively mapping to a standard control scheme of left, down, up and right. The action (or instruction) set consists of placement of the tip of the tongue at the base of the central incisor, left or right first molar and flicking the tongue up (left, bottom or right action) and placing the tip of the tongue against the top of the palate and flicking down (top action). Figure 3.2 (left side) indicates the initial position of the tip of the tongue in relation to the upper and lower jaw while Fig. 3.2 (right side) gives an idea of the trajectory of the tip of the tongue during execution of the action itself.

While the actions feel natural to execute they are not normally occurring tongue motions. This is important as they need to be distinguishable from normal tongue activity that occurs during daily living, for example, those which occur during speech and/or mastication. However, this does also mean that there is a learning period associated with mastering the actions and, while this varies person-to-person, is generally in the range of a few days. The main requirement from this learning period is that an individual can reli-

generated by impulsive flicks of the tongue

[1]The ET connects the nasopharynx to the middle ear

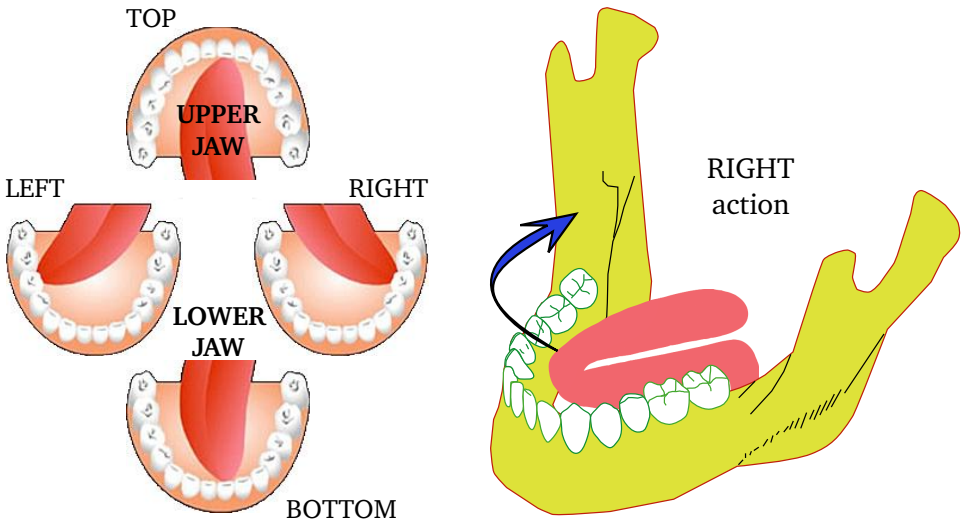


Figure 3.2: Left side - Starting point of the tip of the tongue prior to the execution of the four predefined actions. Right side - 3D representation of the idealised trajectory of the tip of the tongue during execution of the right action

ably perform each action so that the waveform morphology is consistent. They also need to ensure that a certain SNR is achieved during execution so that the TMEP event can be detected within normal bioacoustic activity. Other tongue motions were briefly explored, for instance, rubbing the tongue across the underside of palate or lower jaw. However, these were found to have insufficient SNR and no consistent time or frequency signature.

3.2 Previous work

Some of the research outlined in this thesis builds on previous work, published in 2007, by Vaidyanathan et al. [24, 25]. As such, it is useful to give a brief overview of these two journal papers.

The first paper entitled, '*Tongue-movement communication and control concept for hands-free human-machine interfaces*' introduces the TMEP concept and gives an overview of the system in terms of the four generating actions and the signal acquisition setup [24]. Single channel TMEP signals were collected using custom earpieces in a controlled off-line environment. See Section 3.4.2 for a full description of the earpieces. This meant detection of volitional activity and signal segmentation were also performed off-line and optimised for the data sets. For signal estimation the original pairwise cross-correlation averaging algorithm (developed by Gupta) is presented and will be developed further, in Section 4.4, thus increasing the algorithms generality, robustness and stability. The main emphasis of the paper is to demonstrate the feasibility of discriminating between the four tongue actions and evaluation across various classifiers. The following polychotomous classification

strategies were compared, namely, the matched filter classifier, autoregressive features with a multivariate Gaussian classifier, a nonlinear alignment classifier and a new decision fusion classifier (DFC). Evaluation was based on confusion matrices and average classification accuracy across eight well trained individuals and showed that over 97% correct classification was achievable using the DFC under controlled conditions.

The second paper entitled, '*A dual mode human-robot teleoperation interface based on airflow in the aural cavity*' highlights the feasibility of using bioacoustic signals captured in the ear for command of robotic systems [25]. Specifically, the machine interface system is conceptualised and an initial off-line simulation is presented based on teleoperation of a mobile robot platform maneuvering through a constrained environment. Initially, a simple acoustic model is developed of the ear canal, assuming that the bioacoustic activity is generated through airflow, demonstrating the sensitivity of the ear canal as an acoustic output device for human-machine interfaces. Two operating modes are envisaged for command of the robot platform with the instruction sets consisting of either dragon speech or compound tongue movements. The speech set consists of the words 'up', 'down', 'left', 'right', 'kill', 'pan', and 'move'. The tongue action instruction set involves performing compound sequences of the tongue actions thus enabling more commands to be given. For example, when only two actions are being used the following simple and compound commands could be issued: left, right, left-left, right-right. This increases the size of the instruction set albeit at the cost of the instruction execution time. For speech classification, based on mel-frequency cepstral coefficients (MFCC) and a hidden markov model (HMM), the average off-line accuracy was 92.77% across twenty healthy subjects[1]. For tongue classification, the average confusion matrix from the eight subjects presented in the previous paper were utilised. Alongside these results, dual channel off-line data was collected from a single untrained subject performing two tongue actions. Using a multi-channel DFC algorithm, 98.91% average classification accuracy was achieved during this dichotomous discrimination. The paper concludes with a simulation performed off-line with commands selected based on the probabilities from the average multinomial confusion matrices. The simulation was performed using either the average speech performance parameters or the average tongue action performance parameters across the four actions with two compound movements providing additional control. This simulation demonstrated the feasibility of hands-free teleoperation of a mobile robot based on both speech and TMEP signals.

Ultimately, both papers have provided a 'proof-of-concept' indicating the potential of TMEP signals within HMI applications. In the context of the TMEP system, further research that would extend these initial results include the following:

- Development of a real-time framework based on efficient processing modules
- Full real-time evaluation of the discrimination performance of TMEP signals in col-

[1]MFCC and HMM are common speech processing tools with further details given in Section 2.4.5

laboration with this framework

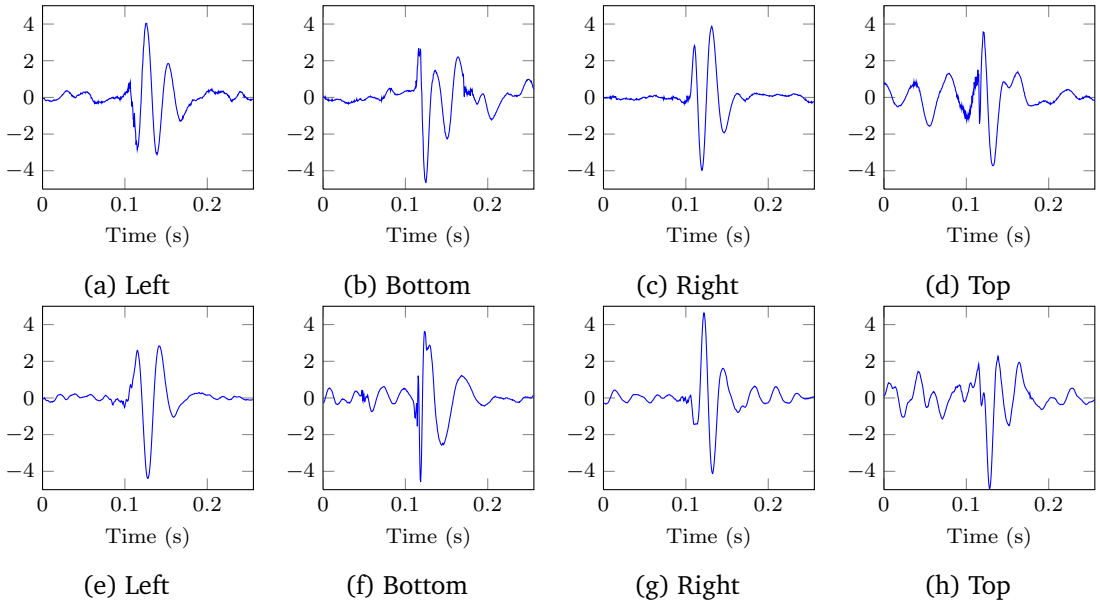
- Testing in challenging situations where internal and external interference merges with the volitional activity
- Characterisation of the performance under various test conditions, for example, number of instances used during training or the amount of training a subject has experienced
- Comparison of alterative decoding techniques which may prove more resilient to the decoding degradation experienced during real-time operation
- Control of a real world peripheral, evaluating its performance during an everyday task
- Development of a completely wireless system for untethered monitoring of the aural cavity

These extensions have been highlighted by the author and to some extent have been addressed in this thesis and associated research.

3.3 Bioacoustic characteristics

The typical duration of a controlled TMEP action is 200 ms with all noticeable oscillations associated with the impulsive waveform completely decayed after this time frame. Controlled (TMEP) actions are differentiated from non-controlled or interfering actions based on whether the user performed a volitional movement of the tongue with the intention of producing an output in the HMI system. If activity is detected but was not intended by the user, that is, a non-controlled action occurred, it is considered as internal interference. Figure 3.3 shows example TMEP segments for single instances of the four previously described actions from two subjects (subject A and subject B). These signals were captured using custom moulded earpieces and sampled at a rate of 2 kHz (see Section 3.4). They were collected in a standard office environment with no inhibition or reduction to the background noise. It is clear to see that there are distinct and unique morphological features associated with each action type. Figure 3.3 not only highlights the distinct nature of the signals action-to-action but also highlights the dissimilarity between the actions subject-to-subject. Regardless of these differences, if the distribution of signals associated with a single action (intra-class variance) is larger than the differences between the actions (inter-class separability) then effective discrimination of the actions would not be possible. To highlight that repeatable signals, exhibiting low intra-class variance, can be produced by a subject, Fig. 3.4 shows a hundred repeated instances of the four actions

for subject A. The distinct repeatable nature of the actions is especially visible if the same signals are aligned to one another, as shown in Fig. 3.5. The inter-class and inter-subject differences reinforces the notion that any modelling of the underlying mechanisms would prove futile due to the sensitive nature of the signals to physiological changes, initial start point of the tip of the tongue and differences in execution of the actions between subjects. This implies that subject-specific training data is required for inter-class classification of the actions based on some kind of pattern recognition machine.



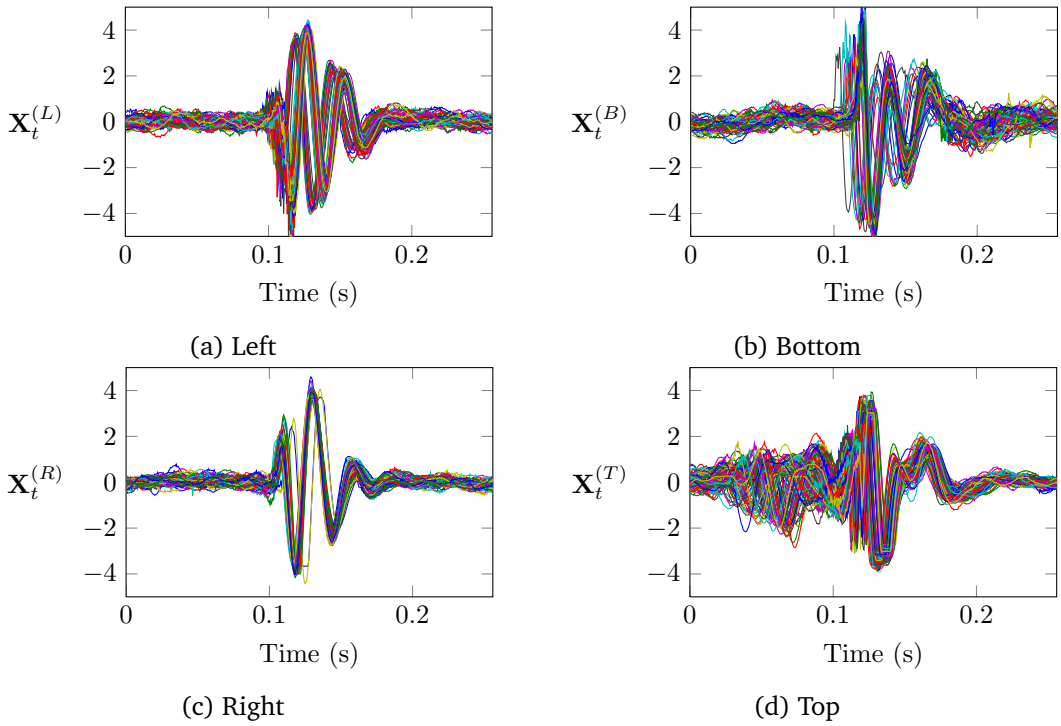
(a) Left (b) Bottom (c) Right (d) Top

(e) Left (f) Bottom (g) Right (h) Top

Figure 3.3: Example TMEP signals from the four prescribed tongue actions (left, bottom, right and top) for two subjects (subject A - top row, subject B - bottom row).

The majority of the frequency content associated with intentional TMEP actions is in the 0 - 100 Hz frequency range. This can be considered the critical frequency band of the TMEP signals. There is little frequency content above 200 Hz which can be generally considered related to ambient noise or non-controlled TMEP activity. Figure 3.6 shows 3D power spectral density (PSD) plots for one hundred instances of each of the four actions for subject A. The PSD was estimated using a modified periodogram based on Welch's method using a Hamming window with eight overlapping sub-segments of the signal. These frequency responses again highlight the intra-class consistency of the data sets. Figure 3.7 shows the average PSD (mean $\pm$ 1 SD) across the data sets for the four tongue actions for subject B. It can be seen, within the 0 - 100 Hz critical band, the peak of the signal power is generally within the low frequency sub-band of 20 - 40 Hz. Furthermore, the shape of the frequency responses generally allude towards a mono-component structure.

These intentional TMEP signals when compared to typically occurring internal acoustic interferences, such as speech, emphasises the lower frequency nature of the volitional



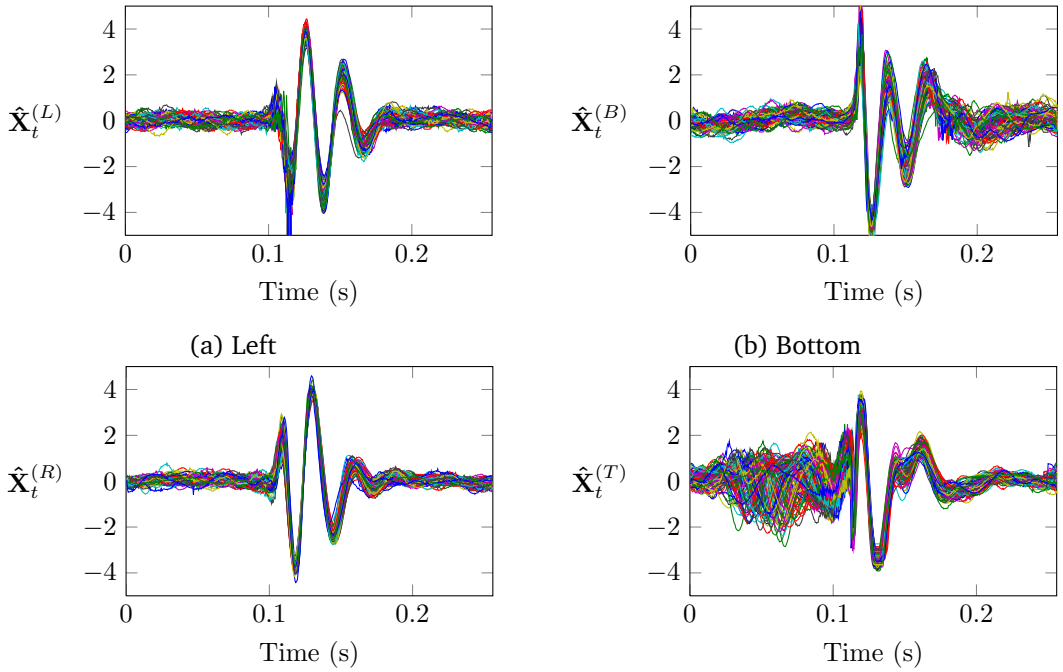
(a) Left (b) Bottom

(c) Right (d) Top

Figure 3.4: Example data sets of four tongue actions (left, bottom, right and top) from subject A. Each data set consists of one hundred instances of each of the actions.

activity. The speech or voice frequency band is generally considered up to 4 kHz when recorded using a microphone positioned in front of the mouth [111]. The fundamental frequency of voiced speech ranges from 85 - 180 Hz and 165 - 255 Hz for a typical adult male and female, respectively. Figure 3.8 shows waveform, spectrogram and PSD plots of the voiced phrase 'tongue-in-cheek' captured simultaneously from a microphone positioned in front of the mouth (3.8a) and one positioned within the ear canal (3.8b) using a generic earpiece. The sampling rate is set at 44,100 Hz and the microphones used during the recording are identical and described fully in Section 3.4.2. Signals recorded in the ear have much of the higher frequency oscillations damped out due to the maxillo-facial region acting as a passive low-pass filter. The frequency range associated with speech captured in the ear is up to 1 kHz due to the effect of the within ear capture. For this reason the fundamental sampling rates chosen during the analysis of TMEP signals is generally set at 2 kHz. Retaining higher frequency components above 100 Hz is important as it is this information that provides a basis for rejection of higher frequency non-controlled TMEP signals such as speech and coughing [27]. The low frequency (0 - 100 Hz) range of the in-ear microphone also shows an increase in power highlighting the enhanced low frequency response that the TMEP concept provides.

The distinct nature of the signals in terms of their time and frequency domain morphologies highlights the potential for inter-action discrimination. This is possible as the



(a) Left (b) Bottom

(c) Right (d) Top

Figure 3.5: Example aligned data sets of four tongue actions (left, bottom, right and top) from subject A. Each data set consists of one hundred aligned instances of each of the actions.

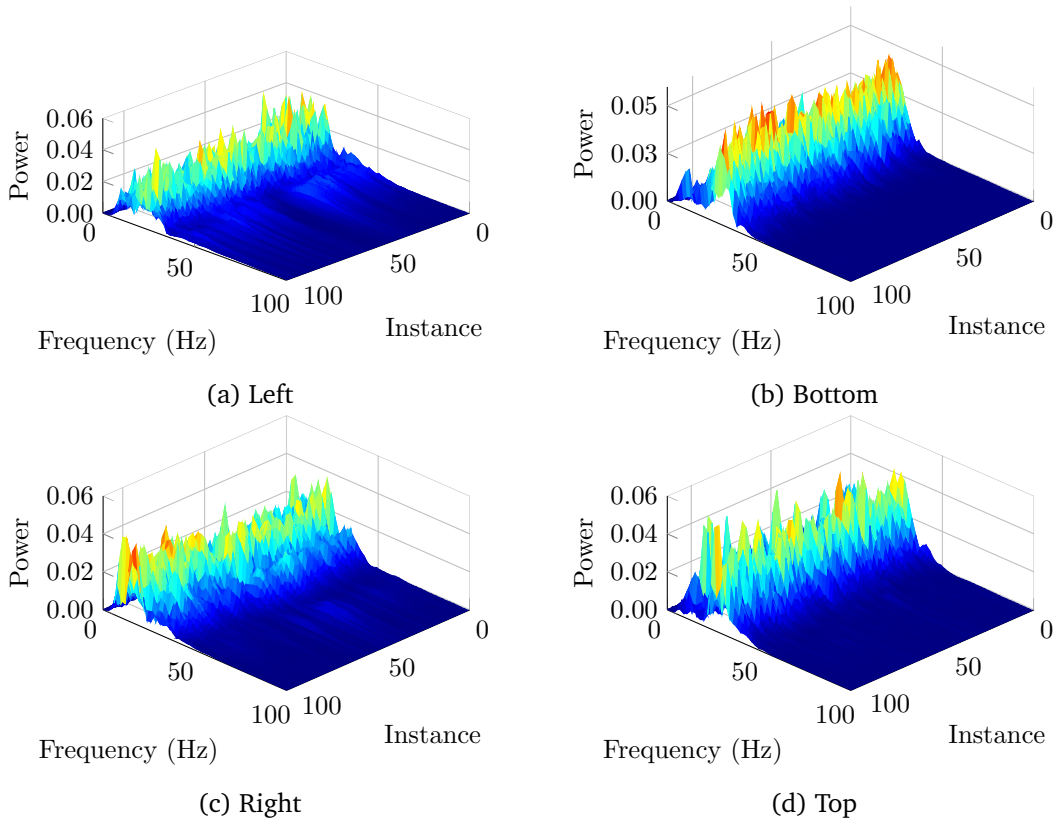


Figure 3.6: Power spectral density plots of data sets associated with the four actions for subject A. Each instance is shown as a separate contour of the PSD surface

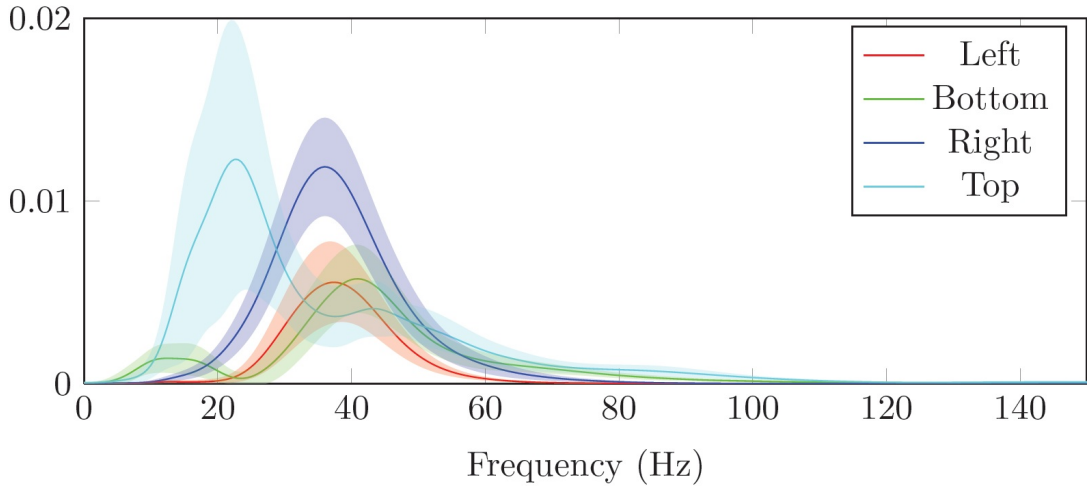
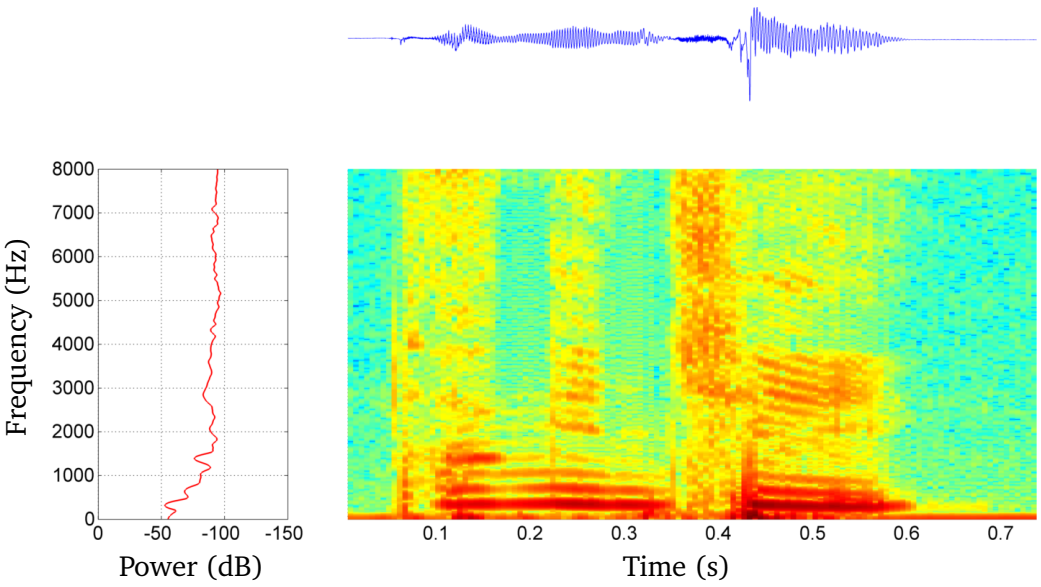
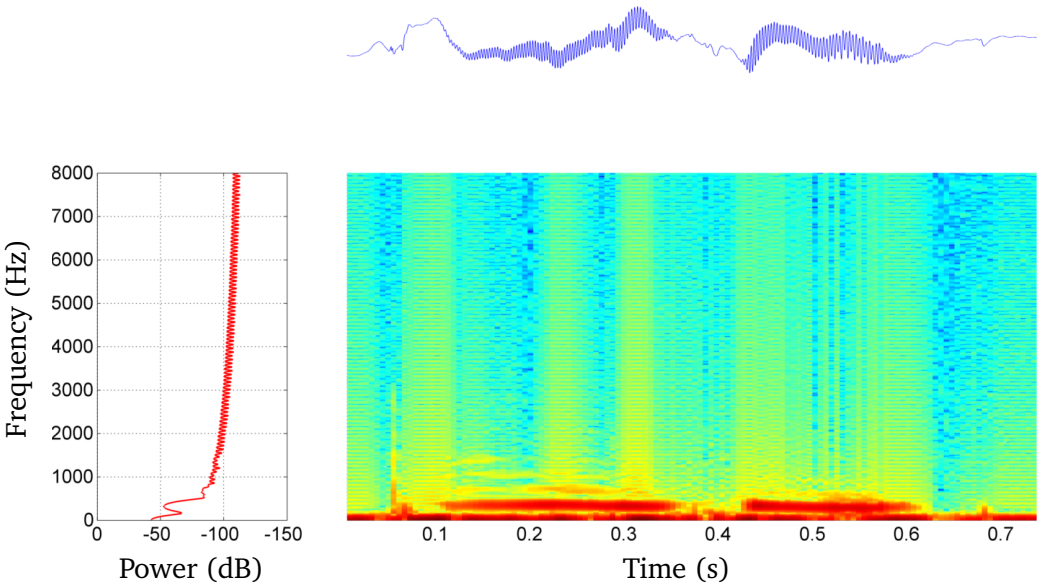


Figure 3.7: Average power spectral density plots with four actions from subject B. Each average PSD line is the mean of the one hundred PSD plots shown in Fig. 3.6



(a) Front of mouth



(b) In ear

Figure 3.8: Comparative plots showing waveform, PSD and spectrogram of voiced speech for the phrase 'tongue-in-cheek' captured simultaneously at a sampling rate of 44.1 kHz by a microphone placed in front of the mouth and one located within the ear canal of an individual

data sets show consistency across instances from the same class. Therefore the combination of both high inter-class separability and low intra-class variance indicate that recognition of the proposed actions is possible.

3.4 System

3.4.1 Hardware

The physical components of the TMEP hardware required to map the continuous stream of bioacoustic information to a useful machine output is shown in Fig. 3.9 and consists of three main parts plus two additional signal conditioning blocks. The main blocks are shown in blue while the signal conditioning blocks are indicated in green. The hardware consists of a microphone transducer to convert the bioacoustic and mechanical vibrations to an electrical voltage, this is followed by the signal conditioning blocks, namely, pre-amplification to boost the audio signal strength followed by a low-pass anti-aliasing filtering. The final two blocks required are the analogue-to-digital convertor (ADC) to digitise the signal and processor which executes the signal processing, extraction and decoding algorithms.

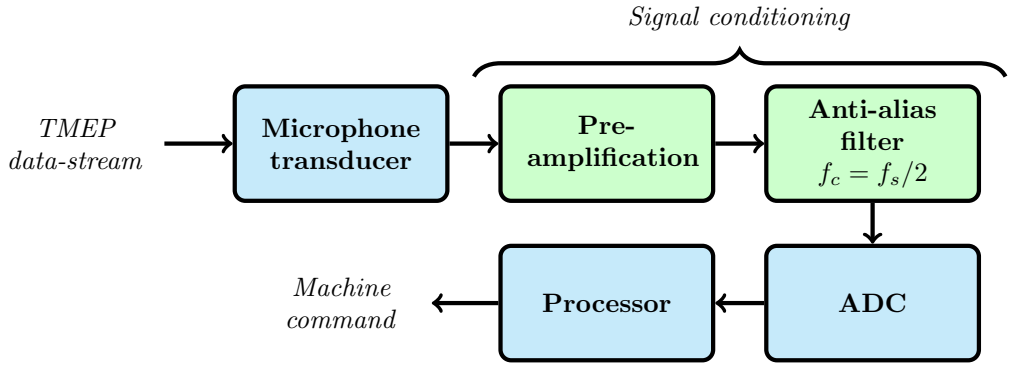


Figure 3.9: TMEP hardware block diagram

In this work, all the digital signal processing has been performed on a standard laptop running MATLAB v7.12. The ADC utilised is a Pure-audio USB-SA as this allows for a true (2-channel) stereo input at 16-bit resolution[1]. The stereo input allows for a dual channel setup, with an earpiece capturing bioacoustic data simultaneously from both ears. Fusing two channels of TMEP data, through a heterogeneous ensemble framework, increases the decoding power in terms of both sensitivity and specificity (see Section 5.11). The USB-SA is an affordable solution (approximately £40), enhancing the cost effective and 'plug and play' benefits of this system even at this early stage of research and development. Hardware specifications associated with the USB soundcard are given in Table 3.1. The

[1]Andrea Electronics, NY, USA

analogue anti-alias filter is implemented onboard the soundcard and is set at the Nyquist frequency.

Specification	Value
Power	USB
Microphone input	2 channel 'true' stereo
Sampling rate	48 k/44.1 k/16 k/8 kHz
Supply voltage	4.5 - 5.5 VDC
Total power consumption	120 mA
A/D conversion resolution	16 bit
THD + N	-84 dB
Supply bias resistor	2.2 KΩ @ 3.3 VDC
Frequency response	20 - 20,000 Hz
Input voltage range	0 - 1.25 Vrms
Dynamic range	95 dB
Record gain range	-6 to 33 dB

Table 3.1: Andrea electronics Pureaudio USB-SA soundcard specifications

As the TMEP signals occur at a much lower frequency than speech, it is important that the sensitivity of the microphone is retained even at these lower frequencies. Although very low frequency microphones (<20 Hz) are available they are prohibitively expensive. Standard microphones will have some roll-off at lower frequencies in the 0 - 100 Hz range. This roll-off is generally unavoidable but should be kept to a minimum. The Knowles FG-23652-C36 was chosen as it exhibits minimal roll-off at these lower frequencies. It is a unidirectional electret condenser microphone and as such, is a low powered and cost effective solution. The frequency response of the microphone is given in Fig. 3.10 and is generally flat above 100 Hz. The roll-off in the 20 - 100 Hz frequency range is approximately -4dB constituting a decrease in the voltage by a factor of only 0.631 over this frequency range.

3.4.2 Earpieces

As stated previously, two main types of earpiece are utilised within this research, namely a custom stock earpiece and a generic stock earpiece. The custom earpiece utilises a subject-specific housing for the microphone, with this casing resembling the well-established housing of a standard hearing aid. In one embodiment, the external portion fits within the concha and is retained, at least in part by the tragus and/or the antitragus. A photo of the earpiece, inserted and removed from the ear is shown in Fig. 3.11. The process of producing the bespoke earpiece initially involves taking an impression mould of the individual's ear using a vulcanising ear impression silicone. The impression task is performed by a qualified audiologist and involves examination of the ear canal using an otoscope, fol-

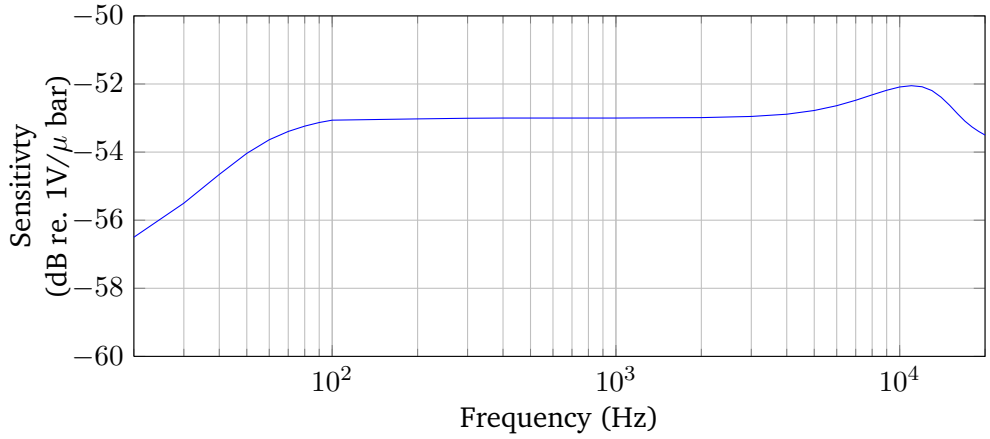


Figure 3.10: Frequency response of microphone used during TMEP signal capture

lowed by insertion of an otoblock to protect the eardrum by preventing contact between the silicone and tympanic membrane. The silicone material is then mixed thoroughly with an activator which begins the hardening process. Before this hardening occurs, it is inserted into a syringe and plunged into the the opening of the ear canal, completely filling the concha. The impression is pressed down firmly ensuring a tight fit, with the individual instructed not to move while the impression sets within their ear.



(a) Earpiece inserted

(b) Earpiece extracted

Figure 3.11: Photos showing the custom earpiece both inserted and removed from the ear

Once the mould has been taken it is sent to a hearing aid manufacturer so that the hollow housing unit can be commissioned[1]. The housing is made from a hard, durable plastic material and is fabricated using a resin casting based on a negative mould of the interior portion of the impression mould. The majority of the housing is wider than the opening of the ear canal, so as to engage the concha or alternatively, to cover the whole

[1]Hearing Technologies International, USA

pinna and in this respect is similar to a hearing aid or hands-free mobile phone device. The wall of the housing is approximately 1.5mm thick and a bore for the microphone outlet is cut into the end that engages the ear canal, with this interior portion reaching a depth of between 6.5 - 12.5mm. The microphone and connecting cable are then positioned inside the housing, with the open port of the microphone placed against the bore of the housing described earlier. The microphone is glued into position and a soft putty is used to fill the remaining internal space inside the housing thus protecting the microphone from picking up any reverberation effects. Finally a backing (with hole for connecting cable) is glued on the exterior portion of the housing sealing the microphone and putty inside it.

The custom mould earpiece fulfills a multi-purpose role, firstly, it allows for the insertion of the microphone deeper within the external acoustic meatus of the individual. This allows the microphone to be placed closer to the S-bend of the ear canal, helping to capture a cleaner signal. Secondly it locates and positions the microphone (in terms of depth and orientation within the external acoustic meatus) very consistently during removal and reinsertion of the earpiece. This is due to the unique shape of an individuals outer ear, which only allows the earpiece to be inserted in one orientation. Both the concha and fossa help to secure the microphone in place once it has been inserted. Lastly, it creates a tight seal within the opening of the ear canal which increases the passive ambient noise attenuation and enhances the low frequency response of the earpiece. However, this has the additional disadvantage of reducing the hearing ability of the individual.

The generic earpiece, provided by Think-a-move LTD, utilises a general housing for the microphone, which allows for a smaller housing to be specified due to the concha and tragus no longer providing retention of the earpiece[1]. Instead the earpiece is held in place via a compliant foam tip, which is compressed before inserting into the ear canal and allowed to expand inside the ear canal thus securing it in place[2]. The housing itself it formed of a hard acrylic shell, with the open port of the microphone extending out of this shell via acoustic tubing. The cable connecting microphone to AD converter protrudes from the other end of the housing. The acoustic tubing is rigid and has an external screw thread cut into it, so that the foam tip can be screwed securely to it. A bore extends the entire length of the foam tip allowing acoustic signals to travel down it and to the microphone. Fig. 3.12 shows the generic earpiece both inserted and removed from an individuals ear.

The use of a generic earpiece has the obvious advantage that it is subject independent and therefore ensures that cost and lead-times are minimised. It also means that the same earpiece (and microphone) can be used by different subjects. Furthermore, the discrete nature of its housing implies that is very unobtrusive in appearance and similar to how a normal music headphone appears when worn. This is due to a much smaller

[1]Think-a-move LTD, Ohio, USA

[2]Comply foam tips, Hearing Components, USA



(a) Earpiece inserted



(b) Earpiece extracted

Figure 3.12: Photos showing the generic earpiece both inserted and removed from the ear

housing, as it is not required to hold the earpiece in position thus enhancing the cosmesis associated with the TMEP concept. The potential drawbacks associated with the generic earpiece include change in position and orientation of the microphone during removal and reinsertion of the earpiece, reduced external noise attenuation and a less enhanced low frequency response. A summary of the main differences between the earpiece types is given in Table 3.2.

	Custom	**Generic**
Benefits	Consistent localisation, better noise attenuation, enhanced LF response	Cheap, easier to manufacture, same earpiece for everyone, cosmetically discrete
Drawbacks	Subject-specific, expensive, harder to manufacture, lead time, reduced cosmesis	Potential for inconsistent localisation, reduced noise attenuation

Table 3.2: Comparison of the custom and generic earpiece

3.4.3 Bioacoustic processing

A general bioacoustic processing framework is proposed for extracting and decoding user intention, expressed though impulsive acoustic sounds. It takes as input, the digital bioacoustic data-stream(s) and correlates this to a set of classification outputs (classes) allowing a discrete command to be sent to the target system. An overview of the processing flow is given in Fig. 3.13. There are a variety of sub-processes that are needed and are represented by a generalised set of sequential subfunctions within a high-level system flow architecture. Mapping of the data-stream(s) consists of two main stages, namely,

bioacoustic extraction and bioacoustic decoding. The extraction stage involves activity detection and segmentation of the signals while the decoding stage decides whether an intentional action was performed and if so from which class it belongs.

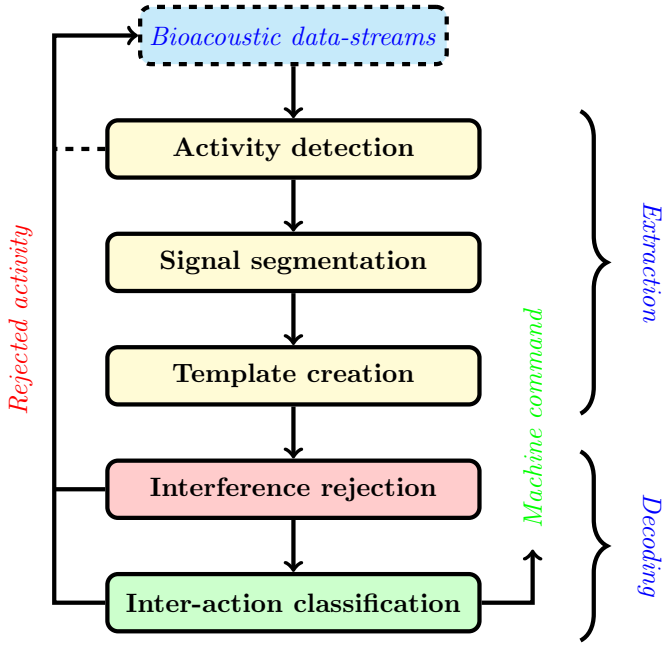


Figure 3.13: Generalised system flowchart showing an overview of the processing needed to translate the incoming TMEP data-stream to a useful machine command

Within the bioacoustic extraction, the activity detection infers from the continuous stream of incoming data whether an event associated with a controlled volitional signal has occurred. To ensure a self-paced HMI system, the monitoring of the signals needs to be executed continually and therefore any computation required at this stage needs to be implemented efficiently. This is especially important when dealing with bioacoustic signals, as high sampling frequencies are generally required. Due to the subject-specific and nonstationary nature of the signals this detection block will generally need to be based on tunable parameters or thresholds. These parameters need to set (or adjusted) in such a way so as to minimise type 1 and type 2 errors. At this stage of the signal processing a type 1 error (or false positive) is defined as events detected when non-controlled or unintended bioacoustic activity occurs. A type 2 error (or false negative) is defined as undetected activity when controlled bioacoustic signals were intended. Any type 1 errors which do occur, can be dealt with during the decoding stage. Once activity has been detected and a bioacoustic event logged, the resulting signal needs to be segmented so that realisable and tractable computation can be performed. This requires that the signal be truncated to a bounded length, with this length dictated by its ability to capture all of the associated signal information. It is obvious that the smaller the length of this segment the better, as

this will minimise computational overheads. However, if too short a segment is defined then valuable information can be discarded. To exemplify the morphology associated with each type of volitional action a template creation process can be used. This can be considered a pattern dependent extraction process and creates representative class-specific waveforms from a set of training signals. These can be used to align new patterns prior to decoding and as model parameters within the inter-action classification block itself.

The bioacoustic decoding stage comprises two main decisions blocks, the first decision block endeavours to dichotomously infer whether a controlled or non-controlled signal has been captured. This block will thus reject any TMEP events that it concludes were not indicative of an intended movement. Once this decision has been established, if it is concluded that a volitional signal was initiated then the second decision process can occur. This final decision process deals with the polychotomous question of which of a predefined subset of actions was issued (inter-action classification) and can be considered a multi-class classification problem. Important steps include, but are not limited to, extraction of appropriate information to efficiently and accurately perform classification with (feature extraction and selection) and mapping of these features to one of M classes (pattern classification). Prior to inter-action classification, segment normalisation is implemented to ensure that no single class dominates the decision process. This ensures that the bioacoustic signals have a standard segment morphology prior to extraction of discriminatory features.

3.4.4 BIOacoustic HMI Software

During the course of this research, software including a GUI frontend was developed to support the capture and implementation of volitional bioacoustic signals especially for the purposes of HMI. The software was developed in MATLAB v7.12 with the frontend shown in Fig. 3.14. The software serves dual functionality, firstly, it allows the capture and storage of bioacoustic signals for further off-line post-processing and secondly, it provides a real-time implementation and testing framework for bioacoustic HMI processing concepts. Although MATLAB is an interpreted language and therefore not considered the most efficient in terms of real-time processing, it is high-level and as such, allows for rapid prototyping and implementation of various bioacoustic processing concepts. The software was written using the object-orientated programming (OOP) capabilities of MATLAB. Encapsulating the program within a set of objects provides a natural and efficient framework for developing complicated software. Furthermore, for asynchronous or self-paced HMI systems, the timing of a volitional action is unknown prior to its occurrence. Event listening provides a useful method for indicating when volitional activity has occurred based on continual and low computational monitoring by the detection subsystem. It can subsequently flag when interesting signal behaviour has occurred and send an event indicating

that the more computationally intensive post-detection subsystems can execute (see Fig. 3.13). The OOP framework provides a neat method for implementing event-based programming, in which objects can trigger events in response to a asynchronous change in state.

3.5 Off-line data sets

3.5.1 Participants

In total fifteen healthy subjects (11 male, 4 female), ranging in age from 20 - 54 years, participated in the off-line data collection procedure. The data-sets were collected under various experimental conditions and are summarised in Table 3.3. As can be seen, they have been grouped according to three experimental conditions namely, earpiece type, whether the participant is trained or untrained and the number of channels recorded. The final column shows the number of subjects associated with each group. The data from Group A was collected in a previous study [24]. Although more data would be desirable, due to the length of time to train the subjects, it was infeasible to collect more subject data within the scope of this thesis. All experiments were approved by the the local research ethics committee with the participants giving their informed consent before taking part in the study.

Group	Earpiece type	Training	Nr. of channels	Nr. of subjects
A	Custom	well trained	1	7
B	Generic	well trained	2	4
C	Generic	minimally trained	2	4

Table 3.3: Overview of all the participants utilised during off-line experiments

For the purposes of one channel testing, all fifteen participants were used, as they all involve at least one earpiece channel. The most effective channel was selected from the subjects who had data recorded from dual earpieces.

3.5.2 Experimental conditions

As stated previously there are three experimental conditions that were varied. The earpiece type can be of either custom or generic build with the reader directed to Section 3.4.2 for a full description of both. The number of channels can be (1) or (2) indicating whether the data was recorded from a single earpiece channel (placement in either ear) or dual earpiece channels (one in each ear). The final experimental condition (training) is used to validate the learning effect of the user and is described below.

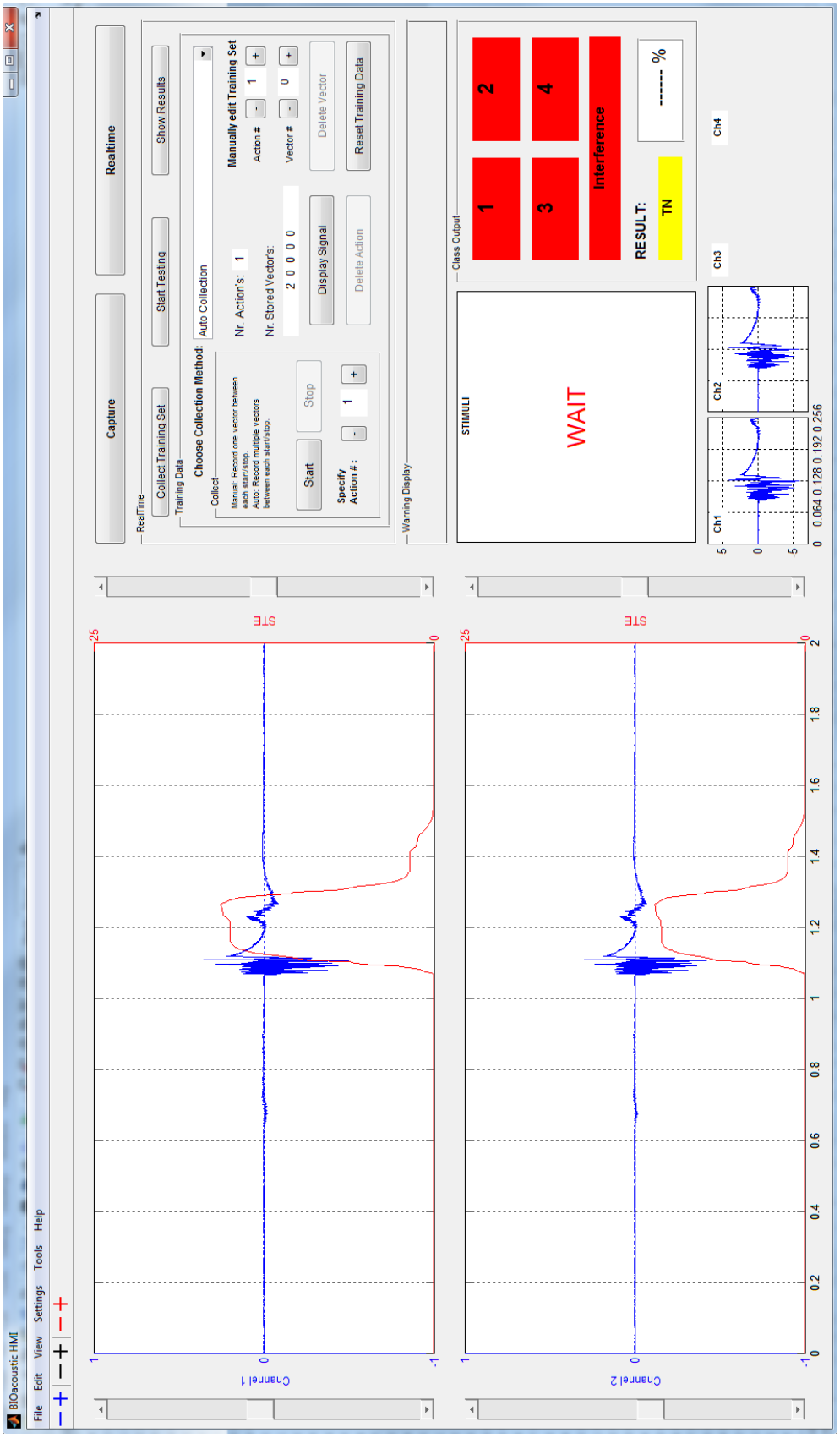


Figure 3.14: Software frontend developed for capturing and realtime decoding of volitional bioacoustic signals

The well trained individuals were instructed over the course of 4-6 sessions, with the gap between sessions consisting of at least two days to allow the participants time to practise making the tongue actions. This is considered important as although the actions themselves feel natural when executing, they are not normally occurring and therefore individuals generally have no prior experience of them. This is especially true of the left and right tongue actions, as during normal activity, for instance talking, eating and drinking, the tongue general only moves in the mid-sagittal plane. These two actions require additional placement of the tongue on the transverse plane prior to performing the movement, with the tip of the tongue moving relative to both the sagittal and coronal planes during execution of the actions. Each session lasted for less than thirty minutes and involved visual feedback from a real-time display of the incoming waveform during the course of the session. This allowed the user to modulate and optimise the manner in which they performed the tongue action under guidance from an instructor, who established the validity of each action in terms of morphology, SNR, inter-action separability and intra-action variance. It should be noted that the time it took individuals to completely grasp the execution of the actions varied subject-to-subject.

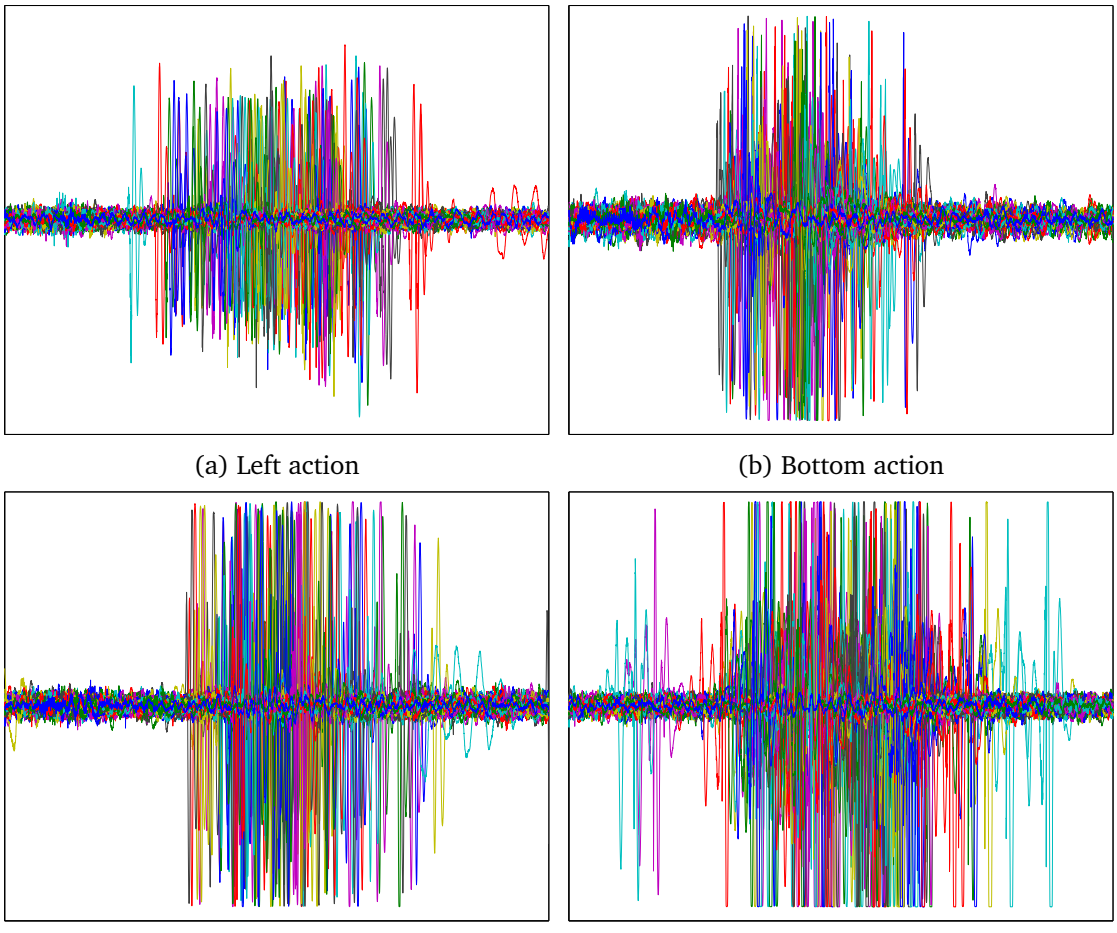
For both the well trained and minimally trained participants, the first session acted as an introduction to the tongue actions as the subjects had no prior knowledge of the TMEP concept. For the minimally trained participants, only one more training session was performed after no longer than a week. The data collection process was then performed after this second session. The second training session was deemed necessary to ensure that the SNR of the signals was high enough to ensure that they could be detected and segmented with some reliability.

3.5.3 Experiment and signal recording

For the single-channel data sets captured using custom earpieces (Group A) the ear was selected based on which was felt to have the better response in terms of amplitude and frequency. Prior to testing, the custom earpiece was designed and developed for each of the subjects using the method described previously in Section 3.4.2. At the time the impression moulding was taken and during the data recording sessions, the subjects were known to have no infections or issues associated with the inner ear, eustachian tube and/or oral cavity. All other subjects utilised a generic earpiece during bi-aural recording.

Data collection was carried out in a standard office environment with the subject sitting comfortably in front of the computer. The experimental procedure consisted of collection of at least one hundred of each action performed in turn. The recording software continually monitored the channel inputs and retained, at least, the last two seconds of data in a constantly updating buffer. When an action had been performed, indicated by the user, the data in the buffer was stored and contained a single TMEP instance associated with

a specific tongue action. This meant that no online detection phase was required during off-line recording. On occasion, the data could be scrapped if the subject felt (1) they had incorrectly performed the action, (2) the recorded segment contained activity other then that associated with just controlled TMEP signals and/or (3) external interference had corrupted the signal sufficiently so as to mask its specific morphology. It should be noted that it was a rare occurrence for a signal to be scrapped by the individual but was a useful mechanism to prevent anomalous recordings and outliers from being included within a data set. When the entire data set associated with an action had been collected it could be saved and labelled with its associated ground truth as indicated by the GUI. Once one hundred instances of each of the actions had been recorded, the collection procedure terminated. Figure 3.15 shows superimposed representations of the TMEP actions associated with a single subject from Group A (well trained, custom earpiece).



(a) Left action

(b) Bottom action

(c) Right action

(d) Top action

Figure 3.15: Raw TMEP signals for tongue movement actions (left, bottom, right, and Top) for 1 subject (amplitude vs. time). The duration of each signal is 2 seconds with each action repeated one hundred repetitions times at a sampling frequency of 2 kHz

3.6 Chapter summary

This chapter has introduced and described background information relating to bioacoustic TMEP signals, originally proposed by Vaidyanathan et al. [24]. This has been in terms of their generation through prescribed volitional tongue movements, sensory detection within the ear canal and general bioacoustic characteristics. These characteristics indicate both temporal and frequency based differences between the four defined tongue actions. The associated hardware has been introduced and consists of both custom and generic earpieces, soundcard and a processor running custom built MATLAB software. The software centres on a GUI interface and enables both bioacoustic data recording and real-time implementation of the proposed bioacoustic processing architecture, with this architecture the main focus of subsequent chapters. Throughout this work, the proposed processing strategies have been validated using off-line TMEP data. Therefore, these data sets in terms of the participants involved, experimental conditions and signal recordings procedure, have also been described.

Chapter 4

Bioacoustic signal extraction

In order to decode user intention from continuous stream(s) of bioacoustic data, areas of intentional activity need to be first highlighted and extracted. Bioacoustic signal extraction performs three key tasks relating to this, namely, volitional activity detection, signal segmentation and exemplification of the generating actions' signatures. The first two tasks can be considered pattern independent and deal with the creation of finite length segments containing *interesting* bioacoustic activity. *Interesting* implies that there is a high chance of volitional activity content but requires further processing to establish its validity. As activity detection is a common process in many diverse signal analysis domains, a short review of general activity detection methods is initially presented. The activity detection subsystem is required to constantly monitor the incoming signal in real-time and has been implemented ensuring efficient computation. The exemplification process can be considered pattern dependent and involves creation of class-specific templates which are representative of the unique signatures associated with each generating action. This is achieved using a new robust template creation process based on randomisation of a partial pairwise cross-correlation averaging tree (PPCCAT). The ability of this method to create stable templates under different data samples is evaluated in terms of the jackknife standard error and absolute bias. Additional processing methods introduced include signal standardisation, ensuring that no single class dominates during the decoding phase.

4.1 Activity detection

4.1.1 Background

Activity detection is an automated technique for detecting the presence, and also absence, of a defined set of activities within a signal. This absence can generally be considered the default state in most detection applications. Most real-time systems require a dedicated activity detection subsystem, for example, voice or speech processing, detection of epilep-

tic or movement-related states in neural signals and monitoring the QRS complex, to name a few [54, 150, 151]. In general, detection is achieved by utilising characteristics of the activity, that differentiate it not only from the background noise but also from additional interfering activities. Transducers are prone to encounter this type of activity, especially in the context of bioacoustic signals. Background noise can be considered a constant random variable sampled from a stationary ergodic process and generally normally distributed, for instance white noise [152]. In contrast, interfering activities although still highly stochastic, can be considered extremely non-stationary, sporadic, of higher energy, and to some extent chaotic. Therefore an optimal detection system will be able to correctly identify the volitional activity from the background noise (true positives) whilst minimising the incorrect identification of interfering signals (false positives).

Due to the similarity in signal and sensors, detection techniques used in speech processing are a useful avenue to explore. The short-term energy (STE) and zero-crossing rate (ZCR) are the most common techniques for detecting the endpoint of utterances and phonemes associated with voiced speech [150, 153, 154]. The STE gives an estimate of the energy in the signal while the ZCR is defined as the number of times the speech signal changes sign within a given time window. As such, the ZCR gives an estimate of the fundamental frequency associated with the signal. The popularity of these methods is attributed to the simplicity of the computation required to compute these measures while the energy contour is an effective detection feature in moderate SNR settings. In real-time applications, the simplicity of the computation during the detection procedure is important, as this is a continual operation which occurs whenever new data becomes available. This metric reduces the risk of computational bottlenecks occurring at this early stage in the processing framework. An additional benefit of these techniques is that there is only one free parameter that needs to be selected, namely, the frame size in which the measures are summated over. Alternative measures that have been used by ASR detection subsystems include the root mean squared energy (RMSE), sub-band energies, spectral-divergence between speech and a background noise model, changes in pitch/frequency and higher order statistics [153, 155, 156]. It should be noted that the majority of these methods have increased computational loads when compared to the STE/ZCR technique.

Detection of the QRS complex within electrocardiogram (ECG) recordings represents another biosignal that provides detection features similar to volitional bioacoustic activity. This is due to similar morphology, time scale and spectral content (approximately 10-25 Hz). Having said this, ECG signals are periodic with many techniques utilising this property. For volitional signals the *interesting* activity is asynchronous. Furthermore, detection of QRS complexes is a vast and independent research field in itself and therefore only a brief overview of relevant methods will be discussed. A comprehensive overview of the main algorithmic approaches used in QRS detection is given in Kohler et al. and for com-

pleteness, are summarised in the remainder of this paragraph [54]. The general structure of the QRS detection subsystem consists of a preprocessing step incorporating a linear and/or nonlinear filtering phase, followed by a decision step comprising a peak detection logic subblock and a final decision subblock. A common approach is to use derivative-based operations which are an effective detection feature due to the sharp transitions at the R-wave. The use of digital filter methods are also used extensively and are designed in such a way as to enhance the QRS waveform while concurrently suppressing noise. This includes various combinations of low pass, high pass, linear and nonlinear filters combined using difference, multiplication and logical operations. At this stage, it is useful to point out that the QRS complex has a fixed morphology that is consistent across subjects and therefore the filters and general algorithmic implementations can be designed in such a way as to enhance this particular form. Another approach utilises a wavelet transform whereby R-peaks are found by scanning for simultaneous modulus maxima in the relevant scales of the transform. Similar methods based on filter banks have also been proposed. In general, all the methods require some kind of threshold to be set, allowing a binary decision to be made based on the detection feature contour. These can be of fixed or adaptive nature but, in general, are adapted due to changing background noise conditions and similarity between QRS complexes beat-to-beat. Other complementary approaches include adaptive filtering, matched filters, Hilbert transforms and classification approaches based on artificial neural networks (ANN), HMM and Bayesian formalisations. Almost all the QRS algorithms use additional decision rules for the reduction of false-positive detections.

An adaptive Wiener filter can be used for detecting a change in terms of the wide-sense stationarity and thus prediction error associated with a signal [151]. The signal can be modelled as an auto-regressive (AR) process which assumes the signal to be both stochastic and stationary. This approach can be set up using two different problem definitions which yield the same results, namely linear finite impulse response (FIR) prediction and AR modelling [152]. In linear FIR filtering the aim is to determine an FIR filter that can optimally predict future samples of an AR process based on a linear combination of past samples. As mentioned previously, this has been applied to the QRS complex detection problem in which a sudden change in the signal, that is when a QRS complex occurs, leads to an increase in the prediction error [54]. In AR modelling, the intention is to determine an all-pole infinite impulse response (IIR) filter, that when excited with white noise produces a signal with the same characteristics as the AR process. This idea has also been applied in the neural domain, whereby a change from quasi-stationary baseline activity to highly non-stationary epileptic activity, in LFP recordings from rats, leads the AR model to begin inaccurately predicting the system behaviour at this transition [151]. Regardless of the school of thought, the AR coefficients are adaptively adjusted as new data enters the subsystem. When a sudden and/or non-stationary change in the signal

dynamics occurs, the associated prediction error suddenly increases. By using this error signal or its envelope as the detection contour, a detection can be made whenever this contour exceeds a fixed or adaptive threshold. An alternative methodology developed for epileptic activity detection utilises a measure similar to the STE but rather than taking the average or sum of the square of the signal it uses the median of the square of the signal [157]. This is considered a much more robust statistic with regard to outliers, with superimposed spiking action potentials an example of this type of interference in LFP activity. The signal is normalised by a slowly changing background sequence allowing it to adapt to gradual and inevitable changes in the ambient neural state.

The characteristics of impulsive bioacoustic activity, such as TMEP signals, which could enable their perception include, energy and frequency content, nonstationarity, statistical distribution and temporal duration. These could be extracted using various implementations or combinations of the previously mentioned methods. Outside of the assistive domain, similar volitional signals have been highlighted based on finger taps and cantilevered piezoelectric sensors [17]. Individual finger taps were detected using the absolute exponential average across ten sensors based on fixed thresholds found experimentally. The absolute exponential average is a measure of the average energy content across the multiple channels. The use of fixed thresholds indicated that the background noise was assumed stationary and there was adequate SNR for detection. This is often assumed to be the case when detecting volitional activity in acoustic signals, such as speech [150]. Furthermore, if the SNR is high enough, then regardless of whether the ambient noise is stationary or non-stationary, a high fixed threshold will ensure that false positives are inhibited. Based on these observations and due to the similar acoustic nature, the STE detection contour is selected for extraction of volitional bioacoustic activity. Furthermore, when compared to the AR error method during analysis of TMEP signals, it was found to have a similar shape whilst requiring a much smaller computational footprint. Additionally, in opposition to neural activity, there is unlikely to be outliers in the signal and therefore using robust statistics would add unnecessary computation due to the required sorting algorithm. Fixed thresholds in both the energy and time directions are utilised and based on the characteristics of the proposed volitional activity. This will ensure that the detection of intentional activity is maximised whilst rejecting potential interference, such as speech, due to different energy content and temporal duration.

4.1.2 Short-term energy contour

The STE contour of a signal x at time instant n is defined as

$$E_n = \sum_{k=n-W_E+1}^{n} x_k^2 \tag{4.1}$$

Where E_n is the STE at time-step n, W_E is an arbitrary finite length window that the STE is calculated in and $x_n \in \mathbf{x}$ is the current signal sample.

4.1.3 Thresholds

Defining fixed thresholds in both the time and energy directions implies that a total of four thresholds can be stipulated relating to lower and upper thresholds in both directions. These are described in Table 4.1. The ***lower energy threshold*** can be considered the most important and indicates that for a certain level of background noise, no bioacoustic events will be detected. The duration thresholds are compared to the distance between the points where the STE contour exceeds (p) and then recedes (q) the ***lower energy threshold***. Without the ***upper duration threshold*** the chance of false positives is extremely high due to longer and higher energy signals, such as coughing or speech, occurring. The ***lower duration threshold*** is less important as the volitional activity is already impulsive in nature and if all impulsive signals require detection then D_L can be set to zero.

Condition	Reason
$[E_p, E_{p+1}, \ldots, E_q] > T_L$	Fixed ***lower energy threshold*** inhibits detection of ambient noise and low energy signals
$[E_p, E_{p+1}, \ldots, E_q] < T_U$	Fixed ***upper energy threshold*** inhibits detection of high energy signals, for example some speech signals
$q - p > D_L$	Fixed ***lower duration threshold*** inhibits detection of very impulsive signals.
$q - p < D_U$	Fixed ***upper duration threshold*** inhibits detection of longer signals, for example motion artifacts

Table 4.1: Description of fixed thresholds associated with volitional bioacoustic activity detection

Figure 4.1 highlights the proposed activity detection process using a single channel TMEP signal. The TMEP signal exhibits activity associated with two events, namely, a volitional tongue action (event **A**) and a speech signal (event **B**). By setting the four thresholds appropriately, it is possible to only detect the first event, effectively rejecting the speech signal. For example, by setting $T_L = 5, T_U = 40, D_L = 0.1$ and $D_U = 0.6$ as highlighted by the horizontal and vertical lines in the figure, event **A** is detected while event **B** goes undetected. Note that in this example, if $T_U = 30$, then regardless of what D_L and D_U are set at, event **B** would remain undetected.

For multichannel detection, separate thresholds can be set for each channel and an event detected if (1) a single channel flags that an event occurred, (2) the majority of channels flag that an event occurred or (3) all channels flag that an event occurred. Across these multichannel schemes, the first method would have the highest true positive rate

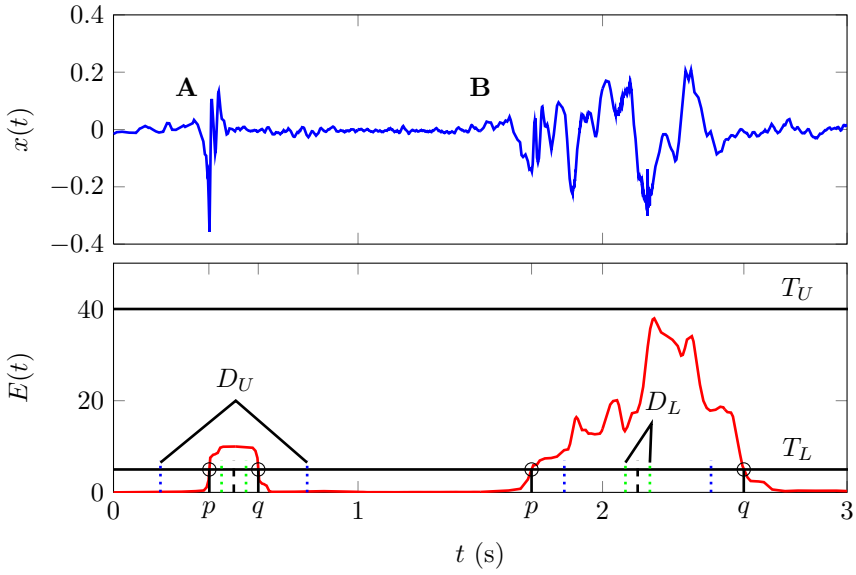


Figure 4.1: Highlighting the detection process: Top plot is an example single channel of TMEP data (blue line) while the bottom plot is the associated short-term energy contour (red line). The event marked A is representative of a tongue action while event B is a typical interference signal. Also shown are example detection parameters including $T_L = 5$ and $T_U = 40$ indicated by horizontal black lines, the exceed and recede points (p and q), $D_L = 0.1$ indicated by dotted green vertical lines and $D_U = 0.6$ indicated by dotted blue vertical lines. The black dashed vertical lines show the midpoint between the exceed and recede points for each event. Based on these detection parameters, event A would be flagged as a TMEP action but event B would not due to $q - p > D_U$

but also the highest false positive rate while conversely the final method would have the lowest true positive rate but also the lowest false positive rate. This is indicative of the general trade-off that occurs in detection systems through 'tightening' or 'loosening' of the threshold parameters.

4.2 Signal segmentation

The following segmentation process is utilised and has been adapted from Vaidyanathan et al. [24]. The signal is initially segmented using the position t, associated with the maximum STE E_t. This is located in the range $[p, q]$ where E_n exceeds and then recedes the minimum STE threshold T_L, while still fulfilling the remaining detection criteria. Thus the initial endpoint is given by

$$t = \arg\max_n(E_p, E_{p+1}, ..., E_q\} \tag{4.2}$$

with the initial start point given by $r = t - N + 1$, where N is the length of the segment to be retained. For independent channels, each signal is shifted so that the midpoint a, of the 1D discrete distribution f, is located at the centre of the segment window using $F = 1/2$. F can be considered the associated cumulative distribution over the window and is given by

$$F_k = \sum_{j=1}^{k} f_j \tag{4.3}$$

where

$$f_j = \frac{x_j^2}{\sum_{i=r}^{t} x_i^2} \tag{4.4}$$

The final end-point is then given as $u = a+N/2$ with the final start-point at $v = a-N/2+1$. This ensures that there is minimal temporal shifting between the various signal instances associated with a particular class. For multichannel systems, if the channels are shifted concurrently, then t and a are taken to be the average position across all channels.

4.3 Real-time considerations

4.3.1 Ring buffer

Ring (or circular) buffers allow for continual monitoring of a real-time input stream without putting unnecessary strain on the memory resources of the processor. It allows for a finite moving window of historical data to be stored, which is necessary for most signal processing applications. For example, real-time filtering requires convolution of the

incoming data stream with a number of filter taps, N_{taps}, implying that at least N_{taps} samples of historical data are needed to compute the causal output related to this filtering operation. Conceptually, as shown in Fig. 4.2 (left side), a ring buffer can be considered a circle containing a fixed number of samples with neither a start or end point. Data is input incrementally into the empty ring, eventually overlapping itself and thus overwriting the data that was stored a fixed amount of time earlier. This fixed amount of time is equal to the length of the buffer L_b. In reality, this circular configuration is unable to be realised on a processor, as memory is never physically created as a ring. Therefore, a linear equivalent implementation, also shown in Fig. 4.2 (right side), is what is actually implemented in practise.

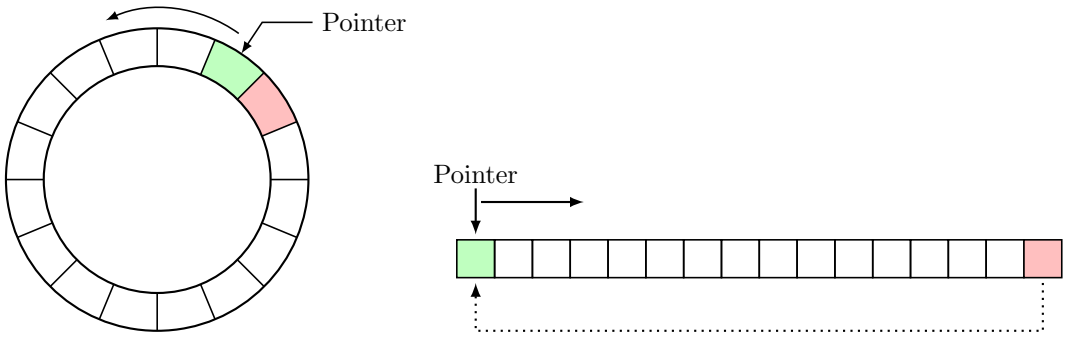


Figure 4.2: Representation of the ring buffer both conceptually in its circular configuration (top) and linearly as it is realised (bottom)

A pointer is utilised allowing the position of the current data sample to be tracked within the circular buffer. This ultimately means that when a new piece of data is inserted, no shifting of the previous data is required. Initially the buffer length is arbitrarily defined and filled with zeros, with the pointer initially set to position 0 in the buffer. When a new piece of raw data arrives, it can be entered into the buffer at this location and the pointer shifted to the next memory slot, that is, position 1. It should be noted that a single piece of data, although ideal, is not necessarily a single sample and depending upon communication latencies between the ADC and processor may contain multiple samples within the data packet. As stated earlier, when the buffer is fully populated the pointer is returned to position 0 and the data begins to overwrite itself. This is achieved using the following clock arithmetic

$$p_k = (p_{k-1} + 1) \mod L_b \tag{4.5}$$

where p_k is the current pointer position, p_{k-1} is the old pointer position and L_b is the fixed length of the buffer. It is this calculation that allows the linear realisation of the circular buffer to act like a continuous ring. The activity detection and signal segmentation processes both require causal data processing and therefore this type of buffer, as with most real-time systems, is necessary for historical signal storage.

4.3.2 Parsimonious computation of short-term energy

The detection contour, based on the STE measure, requires constant updating if the bioacoustic signals are to be continually monitored. Therefore, it is important that it is computed in an efficient and parsimonious manner. To this end, a novel STE update law has been developed. Both the detection and segmentation processes require historical STE values for placement of the accepted signals within the finite length window N. This implies that a secondary buffer needs to be implemented which stores the newly computed STE value whenever new data becomes available to the processor. This buffer is implemented in a similar fashion to storage of the raw data packets, utilising the circular buffer configuration as described in the preceding section. During a communication cycle between the ADC and processor, a new data packet $\mathbf{x}_n = [x_{n-C+1}, x_{n-C+2}, \ldots, x_n]$ will become available and contains multiple contiguous samples, that is $\mathbf{x}_n \in \mathbb{R}^C$ with $C \geq 1$. It makes sense to calculate the STE value at the same frequency as the communication cycles, thus avoiding extra causal computations between the cycles. This would have the effect of slowing the communication process and causing the signals to lag. The calculation of the STE value at consecutive communication cycles is shown conceptually in Fig. 4.3.

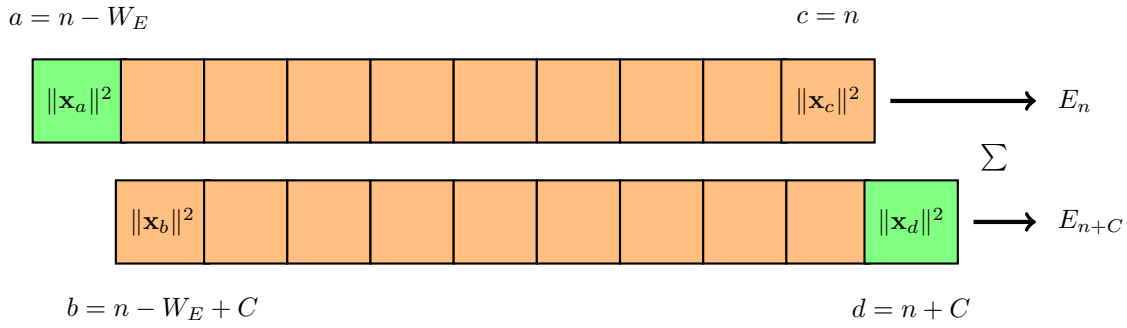


Figure 4.3: Conceptual representation of the STE calculation at time n and at the next communication cycle $n + C$ showing the overlap of information between these two computations (orange blocks). Each block represents a data packet $\mathbf{x}_n = [x_{n-C+1}, x_{n-C+2}, \ldots, x_n]$

As highlighted by the orange blocks in Fig. 4.3, there is considerable overlap in terms of the summated data between consecutive STE computations. Rather than fully recalculating the STE at each communication cycle, an efficient update law can be defined and is given by

$$E_{n+C} = E_n + \|\mathbf{x}_{n+C}\|^2 - \|\mathbf{x}_{n-W_E}\|^2 \tag{4.6}$$

where $\|\mathbf{x}_n\|$ represents the vector norm of a C-dimensional data packet $\mathbf{x}_n \in \mathbb{R}^C$ at communication cycle n. For even quicker implementation, depending on the size of the data packets, a third circular buffer can be implemented which contains the current and historical inner products of each packet. Ignoring read and write operations, this would reduce

the STE computation from $2W_E - 1$ to only 2 mathematical operations. For example, during TMEP capture, $W_E = 400$ samples, implying a reduction by a factor of 400 operations. It should be noted that between cycles the STE contour follows a zero-order hold convention.

4.4 Class-specific signatures

To exemplify a data set associated with a particular volitional action (or class), a robust template creation process has been formulated and evaluated. It is based on a pairwise alignment and averaging process and can be considered an extension to the template creation process proposed by Vaidyanathan et al. [24]. This original template creation process was found to be unstable to different permutations of the data. Therefore, it did not give a robust estimate of the class-specific waveforms. Instead, an extension to this original process has been formulated and is based on a two stage partial tree implementation using random permutations of the data. The various processes required are outlined both mathematically and visually with the stability, associated with the templates, estimated using the jackknife standard error and absolute bias. Evaluation is performed using TMEP data with comparison to the original template creation method.

4.4.1 Cross-correlation alignment

To find the position of best alignment between two vectors, for example a template and incoming test signal, cross-correlation can be used. Alignment is achieved by shifting to the position of maximum cross-correlation, as this value indicates when there is constructive superposition between the two signals. The cross-correlation between two discrete and real vectors $\mathbf{x}, \mathbf{h} \in \mathbb{R}^N$ is given by

$$(x \star h)_n = \sum_{m=-N}^{N} x_m h_{n+m} \tag{4.7}$$

where $\star$ represents the cross-correlation operator and both signals are zero-padded outside of their discrete domains. The position of best alignment $p \in [-N, \ldots, 0, \ldots, N]$ and thus the shift distance is given by the position of maximum cross-correlation using

$$p = \arg\max_n \left((x \star h)_n\right\} \tag{4.8}$$

This implies that vector $\mathbf{x}$ can be shifted onto $\mathbf{h}$ as follows

$$\hat{\mathbf{x}} = [x_p, x_{p+1}, \ldots, x_{N+p}]^T \tag{4.9}$$

If the samples $n < 1$ and $n > N$ are unavailable then alternatively a circularly shifting operation can be performed

$$\hat{\mathbf{x}}_{\mathbf{c}} = \begin{cases} [x_p, x_{p+1}, \ldots, x_N, x_1, x_2, \ldots, x_{p-1}]^T, & \text{if } p \geq 1 \\ \mathbf{x}, & \text{if } p = 0 \\ [x_{N+p}, x_{N+p+1}, \ldots, x_N, x_1, x_2, \ldots, x_{N+p-1}], & \text{if } p \leq -1 \end{cases} \tag{4.10}$$

Due to the causal nature of real-time processing, the following shift can also be defined

$$\hat{\mathbf{x}}_{\mathbf{r}} = \begin{cases} \hat{\mathbf{x}}, & \text{if } p \leq -1 \\ \hat{\mathbf{x}}_{\mathbf{c}}, & \text{otherwise} \end{cases} \tag{4.11}$$

This is due to only historical samples being available during online processing.

4.4.2 Template creation

To obtain a reliable estimate of the representative morphology, equivalent to an exemplar for a given set of signal patterns, a template creation process can be utilised [24]. It is based on a pairwise cross-correlation averaging procedure and allows for better preservation of the morphology of a specific pattern compared to simply taking the aligned global mean. If it is assumed that the number of patterns to be averaged is an integer power of two, then each pattern can be paired off, with each pair aligned and their average taken. The alignment is achieved using the cross-correlation alignment described previously in Section 4.4.1. This process can be repeated using the outcome of the previous pairing until a single template is left. The whole procedure can be represented as a binary tree decomposition as shown in Fig. 4.4 for a data set $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{N_p}], \mathbf{x}_i \in \mathbb{R}^N$ and $N_p = 8$. Each level of the tree implies an alignment and averaging process between two adjoining vectors towards the root of the tree, with the outcome of this process forming the subsequent node. At each stage, the number of nodes is halved, with the number of levels $L = \log_2(N_p)$. By decomposing the alignment stage in this pairwise fashion, it allows for a re-adjustment at each level of the alignment process due to the change in morphology of the signal as a result of the averaging process. This results in local alignments at each level of the binary tree, which would not occur if a single global alignment and averaging procedure was implemented.

4.4.3 Partial-tree implementation

The previous template creation process only works when the number of patterns is an integer power of two. If this constraint is relaxed, allowing for an arbitrary N_p, the same principle can still be implemented based on cropping of the original tree structure. Pairs

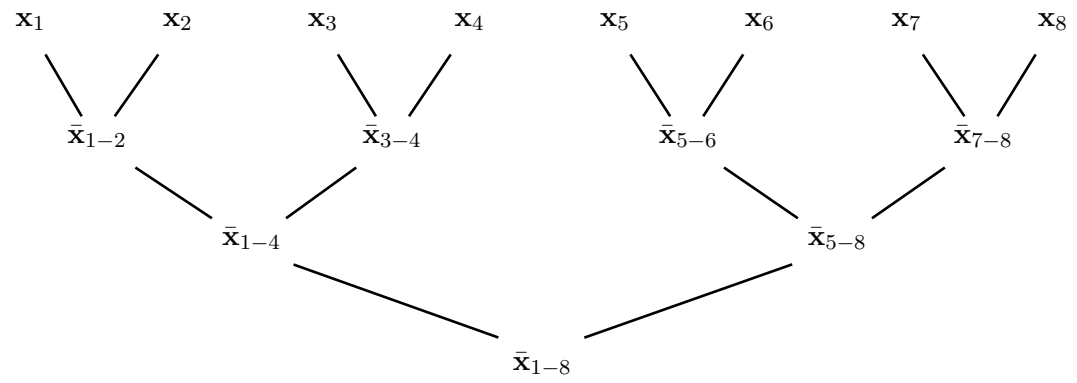


Figure 4.4: Full tree decomposition representation of the template creation process for $N_p = 8$ using the pairwise cross-correlation averaging procedure

of the extreme nodes can be removed from the top layer of the binary tree, thus reducing the number of input nodes by creating a partial pairwise cross-correlation averaging tree (PPCCAT). Figure 4.5 shows this for different size data sets including $N_p = 5, 6$ and 7. The number of levels of the initial binary tree now becomes $L = \lceil \log_2(N_p) \rceil$, with the number of patterns in the first ($N_p^{\{1\}}$) and second ($N_p^{\{2\}}$) layers calculated from

$$N_p^{\{1\}} = N_p - N_p^{\{2\}} = 2N_p - 2^L \tag{4.12}$$

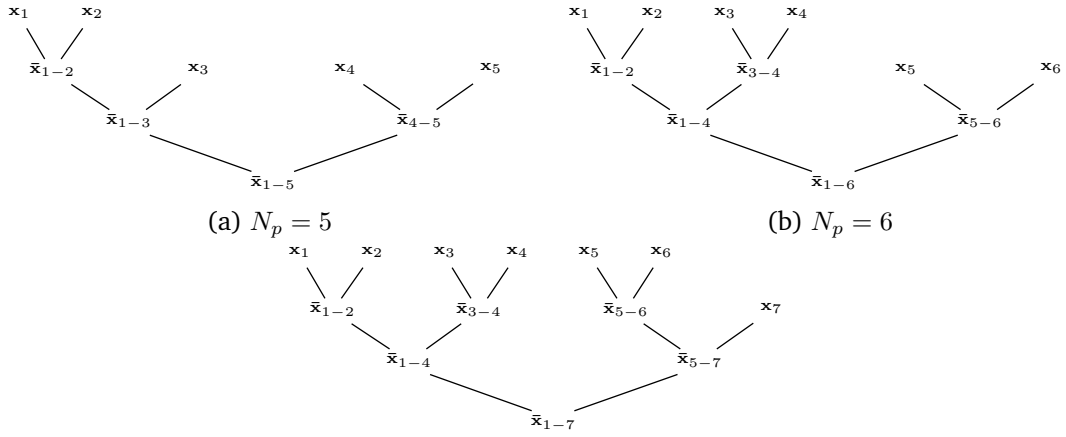
Inputting patterns directly at this second layer creates a bias of the template toward these instances. To remove this bias, the training set can be initially weighted using the following coefficients

$$w = [w_1, w_2] = \frac{1}{N_p}[2^L, 2^{L-1}] \tag{4.13}$$

where the weight indices show association to either the first or second layer, depending upon which layer a specific pattern is input. These weights are required, as the pairwise cross-correlation averaging process can be considered equivalent to finding the expected central tendency of the aligned patterns. The weights are calculated from the following equality

$$\bar{\mathbf{x}}_{1:N_p} = \frac{1}{N_p} \sum_{i=1}^{N_p} \hat{\mathbf{x}}_i = w_1 P_1 + w_2 P_2 \tag{4.14}$$

where $P_1 = \frac{1}{2^L} \sum_{i=1}^{N_p^{(1)}} \hat{\mathbf{x}}_i$ is the pairwise averaging process for the input patterns at the first layer and $P_2 = \frac{1}{2^{L-1}} \sum_{i=N_p^{(1)}+1}^{N_p} \hat{\mathbf{x}}_i$ is associated with the second layer. Inputting these into Equation 4.14 yields the weight ratios outlined in Equation 4.13. Cross-correlation has been ignored in this analysis as it can be considered independent of the expectation operation.



(a) $N_p = 5$ (b) $N_p = 6$

(c) $N_p = 7$

Figure 4.5: Partial tree decomposition representation of the template creation process for (a) $N_p = 5$, (b) $N_p = 6$ and (c) $N_p = 7$ using the PPCCAT implementation

4.4.4 Robust template creation

Encoding the alignment and averaging procedure within this partial tree structure implies that the template creation process is a nonlinear combination of the input patterns. The order that the patterns are input into the tree can therefore have a significant effect on the template that is output. Figure 4.6 (second row) highlights this by showing 32 repetitions of the templates associated with four TMEP patterns from one subject. The raw input patterns are shown in Fig. 4.6 (top row), with each repetition computed using a random permutation of the input data set. As the morphology is the important factor, these templates are further shown aligned to one another in Fig. 4.6 (third row). This implies that the morphology of the generated templates are not only sensitive to the content of the data set but also the initial ordering with this indicative of a stability issue. This instability could be reduced using the following concepts

- Increase the sample size associated with specific patterns. This would mean collecting a larger data set which can generally be prohibitive.
- Remove the initial randomisation process. This would make the template creation process deterministic but is highly unlikely to provide the optimal representation of a template given a sample pattern set.
- Implement a resampling procedure. By random resampling of the input pattern order, it should reduce the unstable effect, creating a more robust template. This would be at the cost of higher computational complexity.

The resampling procedure is of most use, as it is independent of the sample size and the initial training set order. Increasing the computational complexity is generally not

prohibitive during the training stage. The resampling process proposed, utilises a two phase template creation process. In the first phase, B random permutations of the input pattern set are created according to

$$\mathbf{q}_b = \text{randperm}([1, 2, ..., N_p]\}, b = 1, 2...B \tag{4.15}$$

and are used to define the order in which the training vectors are input into the PPCCAT. Each reordering produces a separate template, resulting in B templates. The second phase involves inputting these templates into a secondary PPCCAT with the outcome of this phase providing the final template. Figure 4.6 (bottom row) shows the outcome (after alignment) of this two phase PPCCAT process for 30 iterations of the four different actions shown in Fig. 4.6 (top row). This highlights the increased robustness when compared to the standard template creation process shown in Fig. 4.6 (third row).

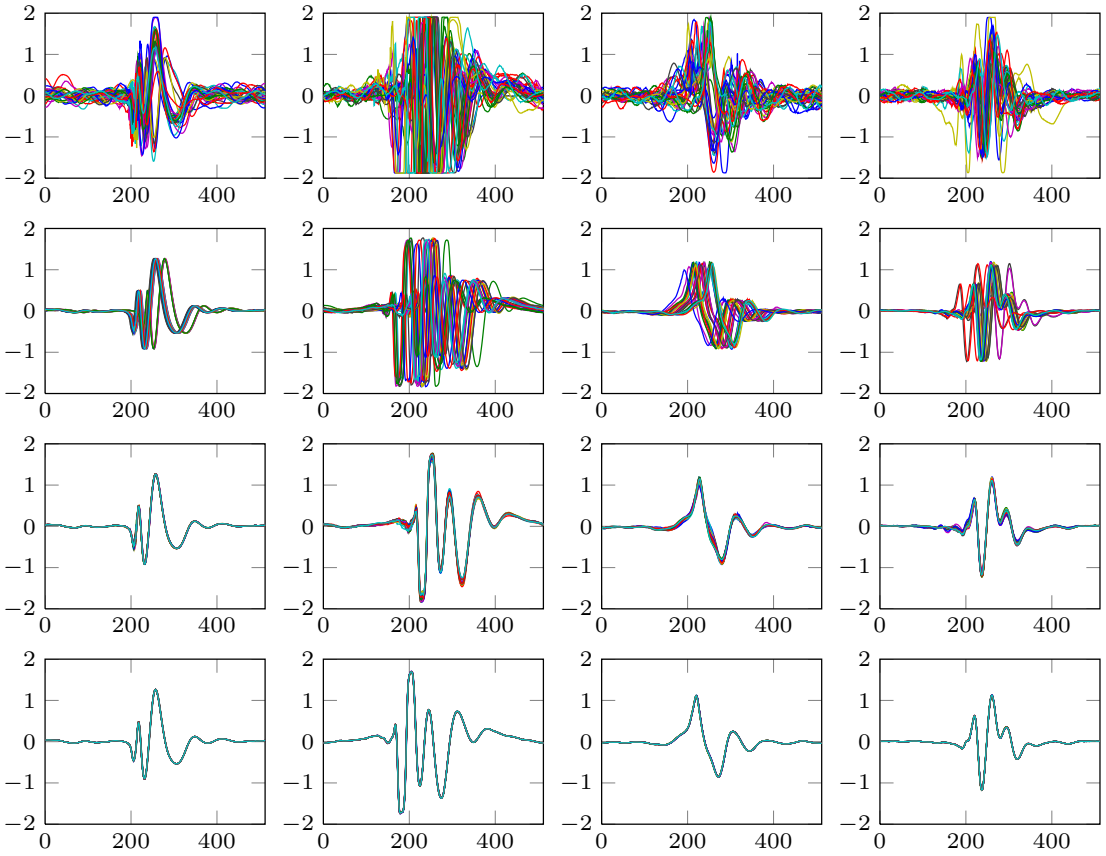


Figure 4.6: The robust template creation process: The input patterns (top row) are randomly ordered B times and the PPCCAT method is used to create B templates (second row) with these templates then used as input to a secondary PPCCAT to create a final robust template. The 32 templates are aligned and shown on the third row allowing for comparison to 32 aligned robust templates created using the randomised PPCCAT method (bottom row) created from random initial orderings of the input patterns.

4.4.5 Evaluation of template creation procedures

To show the effect of the free parameter B and the robust template creation process in general, the jackknife standard error and bias is proposed for comparison across various B. The template can be considered a univariate sample statistic, $\tilde{\mathbf{\Theta}}$, associated with a single channel pattern set and is similar to the expected central tendency associated with the pattern. The standard error of a sample statistic is defined as the estimated standard deviation of the error in the process by which it was generated. In this case the sample statistic is the template creation process (similar to the mean and therefore smooth to changes in data) and the generating process under test is the robust template creation process under varying numbers of random initialisations. The jackknife is preferred to a bootstrap method, as the general template creation process does not use multiple instances of an input pattern. The bias is the difference between the estimators expected value and the true value of the parameter being estimated. In this case, the true value $\mathbf{\Theta}$ of the template is unknown and therefore is also estimated from the robust template creation process using the entire data set and is given by

$$\mathbf{\Theta}_n = \bar{\mathbf{x}}_{n,1:N_p} \tag{4.16}$$

For this same data set ordering, the j^{th} jackknife estimate of the template is

$$\tilde{\mathbf{\Theta}}_{n,j} = \bar{\mathbf{x}}_{n,1:j-1,j+1:N_p} \tag{4.17}$$

and can be considered the template created from all input patterns bar the jth pattern. The number of jackknife samples is therefore equal to the size of the input pattern set ($N_j = N_p$). The jackknife standard error gives an estimate of the deviation between the jackknife samples and is given by

$$s.e_j(\mathbf{\Theta}_n) = \sqrt{\frac{N_j - 1}{N_j} \sum_{j=1}^{N_j} (\tilde{\mathbf{\Theta}}_{n,j} - \bar{\mathbf{\Theta}}_n)^2} \tag{4.18}$$

where

$$\bar{\mathbf{\Theta}}_n = \sum_{j=1}^{N_j} \tilde{\mathbf{\Theta}}_{n,j} \tag{4.19}$$

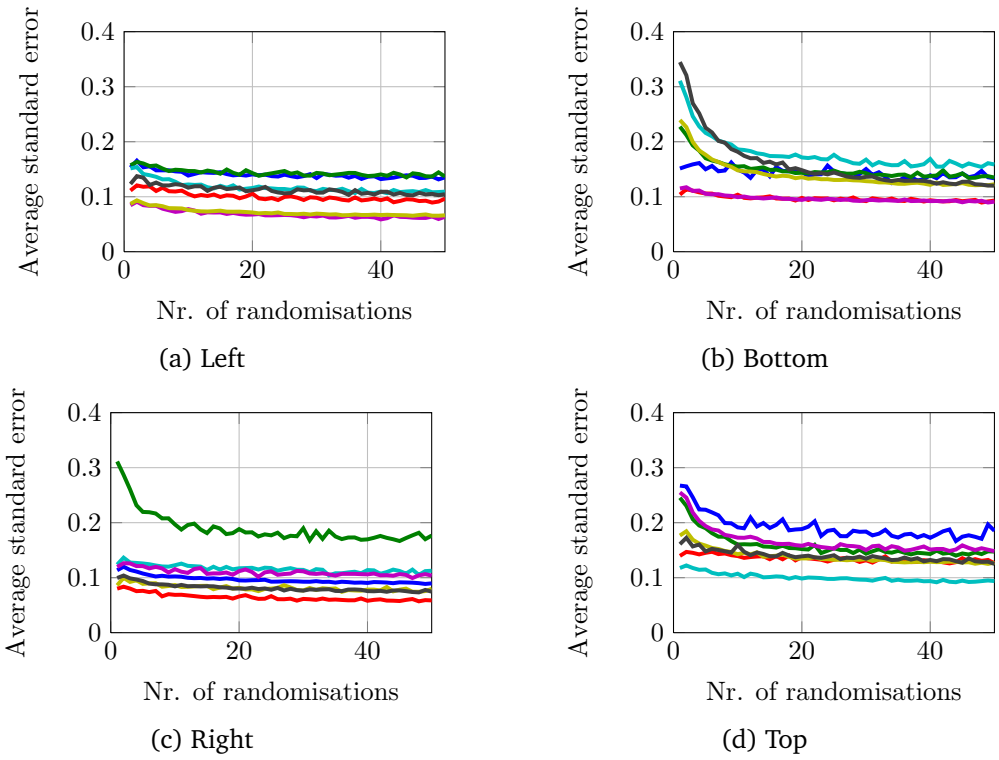
This value can be averaged over the number of samples in each pattern to produce a scalar value indicating the average jackknife standard error $s.e_j(\Theta) = \sum_{n=1}^{N} s.e_j(\mathbf{\Theta}_n)$ across all jackknife runs and samples.

The jackknife bias is given by

$$bias_j(\boldsymbol{\Theta}_n)) = (N_j - 1)(\tilde{\boldsymbol{\Theta}}_{n,j} - \boldsymbol{\Theta}_n) \tag{4.20}$$

and indicates the difference between a template created from the entire data-set and $N_p - 1$ subset of the input patterns. To obtain a scalar output associated with the bias, the average jackknife absolute bias can be defined, that is $bias_j(\Theta) = \sum_{n=1}^{N} \mid bias_j(\Theta(n)) \mid$. It should be noted that all the jackknife estimates of the templates $\boldsymbol{\Theta}_{n,j}$ are aligned to $\boldsymbol{\Theta}_n$ so that it is only the morphology of the templates which is considered.

Figure 4.7 and 4.8 show results for the average jackknife standard error and average jackknife absolute bias against the number of initial randomisations, B. The results are for the four standard actions, shown on separate subplots and for the seven well trained custom earpiece subjects (Group A), shown as separate line series within each graph. Results have been averaged over eighteen different runs using six random permutations of three disjoint data sets associated with a specific action and subject. Each of the three disjoint data sets are of size, $N_p = 31$ and selected from a total of over hundred signals for each subject and action.



(a) Left (b) Bottom

(c) Right (d) Top

Figure 4.7: Average jackknife standard error against Nr. of randomisations. Each line is associated with a subject.

These plots indicate that there is a general trend of reduction in the average stan-

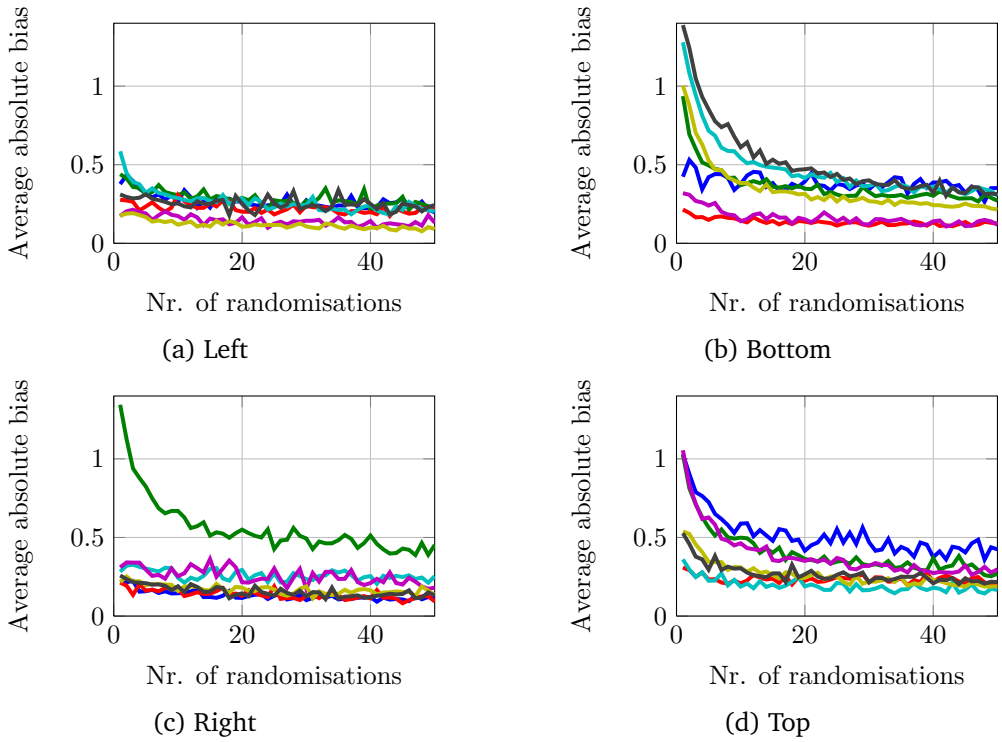


Figure 4.8: Average jackknife absolute bias against Nr. of randomisations. Each line is associated with a subject.

dard error and average absolute bias values as the free parameter, B is increased. $B = 1$ is equivalent to the standard PPCCAT template creation without any resampling. This indicates that the resampling process is increasing the stability of the template creation process by making it less sensitive to the initial training set order. This is especially important when there is higher intra-class variance associated with the initial pattern set. For very consistent patterns, there is sometimes no direct benefit, however the only drawback from using such a procedure is an increase in computation. The asymptotic nature of the results (both in the standard error and the absolute bias) indicate that, for TMEP signals, a reasonable value for the number of initial randomisations is $B \approx 20$. After this, the decrease in the standard error and the absolute bias is minimal and thus the increase in computational overhead becomes less justified. However if computation is not an issue, then increasing B further is a viable option.

4.5 Signal standardisation

Data scaling is an important preprocessing step prior to the implementation of any decoding algorithm. It ensures that no single class dominates the classification procedure due to, for instance, biasing introduced by discrepancies between signal magnitudes as-

sociated with individual classes or from contributions from specific features. Generally, this scaling is achieved using either normalisation within a fixed range (that is [0,1] or [-1,1]) or through standardisation which transforms the data so that it has zero-mean and unit variance. In terms of feature scaling, it is generally implemented to ensure that each feature has a similar contribution and thus the scaling operator is computed across the instance space. For time domain signals (where each sample can be considered a feature) this is not appropriate, as it is the differences in features (or samples) which define the morphology of the signal, especially in the case of impulsive signals. Furthermore, in real-time applications, each new test signal needs to be scaled on an individual basis or in relation to a previously captured training set.

For signal scaling in this work, standardisation is preferred to normalisation for the following reasons (1) it ensures the baseline of the signal remains at zero, (2) class differences between maxima are retained and (3) it is less susceptible to outliers. Standardisation of a vector $\mathbf{x}$ is achieved using the Z-score and is given by

$$\check{\mathbf{x}} = \frac{\mathbf{x} - \mu_{\mathbf{x}}}{\sigma_{\mathbf{x}}} \tag{4.21}$$

where $\mu_{\mathbf{x}}$ and $\sigma_{\mathbf{x}}$ are the associated mean and standard deviation. By performing the scaling on an individual basis, it assumes intra-signal variations are unaffected while inter-signal variations can change over time.

4.6 Chapter summary

This chapter has presented methods associated with extraction of volitional activity from bioacoustic signals. To extract and exemplify waveforms from a continuous stream of bioacoustic data, three processing stages are required, namely activity detection, signal segmentation and template creation. Activity detection can essentially be thought of as flagging when interesting bioacoustic activity occurs. In real-time monitoring of physiological signals, activity detection is a necessary step and therefore a brief review of activity detection in this domain was conducted. Following detection, segmentation extracts a finite length segment of the signal for further processing. This allows segments of volitional activity to be stored, for example, to use as a training set. Template creation exemplifies the class-specific morphologies associated with a set of similar patterns, generated from repetitions of the same action. This can be achieved using a robust template creation process which has been devised based on a randomised two stage PPCCAT implementation. Furthermore, it was evaluated on TMEP data in terms of the jackknife average standard error and average absolute bias to assess its robustness. This indicated that the produced templates were more stable in regards to different permutations of the data set than previous methods. A brief description of the signal standardisation process has also been given

and is utilised prior to decoding the signal to ensure that no single class dominates the classification process.

Chapter 5

Decoding bioacoustic activity

Within assistive HMI applications, the main challenge is determining the user intention from the sensory inputs. This intention is expressed by performing a prescribed intention from a specified alphabet of M actions. Therefore, it can be thought of as a M-class polychotomous classification problem using the extracted signal segments as input. In real-time environments, it is highly unlikely that the activity detection subsystem will be able to reject all nonvoluntary activity. This implies that interfering signals will be prevalent in the segmented bioacoustic signals and thus additional mechanisms for rejecting such activity need to be explored.

In this chapter, tools from the pattern recognition domain are used as the building blocks for decoding intention from bioacoustic segments. A general classification model is presented for performing inter-class classification based on a single channel. Central to this generalised model is the alignment process, whereby unseen test instances are initially aligned to class-specific templates. The standard process associated with pattern recognition include feature extraction, feature selection and classification, and thus various concepts have been explored in regard to each of these. Throughout this chapter, the proposed methods have been evaluated on off-line TMEP data as previously described in Section 3.5. The culmination of this chapter is the formulation and implementation of a heterogeneous ensemble classifier based on multiple classification models and source channels. Using preferential voting, this ensemble framework allows for effective interaction classification whilst providing an additional mechanism for rejecting interfering activity.

5.1 Learning from examples

Broadly speaking any method which incorporates information derived from training examples, when designing a function to perform classification, can be considered a learning problem [158]. There are various forms of learning which can occur and are generally

categorised into the three main groups of supervised, unsupervised and reinforcement learning problems[1]. During supervised learning, the training data is labelled using ground truth which is used to guide the learning process in a supervisory manner. This learning paradigm involves a type of formative assessment which can be used to adapt or estimate the learning model in a way that is analogous to being educated by a teacher. Unsupervised learning has no such ground truth associated with the data and instead the task is based on clustering within the feature space using natural groupings in the data [159]. This type of self-learning, groups similar input patterns together with different algorithms leading to different cluster formations. On the other hand, reinforcement learning involves the search for suitable actions in a given situation [160]. The selection of an action is based on the maximisation of some reward. A desired label is not actually given with the training examples, instead the learning algorithm must discover the appropriate actions through trial and error. Typically, there is a sequence of states and actions through which the algorithm can interact with its environment.

In general, if labelled training data is available, then a supervised learning approach should be employed and only if this ground truth is unavailable should an unsupervised approach be considered. In HMI applications, the range of volitional activity is known a priori and thus a labelled training set representing each class can be collected prior to classification. This implies that a supervised classification approach can be used. Additionally, impulsive bioacoustic signals are subject-specific in nature implying that supervised learning has to be individually performed for each user. These signals can be thought of as time domain patterns and therefore the learning problem can be defined as a pattern recognition task. Pattern recognition is based on the assumption that a pattern shares common attributes with other patterns from the same class, with the hope that they are unique to their specific domain and as such, can be considered a named entity [159]. In reality, it will never be feasible to pick a set of attributes which are truly unique to their specific domain implying that the underlying class specific pattern distributions will overlap, leading to classification errors. A classifier which is able to achieve this minimal error rate is known as the Bayes optimal classifier [161]. This error is minimised when inter-class separability is maximised while intra-class variance is minimised. For practical reasons, many classifier models make a priori assumptions about the distribution of the data. This implies that there will be a bias associated with the model and unless chosen suitably this theoretical minimum will be far from achievable.

The fundamental goal of the classification task is to minimise the difference between a classification models generalisation error and the Bayes optimal error. If discriminative features are extracted from the source data, it can make the job of the classifier trivial. The traditional goal of feature extraction is to characterise the pattern domains so that they

[1]In recent years, cross-fertilisation between these three groups has occurred, leading to new learning paradigms such as semi-supervised learning

are descriptive of their class, while providing a high level of distinction between categories [158]. Feature extraction can be considered a classification pre-processing step and not only introduces invariance to irrelevant differences among the classes but can also speed up classifier computation, which is especially important in real-time applications. Further to this, a subset of features can be selected through feature selection techniques, reducing the original set of extracted features and providing a simpler, more descriptive model. This follows from the law of parsimony put forth by William of Ockham (c. $1285 - 1349$), also known as Occam's razor, which states that preference should be given to the simpler of two competing explanations [162]. In the pattern recognition domain, this implies that a simpler classification model will generalise to unseen test data more effectively than a more complex model.

5.2 Learning challenges

The subject-specific nature of volitional bioacoustic signals implies that unique sample sets are required from each user to train the classifier with. The collection of a large personal training set, prior to using a HMI, can be extremely inhibitory and as such reduces the useability of the device. This implies that generally, only a small sample of instances will be available to train the system on. This is known as the small sample problem and can lead to less robust model estimation. For example, in feature selection, small samples can lead to unstable feature subsets being selected whereby different feature subsets are chosen by different training samples [163]. Stability, in this context, is defined as the robustness of the parameter estimation process to differences in training sets drawn from the same generating distribution [164]. Therefore, the small sample issue is intrinsically linked to the stability associated with classifier model estimation.

The challenge associated with small samples is further exacerbated when dealing with high dimensional feature spaces. Within the signal classification domain of pattern recognition, the initial feature set can be considered the time domain samples and as such, is relatively large. It depends upon the bioacoustic characteristics which dictate both the segment length and sampling frequency. In the case of TMEP signal capture this has been set at 256 ms and 2000 Hz respectively, thus implying an initial feature set size of 512 samples. As class-specific signal characteristics are unknown prior to learning, a large initial feature pool is generally required. Therefore, it is unfeasible to expect an individual to collect enough training signals to match or exceed the dimensionality of the feature subset.

To highlight the combined problem of having both a high dimensionality and small sample size, that is $N >> N_p$, can be highlighted through estimation of the covariance matrix [165]. The covariance matrix and its inverse are essential to the multivariate Gaussian

distribution and as such, play a central role in many pattern recognition models. These include the naïve Gaussian Bayes classifier, principle components analysis and Gaussian mixture models, to name a few [160]. For full estimation, due to its symmetry, the covariance matrix requires $N(N-1)/2$ parameters to be estimated. Therefore, if not enough observations are available compared to the number of parameters to be estimated, it will have linearly dependent rows and columns. This implies that it will be singular, non-invertible and have zero-valued eigenvalues, making estimation of the aforementioned models intractable and erroneous **??**.

5.3 Classifier evaluation

Before presenting the proposed decoding strategies and their constituents, it is useful to describe methods by which the discrimination strategies will be evaluated based on the outcomes from classifying positive and negative instances. A positive instance is an intended action while a negative instance is an interfering signal. Both off-line and real-time evaluation can be based on updated measures of sensitivity and specificity with a cross-validation method employed during off-line performance estimation. Cross-validation allows the ability of a decoding strategy to generalise to unseen data be assessed based on finite data sets.

5.3.1 Performance measures

Within multi-label classification, presenting just a single error rate can be uninformative when assessing the ability of a classification system. Instead, it can prove more useful to highlight the multinomial error rates associated with incorrectly assigning an instance from class $m, m = 1, \ldots, M$ to one of the $M-1$ remaining classes. Confusion matrices provide a useful tool for visualising the performance of a multi-class classifier in this manner [166]. Figure 5.1 shows the standard confusion matrix with an element from row m and column c indicating the probability of labelling an instance from class-m as class-c.

The main diagonal of the confusion matrix represents the true positive (TP) rates while conversely the off-diagonals represent the individual errors or false negative rates. From the confusion matrix, the average classification accuracy across the M-classes can be calculated as follows

$$\text{CA} = \frac{\sum_{m=1}^{M} \text{TP}_m}{\sum_{m=1}^{M} \text{TP}_m + \sum_{m=1}^{M} \sum_{c=1}^{M} \text{FN}_{m,c}} \tag{5.1}$$

Within the context of bioacoustic signal classification, due to the susceptible nature of these signals to merge with interfering signals in realtime, it is common for the decoding strategy to incorporate interference rejection capabilities into the system. Therefore, for multi-class classification strategies with an additional interference rejection mecha-

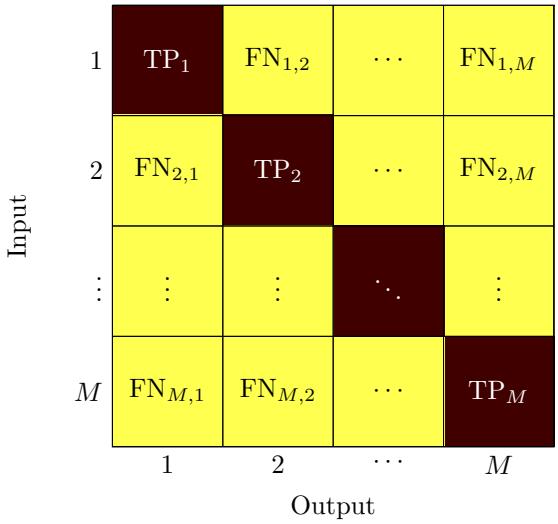


Figure 5.1: Standard confusion matrix for multi-class classification

nism, the confusion matrix has been redefined to incorporate evaluation of this additional mechanism. Figure 5.2 presents this updated confusion matrix. Based on this matrix, an instance with label m is considered a positive instance while anything else, that is interference, is considered a negative instance. Based on this confusion matrix, there are now five different outcomes that can occur, two are associated with being classified as a positive instance including (1) an instance from class m being correctly classified as class m (TP) or (2) an instance from class m being misclassified as class c with $c \neq m$ (FN^+). The three remaining outcomes are associated with assigning a negative class-label, including (1) an instance of class m being rejected as interference by the system (FN^-), (2) an interfering signal being correctly identified as interference (TN) or (3) an interfering signal being incorrectly classified to one of the m positive classes.

As both positive and negative classes are now being considered, it is useful to define the statistical measures of sensitivity and specificity [167]. Based on the updated confusion matrix, two new sensitivity measures can be defined. The first considers only the positive instances, ignoring (FN^-) and is given by

$$\text{SN}^+ = \frac{\sum_{m=1}^{M} \text{TP}_m}{\sum_{m=1}^{M} \text{TP}_m + \sum_{m=1}^{M} \sum_{c=1}^{M} \text{FN}^+_{m,c}} \tag{5.2}$$

This measure can be considered closely related to the CA measure associated with the standard confusion matrix given in Equation 5.1 earlier. If no response from the system, due to rejection of a positive instance, is considered a noncritical issue based on the application, then this is a useful measure. If the rejection of positive instance does have a detrimental effect in terms of the response of the peripheral being controlled, then such

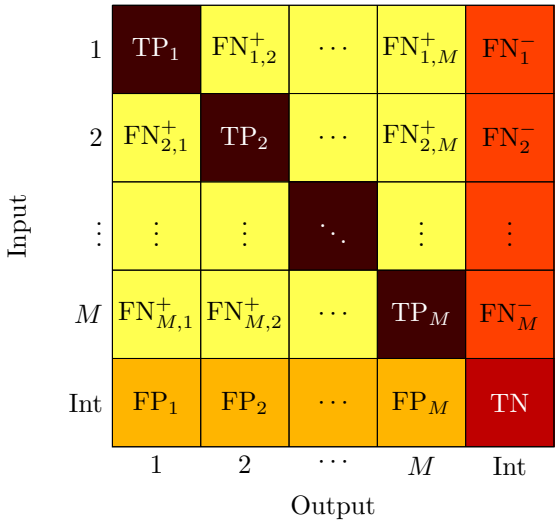


Figure 5.2: Confusion matrix developed for a multi-class classification system with interference rejection capabilities

instances need to be taken into account. The sensitivity can therefore be redefined as

$$\mathrm{SN}^{\pm} = \frac{\sum_{m=1}^{M} \mathrm{TP}_m}{\sum_{m=1}^{M} \mathrm{TP}_m + \sum_{m=1}^{M} \sum_{c=1}^{M} \mathrm{FN}^{+}_{m,c} + \sum_{m=1}^{M} \mathrm{FN}^{-}_{m}} \tag{5.3}$$

Conversely, specificity measures the proportion of negative cases that were correctly identified by the interference rejection mechanisms and is given by

$$\mathrm{SP} = \frac{\mathrm{TN}}{\mathrm{TN} + \sum_{c=1}^{M} \mathrm{FP}_c} \tag{5.4}$$

Finally, the specificity can be combined with the different sensitivity measures to give an overall indication of the performance of a classification strategy. Equation 5.5 below is associated with SN^{+} while Equation 5.6 is associated with $\mathrm{SN}^{\pm}$.

$$\mathrm{CA}^{+} = \frac{\sum_{m=1}^{M} \mathrm{TP}_m + \mathrm{TN}}{\sum_{m=1}^{M} \mathrm{TP}_m + \mathrm{TN} + \sum_{m=1}^{M} \sum_{c=1}^{M} \mathrm{FN}^{+}_{m,c} + \sum_{c=1}^{M} \mathrm{FP}_c} \tag{5.5}$$

$$\mathrm{CA}^{\pm} = \frac{\sum_{m=1}^{M} \mathrm{TP}_m + \mathrm{TN}}{\sum_{m=1}^{M} \mathrm{TP}_m + \mathrm{TN} + \sum_{m=1}^{M} \sum_{c=1}^{M} \mathrm{FN}^{+}_{m,c} + \sum_{c=1}^{M} \mathrm{FP}_c + \sum_{m=1}^{M} \mathrm{FN}^{-}_{m}} \tag{5.6}$$

5.3.2 Off-line performance estimation

During off-line validation of the proposed decoding strategies, there are various methods by which classification performance can be estimated. Ultimately, this performance evaluation is equivalent to populating the confusion matrix using previously collected data set

$\mathcal{D}$. Depending upon the importance of the data set and the constraints associated with specific test conditions, there are various protocols that can be used for classifier performance estimation [168, 169]. Of these, the easiest to implement is the *Resubstitution method* which utilises the whole data set for training the classifier and then reuses the same data during testing. However, this leads to optimistically biased results, as well as little scope for testing different training set sizes without reducing the overall amount of data. To remove the effect of this biasing, a *Holdout method* can instead be used. This is commonly known in the literature as a resampling test and involves partitioning the data into two mutually exclusive and random subsets [168]. One of the subsets is used for training while the other is used for testing. It is considered a more pessimistic estimator, whereby a larger test-to-training set ratio leads to a higher bias of the estimate [169]. Conversely, fewer instances imply larger confidence bounds associated with the estimate. Furthermore, the *Holdout method* doesn't use the data instances uniformly and only asymptotically approaches uniformity as the number of *holdout* iterations tends to infinity. The *Bootstrap method* is similar in concept, except now the training set is sampled using replacement, with the remaining instances used for testing. This allows for a training set of any size to be specified and with enough iterations, can give low variance estimates. The bias can be extremely large on some problems and can be, in part, attributed to repeated instances used during training.

A good solution to the estimation problem is the *Leave-one-out method* which utilises the whole database at training, bar one instance [160]. The scheme is iterated, leaving out a different instance each time. Although, it provides very precise results, this protocol is computationally demanding and additionally doesn't allow for control over the number of training instances. Instead, K-fold cross-validation (K-CV) can provide a compromise between the precision and complexity of the estimation problem. It includes the *Leave-one-out method* as a special case, that is, when $K = N_p$. Rather than leaving out a single instance, CV leaves a number of instances out, with this number equal to $\lfloor N_p/K \rfloor$. It is often used to get an appropriate estimate of predictive accuracy [168]. Ten-fold stratified CV has been highlighted as the most effective during model selection and accuracy estimation [169]. Figure 5.3 illustrates the CV process for $K = 3$ folds with stratification implying that within each fold, the number of instances of each class are equal.

In standard K-CV, the data set is split into K disjoint sets. At each iteration of a CV trial, one fold is used as validation set while the remaining $K - 1$ sets are grouped and used to train the learning algorithm. This is repeated K times, using a different validation fold at each iteration. A confusion matrix is computed based on the performance across the K-folds indicating that each instance has been used once for testing during a single CV trial. The process can be repeated a number of times so that average performance measures and the associated variances can be assessed.

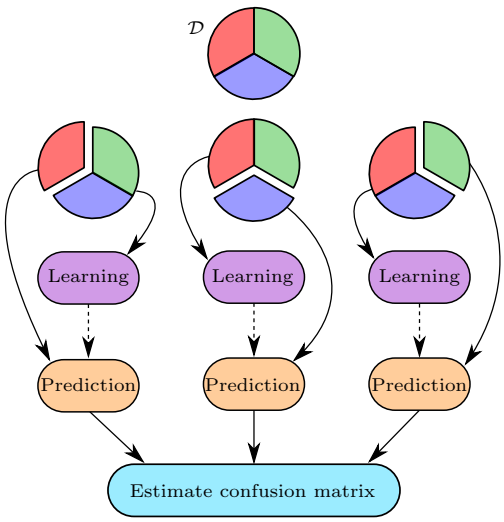


Figure 5.3: Highlighting the CV procedure for $K = 3$, with each segment representing a stratified fold of the current feature subset being evaluated. The same idea holds for any size K.

The following cross-validation procedure is used in this work for classification performance estimation based on off-line data-sets. To allow for different size training sets, the stratified K-CV procedure has been adapted. Figure 5.4 illustrates this modified CV procedure. In each CV trial, $\mathcal{D}$ is randomly permuted and split into K-folds. In the same manner as stratified K-CV, each fold in turn is used as a test (or validation) set. A proportion of the subsequent instances are used for training with the remaining instances left unused. This implies that the size of the training set can be adjusted to suit the purposes of the evaluation process. It should be noted that this modified CV procedure uses all the data samples an equal amount of times during a single CV run and thus retains a good bias-variance trade-off that is associated with standard CV techniques.

5.4 Single channel classification model

This section describes the general classification model developed for individually decoding single channel volitional bioacoustic signals. Initially, single channel classification is performed with the results presented in Section 5.10.4. Stereo channel results based on the multi-channel heterogeneous ensemble framework are formulated and presented in Section 5.11. Figure 5.5 highlights the generalised structure of the proposed single channel classification model developed by the author. This shows the flow of processing for both training set and test instances. During training, a template exemplifying each signal class is created from pre-collected training data-sets. This is achieved using the robust template creation method outlined earlier in Section 4.4. The class-specific templates serve two

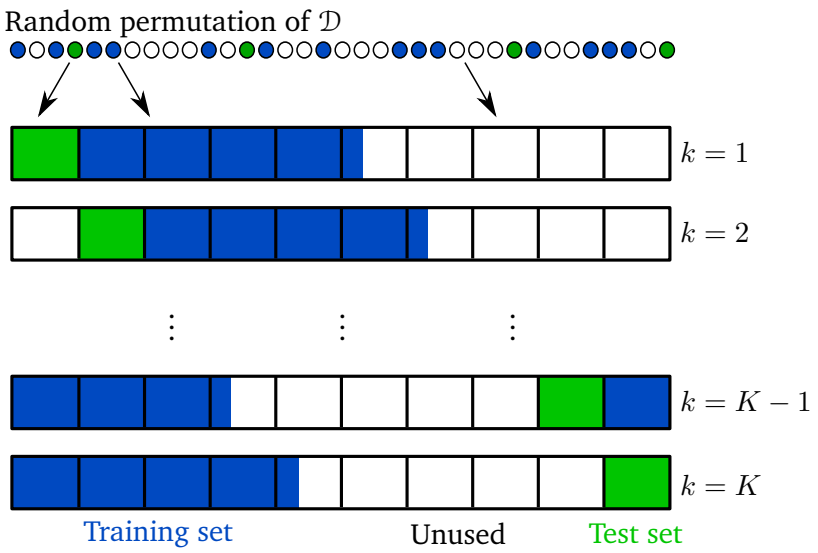


Figure 5.4: Partitioning of $\mathcal{D}$ during a single CV run with each box representing a disjoint set of K-folds.

purposes, firstly, each one can be considered the representative signal associated with a particular class and therefore used during classification as a robust central tendency estimate associated with the signals. Secondly and most importantly, it allows signals to be aligned during classification, significantly decreasing intra-class variance. For bioacoustic and biosignals in general, the time domain signature can be considered a key descriptive feature of the signals. The segmentation approach, although capturing the majority of the energy associated with the volitional activity, is not optimal in terms of placement of the signal segments within the window. Furthermore, by shifting the signals based on morphological commonalities, similar signals will shift coherently while dissimilar signals should shift more randomly. This is illustrated by comparing the unaligned TMEP data sets shown previously in Fig. 3.4 to their class-specific aligned counterparts in Fig. 3.5.

If a large enough data set is available then the instances used for template creation and classifier parameter estimation can be independent, potentially leading to better generalisation by the classification model. These training sets can then be used during feature extraction, feature subset selection and to estimate the class-specific classifier model parameters. During the testing phase, the label of the instance is unknown and therefore alignment to all of the templates is required. This creates M time-shifted replications of the original signal. It is due to these replications that a class-specific classification model is required and implies that many classical black-box classification approaches, for example, ANNs and multi-class SVMs cannot be used [**?**]. These methods are discriminant models and therefore directly estimate the decision boundaries from the features [159]. Features are extracted from each of the replications based on the proposed feature mapping found

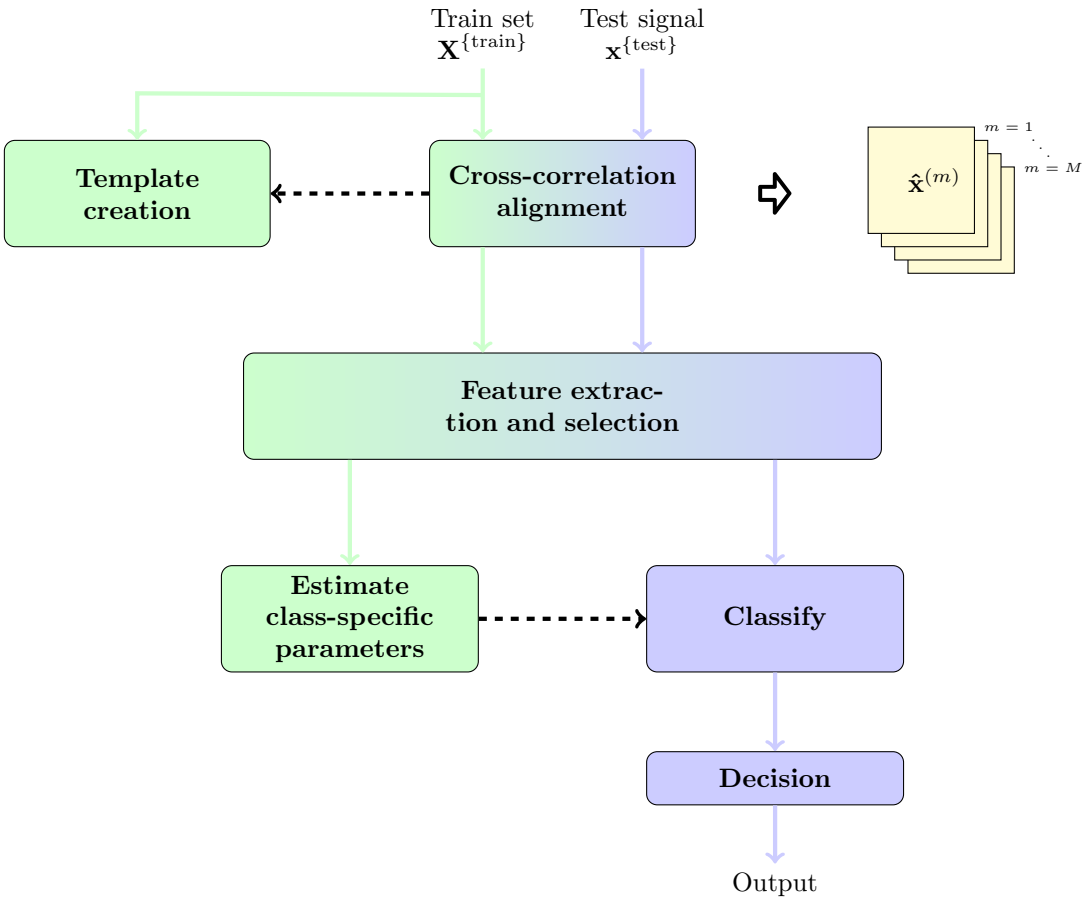


Figure 5.5: Overview of the general classification process. Green lines represent the classification model learning phase flow while the blue line are associated with the prediction phase. Black dashed lines are associated with both the learning and prediction phase. The important template alignment stage creates M time shifted replications of the input signals which means class-specific classification models are required.

during training. The class-specific features are fed into their corresponding class-specific classification models using the parameter values estimated during training. A final decision block is used to combine the output of these M-models, assigning a class label to the test instance.

5.5 Feature extraction

Feature extraction methods create new features based on transformations or combinations of the original feature set [159]. In the case of bioacoustic and bio-signals in general, this original feature space is the time domain samples. Feature extraction can be thought of as the process of mapping the signals to a new subspace whereby class-specific information and thus discriminatory power is extracted from them. In many cases, it is used to reduce the number of features associated with the original patterns. Having said this, it is not necessarily simply a reduction in the dimensionality associated with the signals but can also extract latent information leading to better classification performance and generalisation. It precedes feature selection, which attempts to select the best subset of these extracted features based on their performance under some classification system [163].

For impulsive bioacoustic signals, the unique signature of the generating motion is generally already visible in the waveform. This implies that the time domain features themselves already contain useful discriminatory information due to distinct morphologies. Spectral content of the bioacoustic signals provide another useful avenue for extracting effective features using Fourier analysis. For impulsive bioacoustic signals, class-specific frequency responses are likely to be distinct based on various factors. These include the different physiological mechanisms underlying their generation and translation, the speed of the generating motion and varying resonant frequencies associated with the anatomy. Frequency features can be extracted based on the DFT and discrete cosine transform (DCT) which are implemented using the fast Fourier transform (FFT) [170, 171]. Recently, time-frequency based techniques have received increased attention for extracting discriminative information from bioacoustic and bio-signals in general [15]. These include the short-term Fourier spectrum, multiresolution wavelet analysis and the Wigner-Ville distribution, to name a few. Although considered superior at representing non-stationary signals, they all abide the Heisenburg uncertainty principal stating that you cannot be both localised in time and frequency [170]. Furthermore, short duration signals can be assumed quasi-stationary in small time frames, although this is not always the case. Subdividing the signal further will lead to loss of important spectral content while significantly increasing the number of features. Due to their reduced spectral resolution, time-frequency methods are considered unsuitable for feature extraction of short bioacoustic signals representing volitional activity.

Many speech recognition systems extract frequency based features using the cepstrum which is the inverse Fourier transform of the log spectrum [172]. Cepstral coefficients are useful as they isolate the two components of speech, namely the excitation and vocal tract components. Although they have found applications in other domains, they are generally only suitable for analysing signal models that are generated in a similar manner as the speech model.

Although modelling of the processes underlying a specific bioacoustic HMI concept is infeasible, a general signal model can still be utilised. The model parameters are estimated from the signals themselves with these parameter values used as discriminatory features. A common model employed in bio-signal analysis is the AR model given its suitability for modelling random processes and natural phenomena in general [96]. Another data-dependent transform and probably one of the most widely known feature extractors is based on PCA [159]. It projects the original data onto the most expressive orthogonal subspace by projecting the signals in the directions of maximum variance with this set of uncorrelated variables called the principal components. For extraction of non-Gaussian features, independent component analysis (ICA) can be employed. The idea of ICA is to split a set of mixed signals into its sources without prior knowledge regarding the nature of the signals. The only assumption is that these underlying source signals are mutually independent in statistical terms [96]. Recently, solutions to the under-determined ICA problem have been proposed, whereby the number of mixed signals is much smaller than the number of original sources and can potentially be just a single channel [173, 174]. ICA has traditionally been used as a preprocessing step, for removing mutually independent artifacts from multichannel biosignals. For decomposing a signal into useful features, although not unheard of, can be less effective especially if the signal comes from a single generating action and therefore can be considered a single source.

More complex features can be extracted from biosignals based on nonlinear transformations. These include but are not limited too, kernel PCA, multidimensional scaling, locally linear embedding, isomaps, higher order statistics and features based on nonlinear dynamics and chaos theory [159, 175–177]. Such methods, although having the potential for characterising biological phenomenon more efficiently, have several significant drawbacks. These include free parameters to select, high computational costs, low interpretability and are generally complex models requiring large samples for estimation. Many require intrinsic geometric relationships or manifolds presence within the original data for effective features to be extracted and were intended predominantly for visualisation and analysis purposes. Nonlinear feature extraction methods are therefore generally not very useful for real-time extraction of biosignal features especially if the signals are of shorter duration signals and only a small sample available. In the following sections, various linear feature extraction methods are introduced for utilisation within the bioacoustic

decoding framework.

5.5.1 Autoregressive coefficients

All signals are generated from an underlying model. If the model is known then parametric estimation can be performed, using the data to extract the coefficients that describe such a model. In many cases, the generating physical model is unknown or too complicated, thus requiring over-simplification. Instead, a general signal model can be employed, which is suitable for modelling random processes and thus various types of natural phenomena [96, 152]. A typical model employed in bio-signal processing is the AR model and is given by

$$x_n = \sum_{k=1}^{p} A_k x_{n-k} + e_n \tag{5.7}$$

where $p \in \mathbb{Z}^+$ is the model order, A_k are the real-valued AR coefficients and e_n represents white noise independent of the underlying process. An AR model assumes that the current data point x_n is closely related to the p previous data points. It can be considered an all-pole model and as such, gives higher spectral resolution as opposed to say, a moving average (MA) process.

There are two main methods for estimating the model coefficients A_p, namely, the Yule-Walker autocorrelation method or Burg's method [152]. They both try to minimise the mean-squared prediction error between the output of the AR model and the actual data. The Yule-Walker method minimises the following

$$\xi = \min\{ \sum_{n=0}^{N+p-1} (e_n^+)^2 \} \tag{5.8}$$

where $e_n^+ = x_n + A_1 x_{n-1} + A_2 x_{n-2} + ... + A_p x_{n-p}$ is the forward prediction error. It minimises only the forward prediction error and runs the filter off the end of the sequence (equivalent to zero-padding the data). This ensures the filter is minimum phase but can lead to erroneous estimations, especially for small data segments. Conversely, the Burg method tries to minimise both the forward and backward prediction error without running the filter off the end of the data segment and is given by

$$\xi = \min\{ \sum_{n=0}^{N-1} [(e_n^+)^2 + (e_n^-)^2] \} \tag{5.9}$$

where $e_n^- = x_{n-p} + A_1 x_{n-p+1} + A_2 x_{n-p+2} + ... + A_p x_n$ is the backward prediction error. By using an iterative method, Burg was able to achieve the minimum phase property without padding the ends of the data segments. By incorporating the backward prediction error, Burg's method is considered to give a more accurate estimation of A_p, especially for small

data segments [178]. The model order p, is a free parameter and is generally chosen based on various overfitting criteria, such as the Akaike information criterion (AIC) [151]. Within a classification framework, the model coefficients can be used as features, such that $\mathbf{s} = [A_1, A_2, \ldots, A_p]^T$, with p selected as the one which provides the maximum CA. Finally, it should be noted that the coefficient values are both data and model order dependent and thus using different model orders yields completely different coefficients. This implies that feature selection is not necessary as p needs to be determined prior to feature extraction.

Figure 5.6 shows AR coefficients extracted from 32 instances of four bioacoustic TMEP actions for one subject. Even with this large reduction in the number of features (compared to the time domain) there are obvious differences between the actions. For example, the top action has a distinct shape when compared to the remaining actions. Having said this, some coefficients exhibit higher intra-class variance and therefore this could have a detrimental affect on classification performance. Without a feature selection process these coefficients would remain and could potentially degrade prediction accuracy. It should be noted that the AR model assumes that the signal being modelled is stationary [96]. Although this is probably not true for TMEP signals, it was felt that AR coefficients could still potentially contain discriminatory information. This would depend upon the sensitivity of the coefficients to variations in the signals and may be a contributing factor to the higher intra-class variance of some of the coefficients.

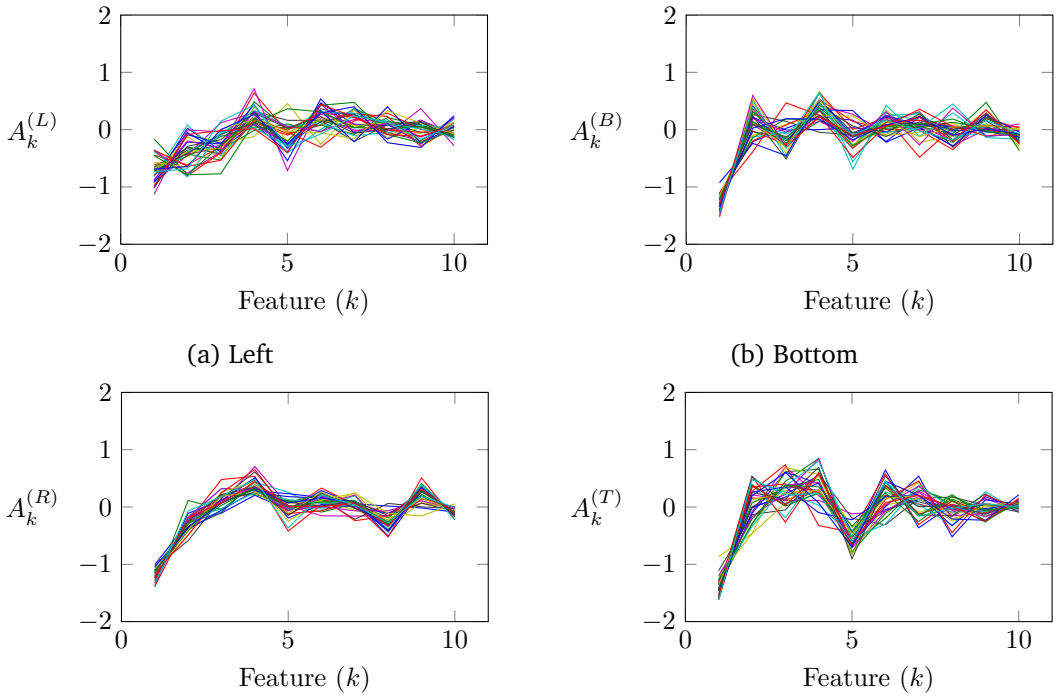
5.5.2 Discrete Fourier transform

Frequency analysis is one of the key tools used in the signal processing domain. It represents the signal in terms of a weighted linear combination of complex exponential signals called a spectrum [179]. For discrete time signals, an approximation of the spectrum can be found using the DFT and is calculated from

$$\mathcal{F}[x_n] = \sum_{n=0}^{N-1} x_n e^{-i2\pi kn/N} = s_k \tag{5.10}$$

for $k = 0, 1, \ldots, N-1$.

By evaluating the spectrum at N regularly-spaced frequency points, that is $f_k = 2\pi k/N$, the transform is machine realisable. In fact, there exists a very efficient algorithm for calculating the DFT of a signal, namely, the FFT [170]. In practise, Fourier transforms operate on a finite domain and implicitly define an extension outside of it. The function $\mathcal{F}[x_n]$ has a periodic extension with a period equal to 2π. Due to the sampling process, the frequency range for a discrete time signal is limited to $0 \leq f < f_s/2$ with frequencies outside this range mapped back onto it. This upper limit is known as the Nyquist frequency [152].



(a) Left (b) Bottom

(c) Right (d) Top

Figure 5.6: Example plots showing autoregressive coefficients for four TMEP actions from one subject found using Burg's method with $p = 10$. Each data set consists of 32 repetitions of each of the actions. The first coefficient is ignored as it is always equal to one and therefore provides no discriminatory information

As the DFT is a complex function, it encodes both the amplitude and phase information associated with a sinusoidal component of the time series. This complex representation allows various features to be computed from them. If the discrete time signal x_n is purely even, then the DFT will reduce to just its real part. Taking only the real part of the DFT gives

$$\mathfrak{Re}\{\mathcal{F}[x_n]\} = \sum_{n=0}^{N-1} x_n \cos(2\pi kn/N) = \mathfrak{Re}\{s_k\} \tag{5.11}$$

On the other hand, if x_n is purely odd, then the DFT will reduce to only its imaginary part. Taking only the imaginary part of the DFT yields

$$\mathfrak{Im}\{\mathcal{F}[x_n]\} = \sum_{n=0}^{N-1} x_n \sin(2\pi kn/N) = \mathfrak{Im}\{s_k\} \tag{5.12}$$

If the signal x_n is neither even or odd then the frequency components will experience both real and imaginary parts. Due to the orthogonality of the real and imaginary axis, the magnitude of each frequency component can be calculated from

$$|\mathcal{F}[x_n]| = \sqrt{\mathfrak{Re}\{\mathcal{F}[x_n]\}^2 + \mathfrak{Im}\{\mathcal{F}[x_n]\}^2} = |s_k| \tag{5.13}$$

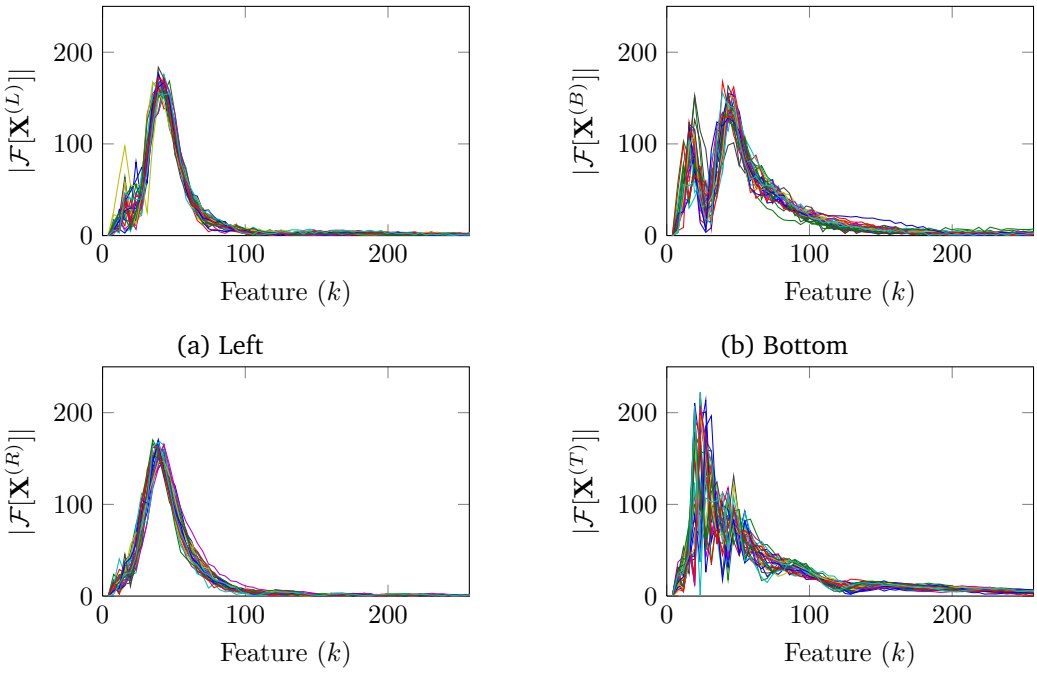
The final property that can be computed, is the phase and is given by

$$\angle\mathcal{F}[x_n] = \text{atan2}(\mathfrak{Re}\{\mathcal{F}[x_n]\}, \mathfrak{Im}\{\mathcal{F}[x_n]\}) = \angle s_k \tag{5.14}$$

and can be considered the shift in each frequency component from the corresponding zero phase wave, that is $\sin(2\pi fn)$.

Frequency information is extracted from the same 32 instances of four TMEP actions as previously. Figure 5.7 highlights the magnitude of the complex DFT coefficients for these instances. As the DFT is a redundant transform and calculates frequency coefficients up to the sampling rate, only half the coefficients need to be retained. This means the initial feature subset size is 256 features for a signal segment consisting of 512 samples. It can be seen that there are distinct areas of the spectrum which could provide a basis for effective action recognition. Take for example, the bi-modal shape of the bottom action, which shows zero information at around the 25 Hz frequency coefficient. This starkly contrasts the other actions which exhibit their critical frequencies around this point. For these signals, it is useful to point out that frequency is exhibited above 100 Hz for some actions and in the case of the top action above 200 Hz. This higher frequency content could also potentially provide discrimination between the actions. Having said this, the frequency responses of the left and right actions are visually quite similar and therefore these actions could be confused for classifiers based solely on these features. The low

intra-class variance across all the frequency responses indicates the consistency of these features.



(a) Left (b) Bottom

(c) Right (d) Top

Figure 5.7: Example plots showing the magnitude of the Fourier coefficients for four TMEP actions from one subject found using the FFT algorithm. Each data set consists of 32 repetitions of each of the actions.

5.5.3 Discrete cosine transform

The DCT is widely used in lossy data compression applications due to its superior compaction ability, including JPEG, MPEG and CCITT H.261 amongst others [171]. It is closely related to the DFT and can in fact be obtained from it. As real-world signals can typically be represented as an even function, the one dimensional DCT allows the frequency information associated with them to be represented on fewer coefficients and has similar properties to the PCA transform (described in Section 5.5.4) [180]. There are eight types of DCT which implicitly assume a left boundary even extension. Similarly, there are eight discrete sine transformations (DST) which assume a left boundary odd extension. The different variants are due to various combinations of even or odd boundary extensions and whether the points of extension are overlapped or repeated for each boundary. Therefore there are a total of $N_{types} = 2 \times 2 \times 2 \times 2 = 16$ types of discrete cosine and sine transformations. The overlapping (or repetition) at the boundaries, leads to symmetry about the boundary or the point halfway between the extensions. In this work a DCT-II variant was

chosen as it assumes both left and right even boundary extensions and therefore is more representative of real bio-signals. It is computed as follows

$$\mathcal{D}[x_n] = \sqrt{\frac{2}{N}} C_k \sum_{n=0}^{N-1} x_n \cos\left(\frac{\pi(2n+1)k}{2N}\right) = \mathbf{s}_k \tag{5.15}$$

for $k = 0, 1, \ldots, N-1$. The boundary conditions allow the transformation to be orthogonal and are given by

$$C_k = \begin{cases} 1/\sqrt{2}, & \text{if } k = 0 \\ 1, & \text{if } k \neq 0 \end{cases} \tag{5.16}$$

leading to an equal number of frequency coefficients as the time domain input vector.

The DCT (as well as the DFT and PCA) transforms can be represented as a matrix operation, between a basis matrix $\mathbf{\Phi}$ and the time domain signal vector $\mathbf{x}$. This is given as

$$\mathbf{s} = \mathbf{\Phi x} \tag{5.17}$$

where the elements of the $N \times N$ basis matrix are as follows

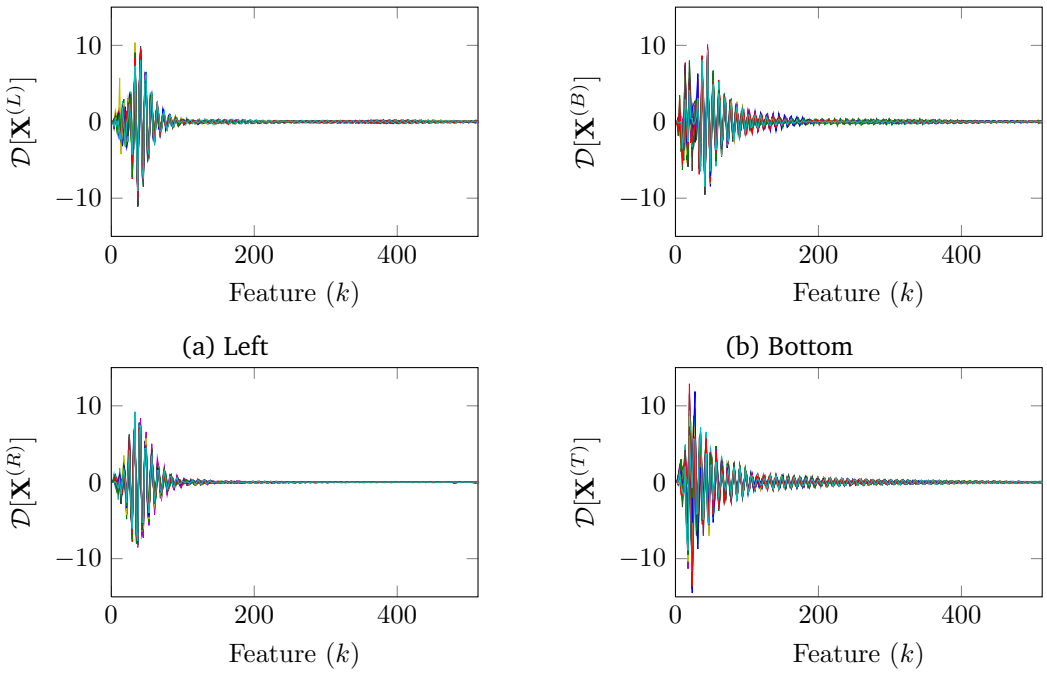
$$\mathbf{\Phi}_{k,n} = C_k \cos\left(\frac{\pi(2n+1)k}{2N}\right) \tag{5.18}$$

The DCT is a data independent transformation and expresses a time domain signal as a linear combination of cosine functions oscillating at various frequencies.

Figure 5.8 highlights the ability of the DCT for extracting features from bioacoustic signals. It shows example plots of DCT coefficients extracted from the four action TMEP data sets used previously. The DCT transform is no longer a redundant transform with all 512 coefficients conveying information. It can be thought of as an extension of the real Fourier transform with double the spectral resolution and a repeated implicit boundary. Inspecting the plots globally, it can be seen that there are distinct morphologies across the actions. The top and bottom actions show clearly unique responses but even the left and right actions are noticeably dissimilar, with the right exhibiting a slightly shorter and wider spectrum. It is also clear to see that there is high level of intra-class consistency within each of these data sets. The increase in resolution, although increasing the initial feature subset size, provides additional features compared to the standard Fourier transform which could prove important in terms of better classification performance.

5.5.4 Principle component analysis

PCA also known as the Karhunen-Loeve transform is a linear transform, frequently used to provide a compact representation of data [159]. It is defined as the orthogonal projection



(a) Left (b) Bottom

(c) Right (d) Top

Figure 5.8: Example plots showing DCT coefficients for four TMEP actions from one subject. Each data set consists of 32 repetitions of each of the actions.

of the data onto a lower (or equal) dimensional linear space. This space is known as the *principal subspace* and is based on maximising the variance of the projected data [160]. As it is a linear transformation, similarly to the DCT and DFT it can be computed using matrix operations, similar to Equation 5.17. The main difference is that the basis matrix, $\mathbf{\Phi}$ is based on the eigenvalues and eigenvectors of the covariance matrix and as such, is data dependent. An alternative formulation considers PCA as the minimisation of the projection error, between the input data and a set of orthogonal basis vectors. This implies that it can be considered optimal in terms of the mean squared error (MSE) as well as in the sense of de-correlation of the input vectors. If $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_{Np}]$ represents a matrix of zero-mean random vectors, where $\mathbf{x}_i \in \mathbb{R}^N$, then its covariance matrix is given by

$$\mathbf{\Sigma}_\mathbf{X} = \frac{1}{N_p - 1}\mathbf{X}\mathbf{X}^T \tag{5.19}$$

If each of the random vectors is projected onto orthonormal vectors, $\mathbf{u}_j \in \mathbb{R}^N$, then its projected variance will be maximised when

$$\mathbf{\Sigma}_\mathbf{X}\mathbf{u}_j = \lambda_j \mathbf{u}_j \tag{5.20}$$

where $\mathbf{u}_j$ is an eigenvector of $\mathbf{\Sigma_X}$ and λ_j is its associated eigenvalue and equivalent to the variance of the projected data along $\mathbf{u}_j$. It is assumed that directions of maximum variance are indicative of useful dynamics associated with the data. Therefore, the larger the value of λ_j, the higher the variance associated with the projection of the data onto the subspace $\mathbf{u}_j^T\mathbf{X}$ and the more informative that projected direction can be considered. The basis matrix is now comprised of the eigenvectors given by Equation 5.21, with its rows sorted in descending order of its eigenvalues. The first principal component is $\mathbf{u}_1$, $\mathbf{u}_2$ is the second and $\mathbf{u}_{Np}$ is the last principal component representing the direction of least variance.

$$\mathbf{\Phi} = \begin{bmatrix} \mathbf{u}_1 & \mathbf{u}_2 & \dots & \mathbf{u}_{Np} \end{bmatrix}^T \tag{5.21}$$

When $N_p < N$ there are dimensionality issues associated with finding the eigenvectors and eigenvalues. This is due to $\mathbf{\Sigma_X}$ having at most N_p independent rows. This implies that it can only be projected onto a linear subspace whose dimension is at most $N_p - 1$ with a minimum of $N - N_p + 1$ zero eigenvalues. To overcome this issue, a method based on eigenfaces has been introduced and is named due to its development within the facial recognition domain [181]. It is based on premultiplying both sides of Equation 5.20 by $\mathbf{X}^T$. The effect of this can be seen by substituting into Equation 5.19 giving

$$\frac{1}{N_p - 1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{u}_j) = \lambda_j(\mathbf{X}^T\mathbf{u}_j) \tag{5.22}$$

This allows for a much smaller $\mathbf{X}^T\mathbf{X}$ matrix of dimension $N_p \times N_p$ to be calculated, whose eigenvectors are now equal to $\mathbf{X}^T\mathbf{\Phi}^T$. To obtain the original eigenvectors associated with $\mathbf{\Sigma_X}$, Equation 5.22 can again be premultiplied by $\mathbf{X}$, thus giving $\mathbf{X}\mathbf{X}^T\mathbf{\Phi}^T$ which can be shown to be an eigenvalue of $\mathbf{\Sigma_X}$ using

$$\frac{1}{N_p - 1}\mathbf{X}\mathbf{X}^T(\mathbf{X}\mathbf{X}^T\mathbf{u}_j) = \mathbf{\Sigma_X}(\mathbf{X}\mathbf{X}^T\mathbf{u}_j) = \lambda_j(\mathbf{X}\mathbf{X}^T\mathbf{u}_j) \tag{5.23}$$

If a time domain signal is considered a single point in N-dimensional Euclidean space, then based on a training set, PCA can be used to find a representative subspace associated with a particular pattern. This is under the assumption that the data is normally distributed and the principle components are orthogonal to one another. Therefore, a new instance can be projected onto this *principal subspace* with the intention of collapsing the important information into fewer components. As it is a linear and orthogonal transformation, similar in this respect transformation to the the DCT, rows of Φ can be removed without affecting the remaining components. Within a class-specific classification framework, there are a number of different ways that the transformation can be implemented. This is due to the data-dependent nature of the transform and the template creation process, which creates class-specific time-shifted repetitions of each signal prior to feature

extraction. Figure 5.9 illustrates the three different implementations of the PCA transform which have varying levels of class dependence.

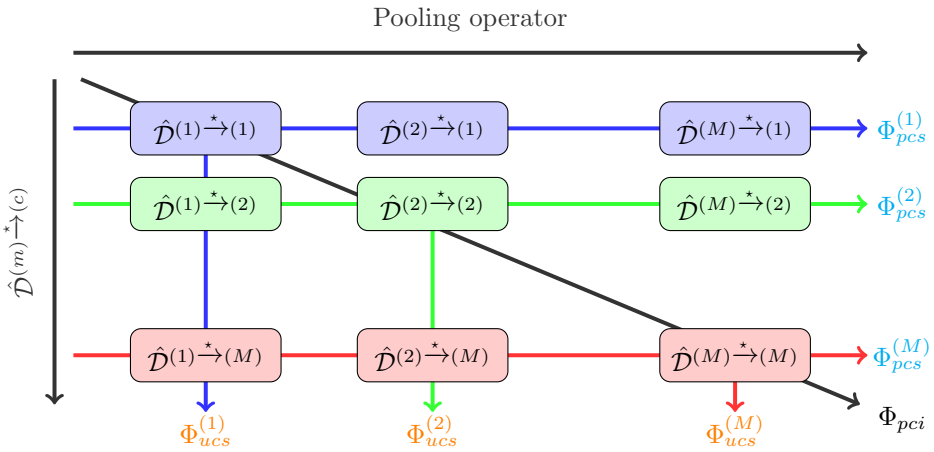


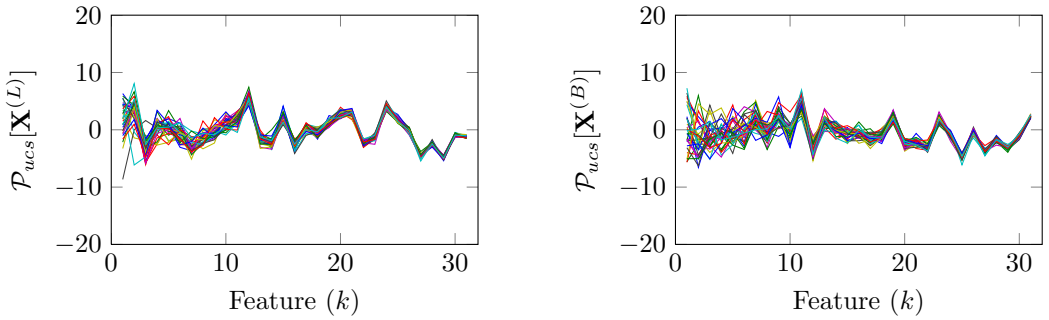
Figure 5.9: The three PCA transforms: Unpooled class-specific (Φ_{ucs}), pooled class-specific (Φ_{pcs}) and pooled class-independent (Φ_{pci}). The pooling operator is indicated horizontally while the template alignment operator is indicated vertically. $\hat{\mathcal{D}}^{(m)\xrightarrow{\star}(c)}$ indicates that the data set associated with class-m is aligned to the template created from class-c.

The unpooled class-specific implementation (PCA$_{ucs}$) uses the training data associated with a specific class, aligned to templates of the same class, creating M class-specific PCA transformation matrices. The initial feature subset size, based on these matrices, is therefore $N_p^{\{m\}} - 1$. Figure 5.10 shows example plots of these PCA$_{ucs}$ coefficients extracted from bioacoustic TMEP data associated with four actions from one subject.

The pooled class-specific implementation (PCA$_{pcs}$) also creates a class-specific transformation matrix. It does this by pooling the data across the various classes for instances aligned to each of the templates. Each of the M class-specific transforms has an initial feature subset size of $M \times N_p^{\{m\}} - 1$. Figure 5.11 shows example plots of these PCA$_{pcs}$ coefficients extracted from the same TMEP data as previously.

The final transform is named the pooled class-independent (PCA$_{pci}$) transform. It is class-independent due to it creating a single transformation matrix for all classes. It does this by pooling the data from each class using only instances self-aligned to their own templates. Figure 5.12 shows example plots of these PCA$_{pci}$ coefficients extracted from the same bioacoustic TMEP data as previously.

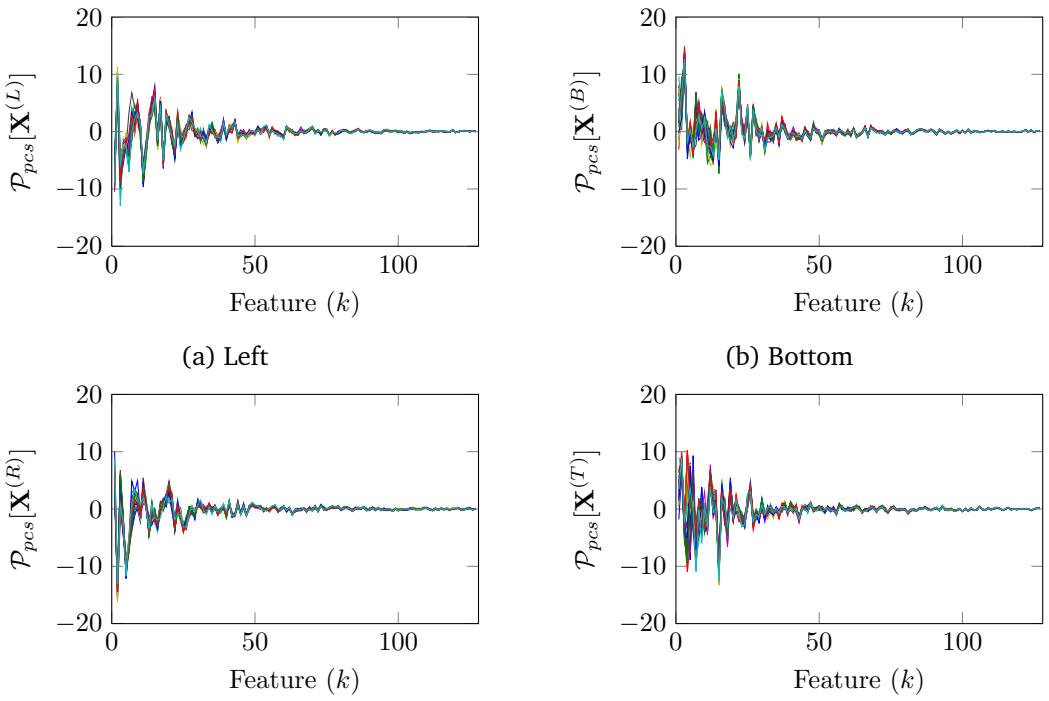
The PCA$_{ucs}$ and PCA$_{pcs}$ due to their class-specific nature show higher intra-class consistency. The transformation matrices have been created from the data and thus are optimal (in the MSE sense) for these data sets. Furthermore, the feature waveforms indicate that the principal subspaces are well separated, especially in the case of the PCA$_{ucs}$ transform. For this particular transform, these plots indicate that most, if not all, of the features would be useful for classification. Although, the PCA$_{pci}$ transform seems to indicate less feature



(a) Left (b) Bottom

(c) Right (d) Top

Figure 5.10: Example plots showing PCA_{ucs} coefficients for four TMEP actions from one subject. Each data set consists of 32 repetitions of each of the actions.



(a) Left (b) Bottom

(c) Right (d) Top

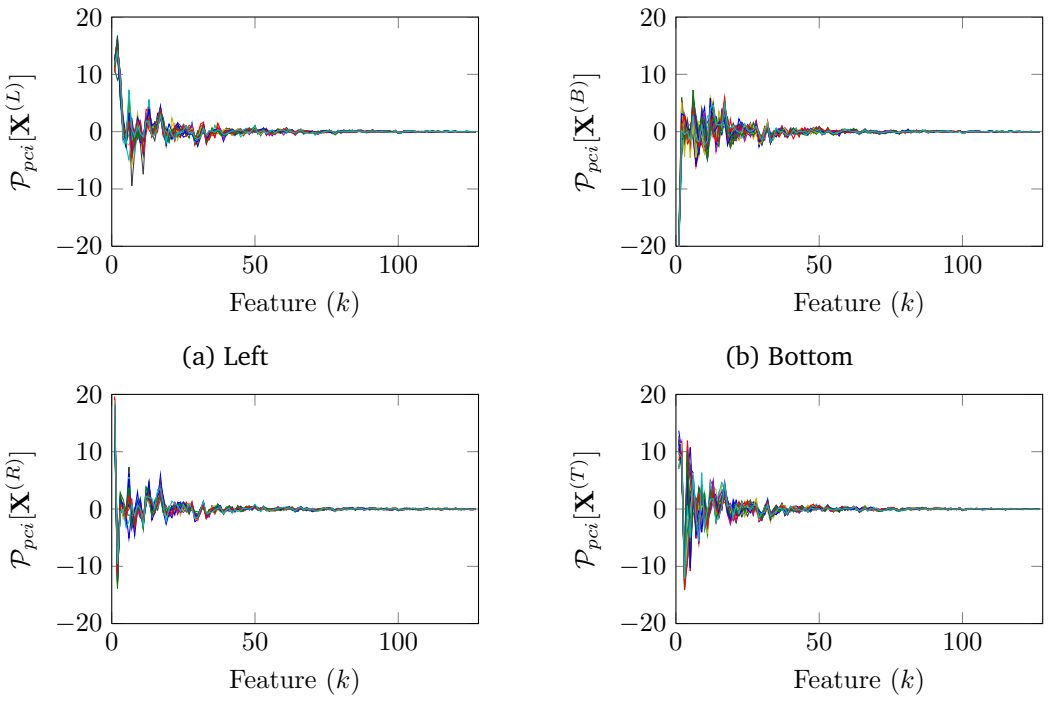
Figure 5.11: Example plots showing PCA$_{pcs}$ coefficients for four TMEP actions from one subject. Each data set consists of 32 repetitions of each of the actions.

separability and consistency between the classes, it is also a more general transform. This implies that it should be able to generalise better, especially if there are deviations of the test features from the training samples. It is still data-dependent and therefore optimised for a particular bioacoustic data-set but this is across all the classes, rather than being optimised for each unique class.

5.6 Feature selection

Feature selection, although an optional precursor to classification, has become tantamount to the learning and classifier itself. It can be defined as, given an initial set of candidate features, select a subset of these features that performs the best according to some evaluation criterion or classification system [163, 182]. When the candidate set of features is of high dimensionality, the use of feature selection methods becomes essential. A feature selection method ϕ can be defined as

$$\phi : \mathbf{S} \rightarrow \mathbf{S}^{*} \tag{5.24}$$



(a) Left (b) Bottom

(c) Right (d) Top

Figure 5.12: Example plots showing PCA_{pci} coefficients for four TMEP actions from one subject. Each data set consists of 32 repetitions of each of the actions.

where $\mathbf{S} \in \mathbb{R}^{N \times N_p}$ is the original feature data set and $\mathbf{S}^* \in \mathbb{R}^{D \times N_p}$ is the selected feature subset with $D < N$. In terms of supervised classification, the objective of a feature selection method is to improve classification performance while avoiding overfitting [183]. It provides a more efficient, faster and cost effective classification model. This implies that feature selection affects both the performance and parsimony associated with a given classification system. It can also lead to a deeper understanding of the underlying processes that generated the data, based on the particular features extracted [97].

Feature selection algorithms fall broadly into two main categories, namely, filter and wrapper models [182, 184]. A filter model evaluates features using general characteristics and associated information content of the training data. It does not incorporate any learning or classification algorithm during its evaluation. On the other hand, the wrapper model uses a predetermined learning algorithm and evaluates the features based on their predictive accuracy. This is usually estimated using an independent validation set through statistical resampling or cross-validation. There is a less widely applicable subgroup of feature selection methods, known as embedded models. These incorporate feature selection as part of the classifier model and thus are intrinsically linked and inseparable from the learning algorithm [184]. Table 5.1 highlights the main differences between filter and wrapper methods.

	filter method	wrapper method
Accuracy	Generally less accurate	Generally more accurate
Generality	Can be used with many different classifier models	Biased toward and thus specific to the classifier model
Execution speed	Fast	Slow
Subset size	Tend to be larger	Tend to be smaller
Ability to generalise	Cannot overfit as no classifier model specified	Overfitting can be reduced based on a resampling mechanism

Table 5.1: Comparing the performance of filter and wrapper models

The manner in which the feature selection algorithm searches through the candidate feature set, can be further categorised and falls broadly into two main groups. These are based on either feature weighting or subset evaluation strategies. They differ in whether they evaluate each feature individually or through selected feature subsets [182]. Feature weighting methods assign weights to each individual feature, with this indicative of its relevance to classification. This effectively ranks the features in order of this relevance and thus a subset of the top D-ranked features can be selected for classification. For example, features can be selected based on their weight value being greater than some predetermined threshold. The major drawback with this approach, is that it cannot account for redundancy among the features. Features that are highly correlated in terms of their relationship to the target concept cannot be ascertained due to the univariate nature of the evaluation strategy. The inclusion of redundant features can lead to an unnecessary increase in the feature subset size while potentially having a detrimental effect on performance due to increased noise in the feature subspace. This is summed up by the phrase 'the D best features are not necessarily the best D features' [185].

On the other hand, feature subset evaluation strategies require a search algorithm to explore candidate feature subsets, guided by some evaluation method [164]. The spectrum of search strategies is wide ranging and includes exhaustive, heuristic, sequential or random search algorithms [182, 185]. These vary in terms of the area and broadness of the feature space they explore, their time and space complexity and algorithmic intricacies. The main advantage of these feature subset methods is that they allow the multivariate correlations and thus the redundancy between the features to be taken into account. However, this is at the expense of much higher computational complexity compared to the feature weighting methods and are generally at least quadratic, if not exponential, with regards to the feature dimension [182]. If the data dimension is very large, these types of methods can be infeasible. Additionally, if this is combined with a small sample size, that is $N_p << N$, then multivariate feature densities cannot be adequately represented [163]. In such cases, it is probably more appropriate to use a univariate feature weighting

method regardless of its aforementioned limitations.

Hybrid methods have been recently introduced and attempt to utilise the benefits of both filter and wrapper methods [164]. Filter methods tend to be faster and have the ability to scale to larger data sets while wrapper methods tend to give superior performance (see Table 5.1). Thus hybridisation attempts to leverage these advantages by initially using a filter method to reduce the candidate feature step, followed by a wrapper to evaluate this smaller subset of features.

5.6.1 Sequential feature selection

In this work, many of the feature extraction techniques proposed require irrelevant features to be removed, using a feature selection technique prior to classification. These include the DFT, DCT and PCA transforms which generally output an initial feature set of high dimensionality. Due to the subject-specific nature of volitional bioacoustic signals, it is infeasible to expect a user to collect a large training set prior to real-time classification, implying that the feature selection task can be considered a high dimension and small sample size problem. A univariate sequential feature selection process is proposed in this work, based on feature weighting (pre-ranking) with its selection based on the following reasons

- Relevance of the features is assumed much more important than redundancy. If the final selected feature set is larger due to no redundancy checking, this should not prove a problem, as predictive accuracy is more important than computational cost
- It is much quicker to train. The proposed feature extraction transforms are of relatively high dimensionality and therefore subset evaluation methods would generally be intractable. Furthermore, this speed allows for comparison among many feature extraction and classification schemes without inhibitory computational overheads.
- The univariate weighting of the features implies no joint probability distributions need to be estimated among the features. When the sample size is small, reliable estimation of multivariate densities is unlikely.

A univariate feature weighting strategy involves a two stage process, initial feature ranking is followed by subset size selection based on this ranking. It therefore naturally lends itself to hybridisation, whereby, a filter model can be used to rank the features and a wrapper to select the final subset size. Such a strategy can be formulated as follows, given the candidate feature set, $\mathcal{D} = \{\mathbf{S}^{(1)}, \mathbf{S}^{(2)}, \ldots, \mathbf{S}^{(M)}\}$ with $\mathbf{S}^{(m)} \in \mathbb{R}^{N \times N_p^{(m)}}$, whereby a feature weighting function (Ω) assigns weights to each feature using

$$\Omega : \mathcal{D}_n \rightarrow \omega_n \tag{5.25}$$

where ω_n is a feature weight or score associated with feature n. The features can then be reordered according to these scores in order of most-to-least relevant, with this ranking based on the following

$$\mathcal{R}_{\omega_n \to j} \Rightarrow \begin{cases} \mathcal{R}_j > \mathcal{R}_n & \text{if } \omega_n > \omega_j, \\ \mathcal{R}_{j=1} & \text{if } \omega_n < \omega_j, \quad \forall n \neq j \\ \mathcal{R}_{j=N} & \text{if } \omega_n > \omega_j, \quad \forall n \neq j \end{cases} \tag{5.26}$$

By reordering the original data set according to this ranking, the top d features can be selected, that is $\breve{\mathcal{D}}_{1,\ldots,d} = (\mathbf{S}^{(1)}_{r,:}, \mathbf{S}^{(2)}_{r,:}, \ldots, \mathbf{S}^{(M)}_{r,:}\}$ with $r = \mathcal{R}_{1,2,\ldots,d}$ and $\mathbf{S}^{(m)}_{r,:} \in \mathbb{R}^{d \times N_p^{(m)}}$. Finally, a subset size function Ψ can score these incremental feature subset as follows

$$\Psi : \breve{\mathcal{D}}_{1,\ldots,d} \to \alpha_d \tag{5.27}$$

with the 'optimal subset' selected according to the one which gives the highest α, that is

$$z = \arg\max_d(\alpha_d), \quad \forall n \tag{5.28}$$

Therefore, the index of the selected feature subset is $\mathcal{R}_{1,2,\ldots,z}$.

5.6.2 Feature weighting filters

Firstly, it is useful to define a simple feature weighting which is actually independent of $\mathcal{D}$. It is based on the assumption that the feature coefficients with a lower index are more relevant than those with a higher index. This assumption can be deemed valid for feature extraction techniques based on frequency components for example, as the lower frequency coefficients can be considered more informative than higher frequencies, which are usually associated with noise. To fit with Equation 5.26, this simple weighting can be defined as

$$\omega_n = -n \tag{5.29}$$

Although generally true on a global scale, this can be considered quite a naive concept, as at the local level features with a higher index may actually be more relevant. Thus it is more suitable to weight the features based on the data itself. An appropriate concept is therefore to rank the features based on their magnitude, with the assumption being that more descriptive and thus discriminative features will have larger values. This can be done by finding the mean magnitude of the features within each class using

$$\omega_n^{(m)} = \frac{1}{N_p^{(m)}} \sum_{i=1}^{N_p^{(m)}} |\mathbf{S}^{(m)}_{n,i}| \tag{5.30}$$

with this used to initially rank the features within each class and is followed by the calculation of the mean of these class-specific rankings. This final feature weighting is calculated using the following

$$\omega_n = -\frac{1}{M}\sum_{m=1}^{M} \mathcal{R}_{\omega_n^{(m)} \to k} \tag{5.31}$$

and has been named the mean rank of the mean magnitude (mRmM). It can be considered a non-parametric measure which is unaffected by differences in the class distributions such as changes in the range of the feature values. This concept can be taken a step further, by initially ranking each feature instance based on their magnitude and then taking the average across these ranks. This is calculated from

$$\omega_n^{(m)} = \frac{1}{N_p^{(m)}}\sum_{i=1}^{N_p^{(m)}} \mathcal{R}_{|\mathbf{S}_{n,i}^{(m)}| \to k} \tag{5.32}$$

with each class-specific weighting combined as before using Equations 5.26 and 5.31. This has been coined the mean rank of the mean rank magnitude (mRmRM) due to the additional ranking process. The extra ranking potentially inhibits the effects of intra-class outliers that could affect the estimation of the mean magnitude.

Another important attribute associated with 'good' discriminatory features is that they should have low intra-class variance indicating them as stable representations of the features. 'Bad' features will generally be unstable and thus exhibit higher variance. The class-specific feature weighting associated with this ranking method can therefore be computed as follows

$$\omega_n^{(m)} = \frac{1}{N_p^{(m)} - 1}\sum_{i=1}^{N_p^{(m)}} \left(\mathbf{S}_{n,i}^{(m)} - \bar{\mathbf{s}}_n^{(m)}\right)^2 \tag{5.33}$$

and again these class-specific weights are combined using Equations 5.26 and 5.31. The negative sign is due to smaller variances now being more beneficial. This scheme has been named the mean rank of the variance (mRV) associated with the features.

The concepts of larger feature values and low intra-class variance leading to better feature choices can also be combined using a single metric. This can be achieved using the F-score, which is a measure of the discrimination between two sets of real numbers. It can be considered a univariate measure which scores each feature based on this discrimination value, calculated from the training instances [184]. There are two ways in which it can be cast into a multi-class environment, namely, extending the original formulation to deal with multiple sets or using pairwise interactions and averaging over these [186]. For clarity, these are named the multi-class F-score (mFS) and pairwise F-score (pFS) respectively. The mFS measure is a generalisation of the two-class situation and is defined for sample

n as

$$\omega_n = \frac{\sum_{m=1}^{M}(\bar{\mathbf{s}}_n^{(m)} - \bar{\mathbf{s}}_n)^2}{\sum_{m=1}^{M}\left(\frac{1}{N_p^{(m)}-1}\sum_{i=1}^{N_p^{(m)}}(\mathbf{S}_{n,i}^{(m)} - \bar{\mathbf{s}}_n^{(m)})^2\right)} \tag{5.34}$$

where $N_p^{(m)}$ is the number of feature patterns associated with class m, $\mathbf{S}^{(m)} \in \mathbb{R}^{N \times N_p^{(m)}}$ is the feature data matrix associated with class m, $\bar{\mathbf{s}}$ is the average of all the feature data sets and $\bar{\mathbf{s}}^{(m)}$ is the average of the feature data set associated with class m [186]. It is similar in formulation to Fisher's discriminant analysis, whereby the features are ranked based on the ratio of the between-class scatter to the within-class scatter [158]. Similarly, pFS is given by

$$\omega_n = \sum_{a=1}^{M-1}\sum_{b=a+1}^{M}\left(\frac{(\bar{\mathbf{s}}_n^{(a)} - \bar{\mathbf{s}}_n)^2 + \bar{\mathbf{s}}_n^{(b)} - (\bar{\mathbf{s}}_n)^2}{\frac{1}{N_p^{(a)}-1}\sum_{i=1}^{N_p^{(a)}}(\mathbf{S}_{n,i}^{(a)} - \bar{\mathbf{s}}_n^{(a)})^2 + \frac{1}{N_p^{(b)}-1}\sum_{i=1}^{N_p^{(b)}}(\mathbf{S}_{n,i}^{(b)} - \bar{\mathbf{s}}_n^{(b)})^2}\right) \tag{5.35}$$

with all the variables having the same meaning as Equation 5.34 with (a) and (b) implying association to class a and b respectively. It should be noted when using the F-score that no intrinsic ranking is required, as the weights are automatically normalised due to the division operator.

5.6.3 Subset size selection

As stated earlier the use of wrapper methods can potentially provide superior performance over filter methods. If the learning algorithm is known *a priori* it makes sense to implement it during final subset size selection. A wrapper evaluates a feature subset based on the predictive accuracy associated with the specified classifier, as estimated from the training set [184].

The subset size function Ψ takes the reordered feature subset as input. Starting from the empty feature set, it then incrementally selects the top d features and evaluates their performance using the CV procedure (see Section 5.3.2). It does this for all possible feature subset sizes and then selects the subset which gives the maximum CV accuracy based on the proposed learning algorithm. When this is combined with a filter feature ranking function, Ω, the overall SFS algorithm can be considered a hybrid method, denoted by the acronym SFS_{hyb}.

To remove the dependency on the learning algorithm and reduce the learning time, the subset size can also be selected using a simple filter metric. This is based on the heuristic that useful features are those whose filter weight ω_n increases the average weighting across all features. This is equivalent to selecting the subset of features whose feature weight is larger than the mean feature weight across all the features. The subset size se-

lection function is now dependent on just the feature weights and not the data itself, such that

$$\Psi : \omega_n \rightarrow \alpha_n \tag{5.36}$$

This subset selection method only makes sense for filter measures which give continuous valued outputs, such as the F-score measures (mFS and pFS). Therefore, it doesn't make sense using this method with the rank-based feature weighting functions. As this strategy only requires a filter and no wrapper method during the feature selection process, it can be denoted by SFS_{filt}.

5.7 Increasing stability during feature selection

As mentioned briefly earlier, when there is a large number of features and only a small number of instances to train on, that is $z >> N_p$, the selection of an appropriate feature subset can be extremely difficult [163, 187]. Many of these features will be redundant or irrelevant and the inclusion of many will have a detrimental effect on the classification. This is partly due to the training set being not fully representative of the feature distributions and thus the feature selection process will be bias towards this small set and therefore not generalise well to new data. Additionally, when the sample size of the training set is small, the stability of the chosen feature set can be severely deteriorated. The stability of a feature is defined as the robustness of the feature preferences it produces to differences in training sets drawn from the same generating distribution [164]. Thus if the generating distribution is not well represented due to a small training sample, the selection of features based on one data set is likely to vary sample-to-sample. This is closely linked to the concept of overfitting, as the selected 'unstable' features perform well during training, but when unseen instances drawn from the generating distribution are classified using the same features, the same feature subset will generally perform with a much lower classification performance. In this section, some measures are introduced to increase the stability of the chosen feature subset.

5.7.1 Weighting envelope

During naive feature selection, the features are initially ranked and then the features added one-by-one according to this ranking. As each new feature is added, the extended feature subset is evaluated according to some criterion. In general, there will be a point where this criterion will yield a maximum value, with the criterion increasing up until this point and then decreasing after it. This behaviour can be modelled using the following

equation

$$\mathcal{F}_w[d] = 1 - \exp\left(\frac{d\log(0.01)}{p}\right) - \exp\left(\frac{(D-d-1)\log(0.01)}{D-p-1}\right) \tag{5.37}$$

where $d \in [1, 2, \ldots, N]$ is the feature subset size, D is the total number of features and p is the 'operating point'. If this 'operating point' is known a priori, then such a function can additionally be used to bias the feature subset size towards this point. It should be noted that this function is not relevant during the feature ranking stage, as generally the mapping from initial feature order to the ranked order will be a highly discontinuous function. By using this function to weight the criterion during feature subset size evaluation, it restricts the feature subset size without actually constraining it with hard bounds. Fig. 5.13 shows the feature weighting function against the feature index, for different operating points. As

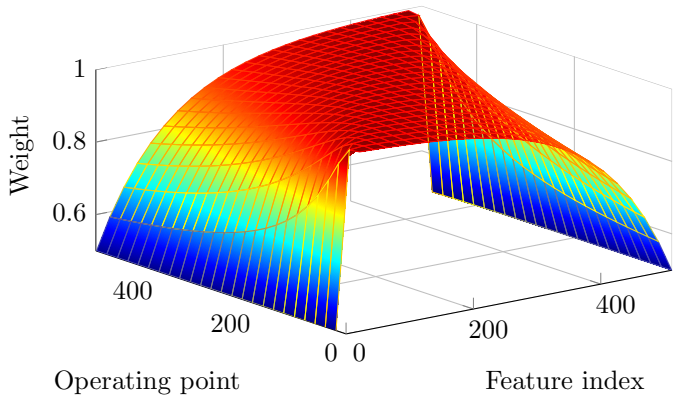


Figure 5.13: Feature weighting envelope for different operating points

can be seen the 'rise-time' and 'fall-time' vary according to the operating point and follow an inverse relationship to one another. This prevents very small or very large feature subsets from being selected regardless of the 'operating point'. Furthermore the region around the 'operating point' is relatively flat, especially at the midpoint, thus giving the feature selection algorithm the flexibility of choosing a feature subset size with higher probability in this flat region.

5.7.2 Weighted sequential feature selection

The weighted sequential feature selection (WSFS) has been recently introduced to select the most relevant features and effectively reduces the number of unstable features included in the 'optimal' feature subset [97]. This is based on removing those features which have little or no contribution to the classification across different training subsets. The idea behind WSFS is to utilise a sequential or univariate feature selection strategy in combination with a resampling procedure, thus allowing the contributions from each fea-

ture to be computed and the stability of each feature estimated. It attempts to overcome the limitations associated with sequential and univariate feature selection processes by selecting the most relevant features independent of changes in the training sets. Therefore, this should increase the robustness of the selected feature set, at the expense of increased computational complexity.

Given the data set $\mathcal{D}$, containing all features and training instances associated with the M classes, it can be divided into K subsets through K-fold stratified cross-validation. Each of these subsets can be labelled $\mathcal{V}^{(1)}, \mathcal{V}^{(2)}, \ldots, \mathcal{V}^{(K)}$ with subset $\mathcal{V}^{(k)}$ consisting of $M \times N_K$ instances, where $N_K = \lfloor N_p/K \rfloor$ and $N_p = N_p^{(m)}, \forall m$. During the cross-validation process, at each fold iteration, subset $\mathcal{V}^{(k)}$ is used as a validation set while the remaining $K-1$ subsets form a single disjoint training set $\mathcal{T}^k = \bigcup_{r=1}^{K} \mathcal{V}^{\{r\}}$ where $r \neq k$. As such, each $\mathcal{T}^{\{k\}}$ contains overlapping subsets with N_K distinct instances. It is these distinct regions which play the most importance role during WSFS, as this is what simulates the differences across training sets drawn from the same generating distribution.

Using the univariate feature selection, Equations 5.25 - 5.28, with the filter weighting function, $\Omega : \mathcal{T}_n^k \rightarrow \omega_n$ (Equation 5.25) and the subset size function, $\Psi : \breve{\mathcal{V}}_{1,\ldots,n}^{(k)} \rightarrow \alpha_n$ (Equation 5.27) allows a feature subset associated with fold k to be selected. Furthermore, this cross-validation procedure can be computed q times using a different random permutation of $\mathcal{D}$, that is, Q repeated K-fold cross-validation, such that $\mathcal{D}^{(q\}} \subset \mathcal{D}$ are created, with $q = [1, 2, \ldots, QK]$. Assuming that Ω is dependent upon $\mathcal{T}$, the proposed procedure should create Q inconsistent feature subsets, which should vary in both size and content. It is these inconsistencies which imply the classification model is biased toward the training set and therefore will give unreliable predictions subset-to-subset. To address this problem, a contribution weight measure ρ_d associated with each feature index d, is defined and determines the consistent features independent of the instances contained within each $\mathcal{D}^{(q\}}$. This weight measure estimates the expected occurrence of each feature within different training sample subsets and is formulated as

$$\rho_d = \sum_{q=1}^{QK} \delta_{d,q} \tag{5.38}$$

where

$$\delta_{d,q} = \begin{cases} 1 & \text{if } d \in \mathcal{R}_{1,2,\ldots,z}^{(q\}} \\ 0 & \text{otherwise} \end{cases} \tag{5.39}$$

$\mathcal{R}_{1,2,\ldots,z}^{(q\}}$ is the selected feature set as found using $\mathcal{T}^{(q\}}$ and $\mathcal{V}^{(q\}}$. To obtain the robust subset of features, the features are grouped into Q subsets $\mathcal{H}^{(q\}}$ based on their contribution weights, according to

$$\mathcal{H}^{(q\}} = (s_d | \rho_d > q\} \qquad \forall d \tag{5.40}$$

where $\mathcal{H}^{\{q\}}$ is the qth feature subset which contains all the features which have contributed at least q times during the initial feature selection process. These new feature subsets can then be evaluated using $\Psi : \mathcal{H}^{\{q\}} \rightarrow \alpha_q$, with the 'optimal subset' $\mathcal{H}^{\{z^*\}}$ selected according to the one which gives the highest score according to $z^* = \arg\max_q(\alpha_q)$.

5.8 Class-specific classifier models

Due to the nature of the proposed classification strategy and the required template alignment stage (highlighted in Fig. 5.5), there is a restriction on the type of classifier that can be utilised within this framework. The template alignment process creates class-specific test instances which implies only class-specific classification models can be utilised. These type of models are based on generative or template matching based classifiers [188]. This rules out discriminative and discriminant based classifier models, as these map the decision boundary directly and as such require a single instance as input. Therefore, many standard classifiers such as ANNs, multi-class SVMs and logistic regression models cannot be utilised [160]. Generative classifiers model the class-conditional densities, that is $p(\mathbf{x}/c_m)$ and then use decision theory to assign a class label. Although this approach is sometimes seen as computationally excessive, it naturally allows for incorporation of risk into the model in the form of a loss matrix (see Section 5.9). Furthermore, this loss matrix can be modified after training, with no further retraining required.

Based on this template model, one of the simplest classifiers that can be implemented is a minimum distance classifier, also known as the nearest neighbour classifier [158]. Assuming the signals to be points in N-dimensional Euclidean space, the distance between an unknown vector $\mathbf{s}$ and the template vector $\bar{\mathbf{x}}^{(m)}, m = 1, 2, \ldots, M$ can be computed using the Euclidean distance, given by

$$d(\mathbf{s}, \bar{\mathbf{x}}^{(m)}) = \|\mathbf{s} - \bar{\mathbf{x}}^{(m)}\|^2 \tag{5.41}$$

This is considered the straight line distance between two N-dimensional points. For reference, this classifier will be denoted the eDC model in later sections. As classifiers generally make decisions through maximisation, the Euclidean classifier class-specific discriminant values can be defined as

$$\mathcal{C}^{(m)}(\mathbf{s}) = -d(\mathbf{s}, \bar{\mathbf{x}}^{(m)}) \tag{5.42}$$

The applicability of distance-based classifier highlights that the proposed model need only be class-specific in nature with generative classifiers considered a subset of this model type. It should be noted that other vector norms can be used, such as the those based on the L_1-norm also known as the Manhattan distance classifier.

The most widely used generative classifier is the naïve Bayes model [189]. It has

various guises based on assumptions regarding the form of the densities. Although in reality, many of the assumptions it makes are violated, in practise it is still a very effective model [190]. Other class-specific classification methods explored include the decision fusion classifier and matched filter with their applicability generally restricted to the time domain [24]. These class-specfic classifiers are introduced in the following sections.

5.8.1 Gaussian Bayes

One of the fundamental statistical approaches in the pattern classification field is Bayesian decision theory [158]. A Bayesian classifier quantifies the trade-off between various classification decisions by modelling the feature space using probability distributions. Bayes' theorem can be used to obtain a posterior probability from a prior probability by incorporating evidence provided by the observed data. If the prior probability of class m, assuming the data is yet to be observed is $p(c_m)$, the evidence of the observed data is $p(\mathbf{s})$ and the likelihood of the data belonging to class c_m is $p(\mathbf{s}|c_m)$. The posterior probability can then be defined as

$$p(c_m|\mathbf{s}) = \frac{p(\mathbf{s}|c_m)p(c_m)}{p(\mathbf{s})} \tag{5.43}$$

The posterior probability is the probability of a class given the evidence, with the aim of a Bayesian classifier to assign a feature vector to the class with the highest posterior probability. When the prior probability is assumed uniform, it is known as a maximum likelihood classifier, as only the likelihood term bears any relevance towards discrimination.

The naïve Bayes' classifier is an effective learning algorithm and is one of the most widely used in the machine learning community [189–191]. It assumes conditional independence between the data, features or attributes. Although rarely true in the real world, this assumption in many cases doesn't have a significantly detrimental effect on the classification performance, even when this assumption is heavily violated. Part of its effectiveness can be attributed to the reduction in the number of classifier parameters that need to be estimated during training, namely the conditional probability distributions. Furthermore, if the data dependencies are distributed evenly among the classes or cancel each other out, then the naïve Bayes classifier can still be considered in some sense optimal [189]. Given the naïve assumption, the likelihood term can now be calculated by

$$p(\mathbf{s}|c_m) = p(s_1, s_2, \ldots, s_D|c_m) = \prod_{k=1}^{D} p(s_k|c_m) \tag{5.44}$$

Substituting this back into Equation 5.43 and taking the $\log$ yields the following class-

specific function

$$\mathcal{C}^{(m)}(\mathbf{s}) = \log\left(p(\mathbf{s}|c_m)p(c_m)\right) = \sum_{k=1}^{D} \log\left(p(s_k|c_m)\right) + \log\left(p(c_m)\right) \tag{5.45}$$

where the evidence term ($p(\mathbf{s})$) has been dropped as it is class independent and therefore only acts as a normalising factor.

A standard assumption that can be made is that the features are distributed normally. Although this may not be exactly true in reality, especially with small samples, it provides a computationally tractable solution. The alternative would be to estimate the feature distributions non-parametrically from the data, requiring a Parzen-window density estimator for instance, which based on a small sample can ultimately overfit the classification model and requires tuning of an additional smoothing parameters [158]. Furthermore, the central limit theorem states that as more independent and identically distributed features are available the more normally the sum of the data will be distributed.

The multivariate Gaussian feature model, conditioned on class m, is given as

$$p(\mathbf{s}|c_m) = \frac{1}{(2\pi)^{D/2}|\mathbf{\Sigma}^{(m)}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{s}-\mu^{(m)})^T\mathbf{\Sigma}^{-1}(\mathbf{s}-\mu^{(m)})\right) \tag{5.46}$$

where $\mu^{(m)}$ is the mean vector of the entire feature set for class m and $\mathbf{\Sigma}$ is the covariance matrix. When $D > N_p^{(m)}$ there are associated problems with taking the inverse of the covariance matrix as it won't be full rank and is indicative of the curse of dimensionality [160]. To alleviate this constraint, it has been proposed that a single covariance matrix be estimated for all the classes by pooling across the training sets [158, 165]. This implies that the number of features that are used to estimate the covariance matrix is only limited by $D < \sum_{m=1}^{M} N_p^{(m)}$, however the quadratic decision boundaries are lost as the discriminant terms are now linear. Alternatively, constraints can be placed on the covariance matrix itself, namely, it can be forced to be diagonal and thus the features assumed to be uncorrelated. The diagonal covariance matrix is defined as

$$\mathbf{\Sigma}_{ij}^{(m)} = \begin{cases} \sigma_i^{(m)} & \text{if } i = j \\ 0 & \text{otherwise} \end{cases} \tag{5.47}$$

and implies that the principle axis of the hyper-ellipsoid decision boundaries are aligned with the global feature frame. Furthermore, this combined with the feature independence assumption suggests a univariate feature model. The univariate feature model for class m is given by

$$p(s_k|c_m) = \frac{1}{\sqrt{2\pi\sigma_k^{(m)}}} \exp\left(-\frac{s_k - \mu_k^{(m)}}{2\sigma_k^{(m)}}\right) \tag{5.48}$$

where $\mu_k^{(m)}$ and $\sigma_k^{(m)}$ are the mean and standard deviation associated with the kth feature from class m. By using the univariate model, any problems associated with performing large and redundant matrix calculations, specifically, taking the inverse of the covariance matrix are mitigated.

As $\mu_k^{(m)}$ and $\sigma_k^{(m)}$, are really only measures associated with the central tendency and spread of the normal distribution, they can instead be estimated using the concept of robust statistics. These include the median or trimmed mean for estimating the central tendency and the median absolute deviation (MAD) or interquartile range (IQR) for estimating the spread. The median is a special case of the trimmed mean and is preferred due to a higher breakdown point and has no additional free parameters to tune. If outliers are present in the training sample, the use of robust statistics can potentially make the estimation of $\mu_k^{(m)}$ and $\sigma_k^{(m)}$ more reliable. When the training set is small, the normality of the features is reduced and thus robust measures could also provide better estimation of these statistics. To differentiate and clarify the various model assumptions, the following notation can be defined

sMGC Multivariate Gaussian classifier using standard statistics. Naïve Bayes classifier conditioned on multivariate Gaussian models estimated using standard statistics.

sUGC Univariate Gaussian classifier using standard statistics. Naïve Bayes classifier conditioned on univariate Gaussian models estimated using standard statistics.

rUGC$_{\text{mad}}$ Univariate Gaussian classifier using robust statistics. Naïve Bayes classifier conditioned on univariate Gaussian models estimated using robust statistics of the median and MAD.

rUGC$_{\text{iqr}}$ Univariate Gaussian classifier using robust statistics. Naïve Bayes classifier conditioned on univariate Gaussian models estimated using robust statistics of the median and IQR.

5.8.2 Decision fusion

The decision fusion classifier (DFC) is based on a two stage classification process and has been previously introduced [24]. It is reformulated here in a concise, modified and general mathematical manner. Initially each sample of a discrete signal vector are classified independently, using a minimum M-class discriminant. This is based on a pointwise distance function between itself and the associated samples of M-template vectors. A discrete likelihood matrix is constructed based on the probability of classifying a sample from class-c as class-m using a training data-set. Each sample is ranked based on its average classification accuracy, using the pointwise distance measure and the M-templates. An unseen instance is then initially re-ranked and classified using the same pointwise distance function and

M-templates. The samples with the highest ranks can then be fused into a single decision fusion vector for each class, based on the probability of a sample, initially classified as c, actually belonging to class-m. The final discriminant value is then based on the fusion of these decision fusion vector samples under naïve assumptions.

A pointwise distance function is defined as

$$d : (x_n, y_n) \mapsto \mathbb{R} \tag{5.49}$$

with an example of a such a measure being the Euclidean norm nearest-neighbour distance function, given by

$$d(x_n, y_n) = 0.5y_n^2 - x_n y_n \tag{5.50}$$

This is a quadratic function, with its minima at x_n. Therefore if $y_n \in \mathbf{y}$, the instance with the closest value to x_n will have the minimum distance of the set $\mathbf{y}$.

A set of training vectors $\mathbf{X}^{(m)}$ associated with class m is used to make M-templates $\bar{x}_n^{(c)}$. Using Equation 5.50, the minimum distance between each point and the corresponding sample from one of the M-templates can be calculated as follows

$$\omega_{n,j}^{(c)} = \arg\min_m \left(d(\hat{\mathbf{X}}_{n,j}^{(c)}, \bar{x}_n^{(m)}) \right) \tag{5.51}$$

$\forall n = 1, 2, \ldots, N$. Furthermore $[c, m] = 1, 2, \ldots, M$ and $j = 1, 2, \ldots, N_p^{(c)}$. The $\hat{}$ indicates that each vector $\mathbf{x}_j^{(c)}$ has been aligned with $\bar{\mathbf{x}}^{(m)}$ using cross-correlation alignment as outlined in Section 4.4.1. The log-likelihood that a signal value at sample n, associated with class c, is closest to the corresponding point of one of the M-templates is found using

$$\mathcal{L}_{n,m}^{(c)} = \log \left(\frac{1}{N_p^{(c)} + M} \left(\sum_{j=1}^{N_p^{(c)}} \delta_{n,j}^{(c)} + 1 \right) \right) \tag{5.52}$$

where $\delta_{n,j}^{(c)}$ represents the kronecker delta function at sample n associated with a vector instance j from class c and is given by

$$\delta_{n,j}^{(c)} = \begin{cases} 1 & \text{if } \omega_{n,j}^{(c)} = m \\ 0 & \text{otherwise} \end{cases} \tag{5.53}$$

A Laplace correction term is included in Equation 5.52 to prevent zero probabilities and thus infinite log likelihoods from occurring. The samples can then be ranked based on the expected likelihood of classifying a sample from class-m as class-m, also known as the average correct classification accuracy for each sample across the M-classes. This ranking

value is then computed as

$$\alpha_n = \mathbb{E}\left(\mathcal{L}_{n,m}^{(c=m)}\right) = \frac{1}{M}\sum_{m=1}^{M}\mathcal{L}_{n,m}^{(c=m)} \tag{5.54}$$

with the samples then ranked from high-low, in a ordinal fashion in a similar manner as Equation 5.26. Based on this ranking, a vector can be reordered and the top D most influential samples extracted from it, as dictated by this ranking order that is $\hat{\mathbf{x}}^* = \hat{\mathbf{x}}[r_{1,2,\ldots,D}]$ with $D \leq N$.

A subset of the most influential samples is extracted from an unseen instance $\hat{\mathbf{s}}$ and is initially classified to the M-templates based on the minimum distance, computed using Equation 5.50 and is given as

$$\omega_n^* = \arg\min_m \left(d(\hat{\mathbf{s}}_n^*, \bar{x}_n^{*(m)})\right) \tag{5.55}$$

The log-likelihood that this vector belongs to each of the M-classes, based on the fusion of the decisions at each sample is given by the following class-specific discriminant value

$$\mathcal{C}^{(m)}(\mathbf{s}) = \sum_{n=1}^{N}\mathcal{L}_{n,m}^{(c=\omega_n^*)} + p(m) \tag{5.56}$$

where $p(m)$ is the prior probability of class m. Therefore, the larger $\mathcal{C}^{(m)}(\mathbf{s})$ is, the more likely $\mathbf{s}^*$ belongs to class m.

5.8.3 Matched filter

Matched filters are designed to locate a signal in additive stochastic noise under the assumption that the noise is stationary and the signal is known [192]. It is designed to maximise the SNR by correlating an unknown signal with a known signal and is considered linearly optimal [96]. If it is assumed that each template is the known signal, then the matched filter is given by the time-reversed waveform of each of these templates, with the conventional output of the matched filter process given by the convolution of this signal and the unknown test signal. Furthermore, if the unknown signal is already aligned with the template, then the two signals will already be in position of maximum correlation. In this case, the maximum output of the matched filter will be when $n = N$. Therefore an unknown signal associated with a particular class, should show higher correlation to a template created from signals from the same class, then with alternative classes. Therefore, using Equations 4.7 and 4.8 from the previous chapter, the class-specific discriminant

values for the matched filter classifier (MFC) are given by

$$\mathcal{C}^{(m)}(\mathbf{s}) = (s \star \bar{x}^{(m)})_p \tag{5.57}$$

where p is the position of maximum cross-correlation and $\mathbf{s}$ is the unknown signal. If cross-correlation alignment between the unknown signal and the M-templates has already been implemented, then the MFC is an extremely trivial extension requiring little extra computation.

5.9 Decision theory and expected loss

In many classification applications, the consequences of misclassifying instances from a particular class may be much more detrimental then misclassifying instances from the alternative classes. The converse is also true, whereby misclassifying to a particular class when its true nature is different maybe equally reckless. This risk can be formalised through the use of a risk vector and the introduction of a loss (or cost) matrix [158, 160]. If it is assumed that the output of the classifier $\mathcal{C}^{(m)}(\mathbf{s})$ reflects the posterior probability $p(c_m/\mathbf{s})$, then the uncertainty or expected loss in the true class can be expressed as

$$\mathbf{r}_c = \sum_{m=1}^{M} (\mathbf{L}_{c,m}\mathcal{C}^{(m)}(\mathbf{s})) \tag{5.58}$$

where $\mathbf{r} \in \mathbb{R}^M$ is a vector conveying the risk associated with classifying to class c. Therefore, the optimal solution for classification is based on minimisation of this risk and is given by

$$\omega_c = \arg\min_c(\mathbf{r}_c) \tag{5.59}$$

In this work, $\mathbf{L} \in \mathbb{R}^{M\times M}$ which assumes the number of decision outputs is equal to the number of classes and reflects the cost of classifying class m as class c. The most common loss matrix is the symmetrical zero-one loss and is given by

$$\mathbf{L}_{m,c} = \begin{cases} 0 & \text{if } m = c \\ 1 & \text{if } m \neq c \end{cases} \tag{5.60}$$

This is equivalent to minimising the misclassification error, as all of the errors are equally costed and as such, is identical to maximising $\omega_c = \arg\max_m(\mathcal{C}^{(m)}(\mathbf{s}))$. Risk is therefore portrayed by modifying this loss matrix, for example, by making $\mathbf{L}_{1,2} >> 1$ whereby the risk associated with classifying class 1 as 2 would now be considered much higher and therefore has a higher chance of being avoided.

5.10 Validation of proposed methods on TMEP data

This section deals with the comparison of various strategies for decoding volitional bioacoustic signals with application to assistive HMI systems. These decoding strategies are based on various combinations of the proposed feature extraction, feature selection and class-specific classification models discussed previously. Validation is based on single channel TMEP data sets as described in Section 3.5 for four-class ($M = 4$) volitional tongue-movement discrimination. Data from a total of fifteen subjects was therefore utilised during this evaluation. TMEP events are detected in the raw bioacoustic data sets, using the detection algorithm described in Section 4.1. Although all thresholds can be optimised for each subject individually, only the the lower threshold T_L remained subject-specific during this activity detection. The upper threshold and duration window are effectively ignored by setting them to $T_U = \inf$ and $D_U = \inf$, respectively. During the off-line validation process it is assumed that only events associated with intentional tongue movements are superimposed onto each bioacoustic signal. The lower duration window is set to $D_L = 0.1$sec to prevent detection of shorter unintentional signals which on occasion can be superimposed onto the signals. The lower threshold T_L is determined from the maximum average energies of the STE contour associated with each subjects data set. That is, T_L is given by

$$T_L = \frac{1}{2}\min\left(\frac{1}{M}\sum_{m=1}^{M}\frac{1}{N_p^{\{m\}}}\sum_{i=1}^{N_p^{\{m\}}}E_t^{\{m,i\}}\right) \tag{5.61}$$

where $E_t^{\{m,i\}}$ is the maximum energy of the ith signal from class m. After detection of TMEP events, the signals are segmented and shifted within a finite length window of 0.256 seconds (equivalent to 512 samples at 2 kHZ) using equations 4.2 - 4.4.

A 10-fold stratified CV procedure was used to obtain confusion matrices for either twenty or a hundred randomised trials, with the number of trials dependant on the computational load required for specific tests. For each subject, the same data partitions were used at each trial iteration implying the same training and data-sets were used by the proposed discrimination strategies. The CV procedure allows for a confusion matrix to be obtained for each trial, with each data instance used only once during testing within a single CV trial. Conversely, they are used multiple times for training with this dependent on the training subset size. At each trial, performance measures such as classification accuracy (Equation 5.1) can then be extracted from these confusion matrices. Across the trials, the average and spread of these measures is found across all trials, using for instance, the mean and standard deviation (SD). Other measures that can be used include the median and IQR if the distribution across trials was found to be non-Gaussian. For the TMEP data, the distribution across trials was for the most part found to be normally distributed based

on a Kolgomorov-Smirnov test and as such the mean and SD are used throughout this analysis. Furthermore, they are felt to provide more meaningful interpretation than the corresponding median and IQR statistics.

As mentioned previously, the template alignment and signal standardisation process proceeding feature extraction are important for effective TMEP classification. Each segment is therefore standardised using the Z-score given by Equation 4.21. Within each test fold, the robust template creation process is utilised to make exemplars of each of the M training sets. Each test and training instance is aligned to each of the templates, creating M shifted replicas of each signal instance. During training, the instances from class-m aligned to template-m, are used to estimate the parameters of the class-specific classifier models. At the prediction phase, the unlabelled test instance aligned to class-m are classified using the corresponding classification model.

5.10.1 Bioacoustic discrimination strategies

Table 5.2 highlights the various combinations of feature extraction transforms, ranking filters (where applicable) and classification algorithms compared during offline decoding of inter-action TMEP signals. The eDC, MFC and DFC classification algorithms are all based on template matching or correlation and as such, only really make sense in the time domain. For AR features, the model order is specified prior to feature extraction and therefore can be selected based on the chosen classification algorithm among other factors. This implies both, multivariate as well as univariate Gaussian classifiers, can be implemented in conjunction with the AR features without encountering problems associated with estimation of the covariance matrix [165]. This is under the assumption that the number of AR features required to adequately represent the TMEP signals is less than the training set size. Furthermore, specifying too large a model order leads to a a set of features which would not generalise well anyway [164].

As the DFT, DCT and PCA transforms all have a larger number of features initially, this prohibits the estimation of a nonsingular inverse covariance matrix. By assuming univariate conditions, that is, a diagonal covariance matrix, it allows the inverse to be found regardless of the number of features selected. For feature extraction methods which are used in conjunction with a feature selection method, the univariate conditions have been assumed. Furthermore, the effect of using robust statistics (median/IQR/MAD instead of mean/SD) for estimation of the univariate Gaussian classifier have only been tested in conjunction with the DCT features. This was due to the rUGC performing less effectively than sUGC in the case of DCT features (see Appendix B.3). Initial validation is implemented and tested across all feature subset sizes allowing comparisons between just classifiers and feature extraction methods to be drawn. Thus initial comparisons across classification

[1] based on a training set size of 32

Feature extraction	Initial Nr. features	Ranking filter	Classification algorithm
Time domain	512	n/a	eDC MFC DFC
$AR_{\{yule,burg\}}$	1-30	n/a	sUGC sMGC
$DFT_{\{mag,real,imag,phase\}}$	256	simple mRmM mRmRM mRV mFS pFS	sUGC $rUGC_{iqr}$ (DCT only) $rUGC_{mad}$ (DCT only)
DCT	512		
$PCA_{\{pcs,ucs,pci\}}$	$\{127,31,127\}$[1]		

Table 5.2: Discrimination strategies employed for decoding volitional bioacoustic activity, specifically TMEP signals, in the initial validation

strategies is the best achievable, independent of the feature selection algorithm.

5.10.2 A note on statistical comparisons

The statistical analysis performed in this work is based mainly on the procedures outlined in Demsar [193]. A general discussion is described in this section for the sake of clarity. Although classification accuracy and/or error rates do give an indication of overall classification performance, this is under the assumption that the data-sets being averaged over are commeasurable and the effect of outliers can be ignored. Take for example the situation where a classifier performs perfectly on one data-set (or subject) and the worst on the rest, it could potentially be highlighted as a relatively fair performer although from the reader's point of view it would probably be considered the worst predictor. As such, it is more appropriate to use statistical tests to ascertain whether their are significant differences between various learning machines and/or other comparative factors. A statistical test is centred around a null hypothesis and determines if a result is unlikely to have occurred based on chance alone. This is according to some pre-determined significance threshold and thus leads to a rejection when this threshold is exceeded. In terms of classifier comparisons, a typical null hypothesis would be, for example, that the average or expected performance of the various classifiers on different data-sets are all equal. Therefore the rejection of this null hypothesis is indicative that some strategies outperform others.

For comparison of two factors or classifiers, the parametric paired t-test is a common way of testing for differences between results. This is used when the same data sets are employed during analysis of the results and therefore the observations can be considered

repeated measures and thus paired. When this is not the case, that is, when testing factors are associated with different subject groups potentially of varying size, the less powerful unpaired t-test can instead be used. T-tests inherently assume that the distribution of your results is normal, however, this is often not the case and for small data sets is generally violated. When the distribution are non-normal, the nonparametric equivalents, namely the Wilcoxon signed-rank (paired) or rank-sum (unpaired) tests can be utilised [193]. These non-parametric test are based on ranking the observations across the factor and as such make no a priori assumption about the data. They are less susceptible to outliers due to not relying on actual performance measure values. Although less statistically powerful, these nonparametric tests will generally be utilised in this work, due to the relatively small number (< 30) of subject data sets available. Furthermore, the Kolgomorov-Smirnov test indicated both normal and non-normal distributed results across the subjects for various discrimination strategies. When comparing measures from CV trials, the results from the individual folds cannot be directly compared due to violation of the independence assumption which underlies the majority of these techniques. Therefore, the average CV classification accuracies are used for comparison.

The previously mentioned statistical tests were not intended for performing multiple (> 2) comparisons. The reason they breakdown is because when evaluating all pairwise combinations of the factors using these methods, there is an elevated risk due to the pairwise combinations, of rejecting a certain number of the null hypotheses based on random chance. The usual way to circumvent this issue is to control the family-wise risk which is the probability of making a Type 1 error. A general solution is to modify the significance threshold to account for the multiple comparisons with the most conservative (but easiest to implement) being the Bonferroni correction [193]. Instead, more specialised techniques have been developed specifically for testing across multiple hypotheses and include parametric repeated measures analysis-of-variance (ANOVA) and its non-parametric equivalent, the Friedman test [194]. For the same reasons above the non-parametric Friedman test will be utilised in this work. The Friedman test highlights if there is a family-wise difference and not the specific factors or classifiers which are significantly different. When the family-wise hypotheses has been rejected, post-hoc tests based on pairwise combinations and a correction factor can be performed to highlight specific differences. This can be done using the correction factor mentioned previously specifically, the Bonferroni-Dunn test. However, due to its cautiousness, it is only useful when comparing a fixed (control) factor to each alternative in turn. For comparing all combinations of classifiers (or factors) against one another, the less conservative Nemenyi test can be performed and is analogous to the Tukey test used in ANOVA. The main premise of such a post-hoc test, is that two classifiers are considered significantly different if their average ranks differ by at least the critical difference (CD) computed from the corrected significance threshold.

A useful way of visualising the results from a post-hoc test is using a critical difference diagram [193, 194]. Figure 5.14 gives an example of a critical difference diagram based on seven factors. In this example, seven factors (A,B,C,D,E,F,G) are being compared, with their average ranks highlighted on the scale. The CD value based on the Nemenyi test is shown above the scale and thus any two factors which have an average rank greater than this value are considered significantly different. Furthermore, the factors which are not significantly different can be grouped together as indicated by the thick black lines. In this example, factor C is considered significantly better than E, F and G, but insignificantly different when compared to factors A, B and D.

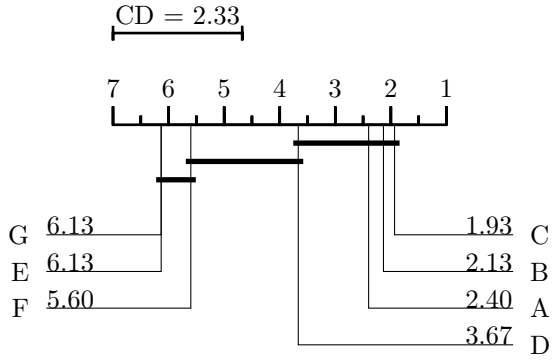
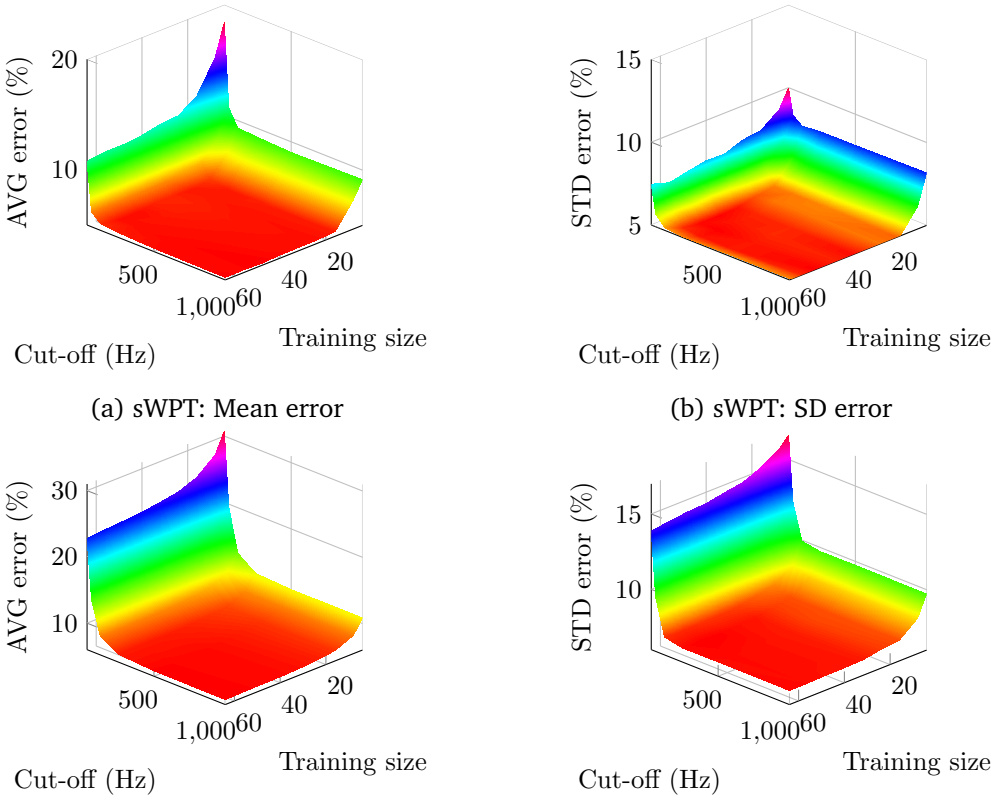


Figure 5.14: Example of a CD diagram. The factors are labelled A-G with their ranking value indicated next to them and on the scale. Factors joined by a thick black line show insignificant differences while unjoined factors are significantly different. This significance is based on the critical difference shown above the scale.

5.10.3 Effect of training set size and filtering on performance

Prior to evaluation of the proposed discrimination strategies in terms of decoding TMEP activity, it is useful to analyse the effect of training set size and filtering in terms of performance. These effects have been tested using the eDC classifier, as it has a low computational cost and thus is appropriate when estimating average error surfaces over different training sizes and filter cut-offs. The following training set sizes have been tested, namely, 4, 8, 16, 24, 32, 40, 48, 56, 64 associated with each class. For each training size, the following filter cut-offs were tested 31.25, 62.5, 125, 250, 500, 1000 Hz. For each training size and cut-off pair, twenty trials of 10-fold CV was used to estimate the error rates for each of the fifteen subjects. Filtering was implemented using both the WPT and sWPT (Appendix A) as they give better signal representation for non-stationary signals. The sWPT results highlight if the reduction in coefficients associated with the WPT has any effect on classification. Figure 5.15 show the mean and standard deviation of the error surfaces across the fifteen subjects using the eDC classifier.



(a) sWPT: Mean error

(b) sWPT: SD error

(c) WPT: Mean error

(d) WPT: SD error

Figure 5.15: Euclidean distance classification performance for different training set sizes (4, 8, 16, 24, 32, 40, 48, 56, 64) and (s)WPT filter cut-offs (31.25, 62.5, 125, 250, 500, 1000 Hz)

The surfaces show that as the filter cut-off and training set size are decreased, there is a general increase in the error rate. The error rate increases visibly as the training set size decreases between sizes 20 - 30. The filtering does not have an effect on the error till below about the 250 Hz cut-off for the WPT case. Conversely, for the sWPT the effect of filtering on the error rate is reduced at lower cut-offs due to higher information content from retention of all WPT coefficients associated with each filter level. This highlights that the majority of discriminatory information is contained at frequencies less than 200 Hz which is also corroborated by the PSD plots in Fig. 3.6. Furthermore, these results indicate that a training set size of 32 instances should be sufficient for effective decoding of TMEP actions and are deemed an appropriate size for prior collection.

5.10.4 Comparison of discrimination strategies

To compare across bioacoustic decoding strategies, out of those tested, the most effective ones were selected for each feature extraction method (AR, DFT, DCT, PCA) in terms of type, ranking filter and classifier. These were evaluated against all the time domain classifiers, namely, eDC, MFC and DFC as, although, based on the same features, the classification process associated with each of these methods can be considered significantly different. The selection of the AR strategy was from a pool of four, based on the method of extracting the features, that is Burg or Yule's method and the classification models sUGC and sMGC. For the DFT strategy the initial pool consisted of 28 combinations, comprising the four specific types (magnitude, phase, real or imaginary parts) using each of the seven ranking methods. The DCT strategy was selected from different combinations of the seven feature ranking methods and three classifier models, namely sUGC, rUGC_{iqr} and rUGC_{mad}. The final PCA strategy was selected from an initial pool consisting of 21 combinations comprising the three specific types (PCA_{ucs}, PCA_{pcs} and PCA_{pci}) using each of the seven ranking methods. The feature ranking methods incorporate the six filter methods of simple, mRmM, mRmRM, mRV, mFS and pFS, and a single CV wrapper. For each feature ranking based discrimination strategy, after feature extraction, the features are ranked according to the stipulated ranking method. Based on this ranking, the top d-features were classified incrementally from $d = 1, 2, ..., D$ and the number of features that gave the best accuracy in each fold selected. Appendix B shows the results for each feature extraction discrimination pool highlighting why the selected strategy was chosen from each feature extraction grouping. For clarity the final discrimination strategies selected for comparison are described below

eDC Euclidean distance classifier based on time domain features

MFC Matched filter classifier based on time domain features

AR_{burg} Autoregressive features based on Burg's method and a standard multivariate Gaussian classifier

DFC Decision fusion classifier based on time domain features

DFT_{real} Discrete Fourier transform (real part), simple feature ranking and standard univariate Gaussian classifier

DCT Discrete cosine transform, Mean rank mean magnitude (mRmM) feature ranking and standard univariate Gaussian classifier

PCA_{pci} Principle component features (pooled-class independent type), simple feature ranking and standard univariate Gaussian classifier

The testing of each strategy was done using one hundred randomised trials of ten-fold CV, based on the same partitions of training and test data. The training size was set at 32 instances for each class and no filtering was implemented. Figure 5.16 highlights the classification results for the best achievable performance from each of the seven strategies. The best achievable performance implies that for strategies involving feature ranking, that within each fold, the feature size that gives the highest accuracy has been selected. This indicates the highest overall classification accuracy that the discrimination system can achieve, independent of the feature selection algorithm. The results are presented in terms of mean classification accuracy for each subject, averaged over all trials as represented by each dot. The pooled mean across all subjects is also shown and indicated by the overlayed black symbol. For each result associated with a specific discrimination strategy, the results have been further separated into four different subject groupings based on the groupings given in Table 3.3. For clarity these are described as follows: All fifteen subjects from groups A, B and C (indicated by the red plot marking, diamond symbol, left side), the seven well trained subjects using the custom earpiece from group A (green plot markings, square symbol, centre left), four well trained subjects using the generic earpiece from group B (blue plot markings, circular symbol, centre right) and four minimally trained subjects also using the generic earpiece from group C (pink plot markings, triangular symbol, right side). For groups B and C the earpiece channel which gave the best results have been used.

It is clear to see that the relative position of the markers indicate that the DCT gives the best relative performance, especially demonstrated by the enhanced performance on the minimally trained subjects. The worst performer is the MFC, which is epitomised by its wide and lower performance also on this subject group. Both the DCT and PCA_{pci}-based classifiers show near perfect performance across all well trained individuals. To further compare the methods, the nonparametric Friedman test, as described in Section 5.10.2 was applied to the results. Within each of the subject groupings, the Friedman

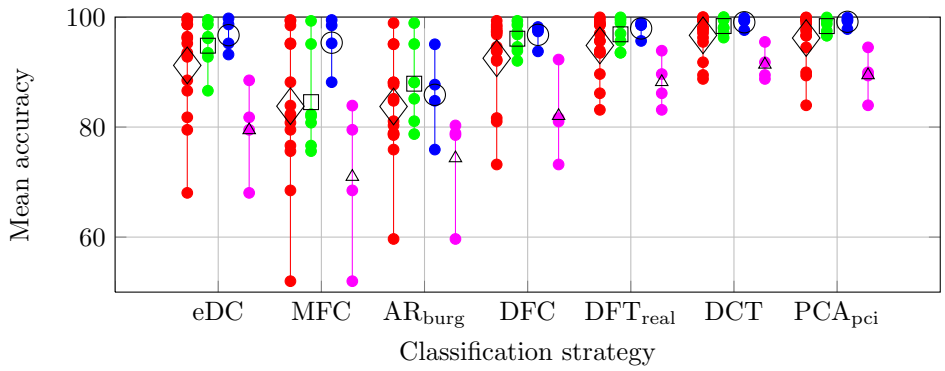
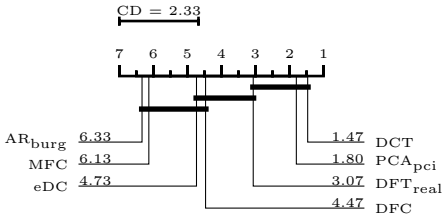


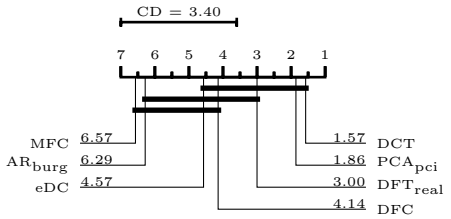
Figure 5.16: Mean classification accuracy between the seven classifier models indicated on the x-axis. The classification accuracy in each fold is taken to be the maximum achievable across the different subsets of features used. Results are presented for four different subject subsets including all subjects, custom well trained, generic well trained and generic minimally trained subject groupings. For each method and subset, the dots joined by a vertical line, show the mean accuracy for each individual subject while the black symbol on the same line shows the pooled mean accuracy across all subjects in the subset. Key: all subjects - red (diamond), custom well trained - green (square), generic well trained - blue (circle) and generic minimally trained - purple (triangle).

test rejected the hypothesis that all seven methods had the same performance. Post-hoc Nemenyi tests were then performed for each subject grouping with the results summarised in Fig. 5.17 using critical difference diagram. The p-value associated with the Friedman familywise significance is also given in the captions.

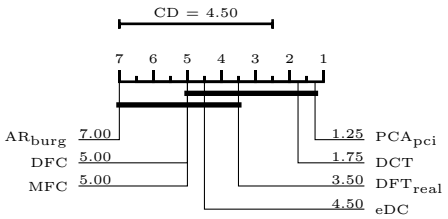
Across all subjects, the DCT and PCA$_{pci}$ have the potential for performing significantly better than all classifiers except the DFT$_{real}$. The MFC and AR$_{burg}$ strategies perform similarly and are significantly worst than the DCT, PCA$_{pci}$, DFT$_{real}$. The eDC and DFC strategies also have similar performance and are only significantly worst than the DCT and PCA$_{pci}$ strategies. The DFT$_{real}$ strategy lies between these two groupings and can be considered significantly better than the MFC and AR$_{burg}$ methods. Apart from the well trained generic earpiece group, the DCT scheme is always ranked highest and indicates its potential for effective decoding of volitional TMEP signals. For the minimally trained subject groupings, the rankings follow a similar structure. The eDC has its worse ranking in this group and although only a small number of subjects are available, it is still indicates where such a simple classifier can break down, that is, when the intra-class variance increases. The DFC can be considered an extension of a distance classifier albeit based on individual samples and with an additional Bayesian fusion. Although, this extra processing makes for a more complex algorithm, it also helps to make it more robust especially when there is higher intra-class variability in the data-set. As no feature selection mechanism are currently implemented at this stage of the analysis, this can only be considered a comparison of the ideal results. Only the eDC and MFC strategies require no feature selection and thus their



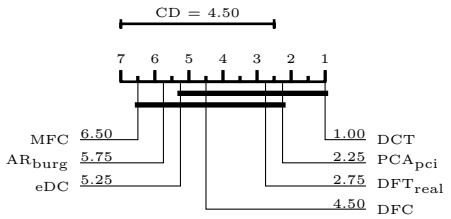
(a) All subjects (p = 7.7e-14)



(b) Seven well trained subjects using custom earpiece (p = 3.4e-6)



(c) Four well trained subjects using generic earpiece (p = 0.0021)



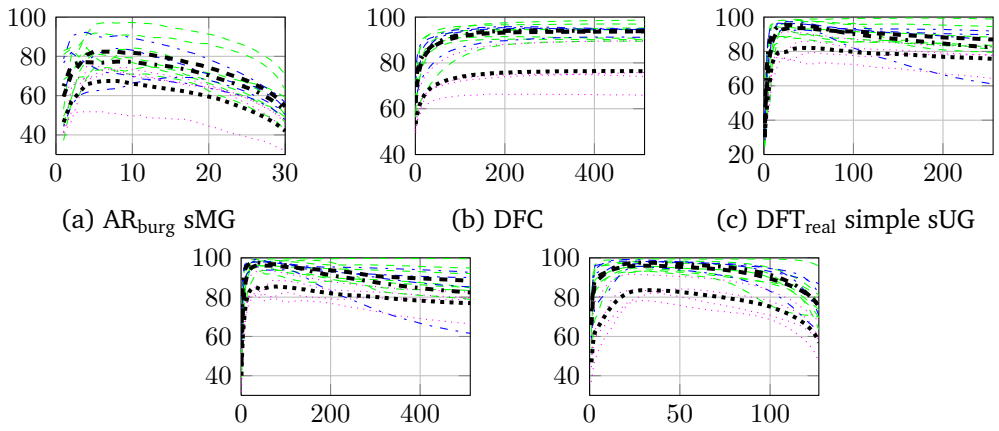
(d) Four minimally trained subjects using generic earpiece (p = 0.0017)

Figure 5.17: Critical difference diagrams portraying the post-hoc Nemenyi tests between the seven classification models for different subsets of the subjects. The familywise significance level using the Friedman test is given in brackets.

results indicate their true prediction power on unseen data instances. Having said this, most of the remaining methods bar the AR-based strategy have the potential for much more accurate decoding of volitional bioacoustic activity.

Figure 5.18 highlights, where appropriate, how the accuracy changes over various subsets of features for the five classification algorithms (AR_{burg}, DFC, DFT_{real}, DCT, PCA_{pci}). Each line represents the average accuracy over CV trials for each subject. The line colourings and markings are as follows, green dashed lines are associated with subjects from the well trained custom earpiece subject group, blue dashed-dotted lines are associated with the well trained generic earpiece subject group and pink dotted lines indicate the remaining minimally trained generic earpiece subject group. Additional thicker black lines of the same styling show the pooled average across each subject grouping.

As expected, the AR_{burg} scheme shows the worst performance across all subject groupings. Its best accuracy is achieved for a model order of between 5 and 10 and requires selection prior to classification. A model order of $p = 6$ proved most efficacious across all subjects and is the suggested model order for classification of volitional TMEP signals. The comparatively poor performance of the AR model may be due to it assuming the signal is (quasi-)stationary over the length of the segment [96]. This is probably violated due to the impulsive nature of the TMEP signals. The DFC exhibits asymptotic behaviour towards its maximum average classification accuracy as the number of features is increased. This is attributed to the weighting effect from the summation over the likelihood terms and in-



(a) AR_{burg} sMG (b) DFC (c) DFT_{real} simple sUG

(d) DCT mRmRM sUG (e) PCA_{pci} simple sUG

Figure 5.18: Mean classification accuracy against the size of the feature subset used for the various feature extraction methods and/or classification models. If applicable, the features are first ranked according to a feature weighting strategy as indicated by the sub-caption. Each coloured line represents the mean classification accuracy associated with one subject with green dashed indicating the subject belongs to the well trained custom earpiece subset, blue dashed and dotted lines indicating well trained generic earpiece and purple dotted indicating minimally trained generic earpiece. Furthermore, the 3 black thicker lines are the mean across each of these subject subset groups as indicated by the same styling.

dicates that feature selection is not important in this case. Instead, either all the features can be used during classification or if a smaller subset is preferable, the smallest number of features which achieves 99% of the maximum classification accuracy during training could, for example, be selected. The behaviour of the three remaining schemes (DFT_{real}, DCT, PCA_{pci}) show a maximum operating point generally for a much smaller subset of the top ranked features. There is rapid acceleration up to this point with the addition of each new top ranked feature, starting from the empty subset. After this maximum point, the accuracy declines at a much slower rate and implies that there is a relatively large band of feature subset sizes which produces high classification accuracy. For these discrimination strategies, feature selection is important so as to ensure that the most effective subset of features is selected during classification.

5.10.5 Comparison of feature selection methods

Figure 5.18 highlights that feature selection is required for the decoding strategies based on DFT, DCT and PCA extraction methods. The following five feature selection methods were compared, SFS_{filt}, SFS_{hyb}, SFS_{hEnv}, $WSFS_{filt}$, $WSFS_{hyb}$ for these feature extraction methods and are described below for clarity.

SFS_{filt} Sequential feature selection using a filter method to rank features and selection of

the feature subset based on those whose value is greater than the average across all features (see Section 5.6.1).

SFS_{hyb} Sequential feature selection using a filter method or CV to rank features and classification accuracy for feature subset selection (see Section 5.6.1). Classification accuracy is computed from ten-fold CV using the specified learning algorithm.

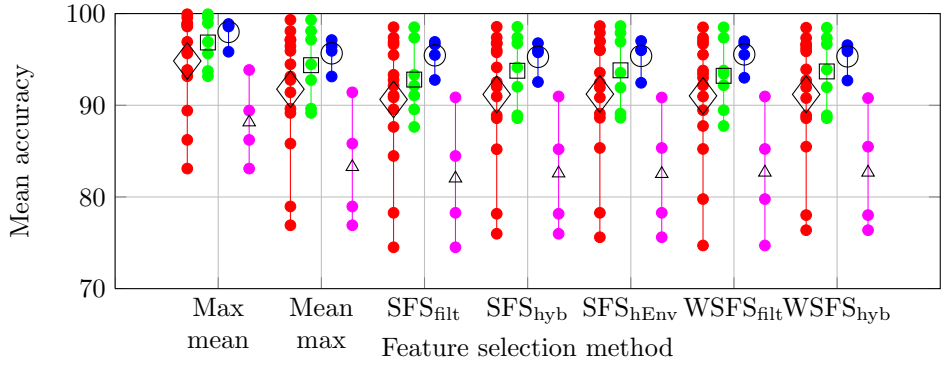
SFS_{hEnv} Sequential feature selection using a filter method or CV to rank features and classification accuracy for feature subset selection using an a priori weighting envelope over different subset sizes (see Section 5.7.1). Classification accuracy is computed based on 10-fold CV using the specified learning algorithm.

$\text{WSFS}_{\text{filt}}$ Weighted sequential feature selection using a filter method to rank the features and initial selection of feature subsets based on those whose value is greater than the average across all features (see Section 5.7.2). Final evaluation of the selected feature subsets is done using classification accuracy computed using a specified learning algorithm based on 10-fold CV. The number of WSFS iterations is set to $Q = 5$.

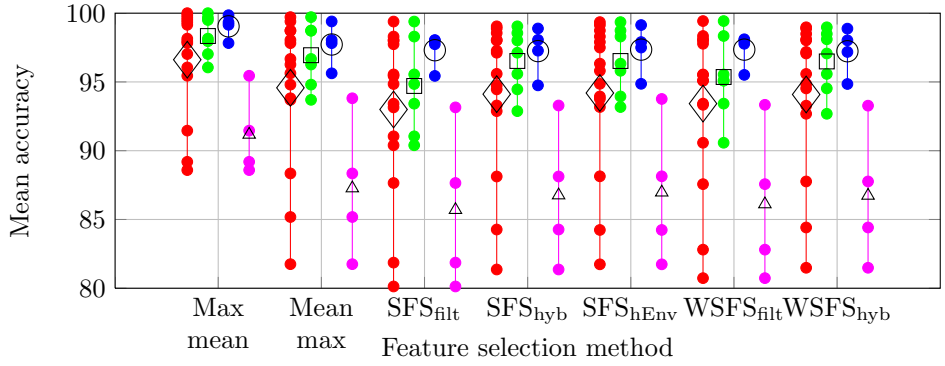
WSFS_{hyb} Weighted sequential feature selection using a filter method to rank the features and classification accuracy for initial subset evaluation (see Section 5.7.2). Final evaluation of the selected feature subsets is also done using the classification accuracy calculated from a specified learning algorithm based on 10-fold CV. The number of WSFS iterations is set to $Q = 1$.

Figure 5.19 shows the results of these five feature selection methods using features extracted using the DFT_{real}, DCT and PCA_{pci} transforms. In each case, the learning algorithm employed is a univariate Gaussian Classifier based on standard statistics (sUGC). For comparison, the mean max and max mean accuracies are also shown and can be considered benchmarks. The mean max is the maximum classification accuracy over different feature subset sizes within each fold. This is averaged over all CV trials and corresponds to the result shown previously in Fig. 5.16. The max mean is the maximum classification accuracy across different feature subset sizes after first averaging across all CV trials. For each subject, this indicates the subset size which gives the best performance. Furthermore, it is the average of these values across the subjects which motivates the choice of operating point required by the weighting envelope used by SFS_{hEnv}. This operating point is set at 30/256, 40/512, 20/127 for the three DFT_{real}, DCT and PCA_{pci} feature extraction methods. The same ranking filters as previously used by each feature extraction method are used in this analysis.

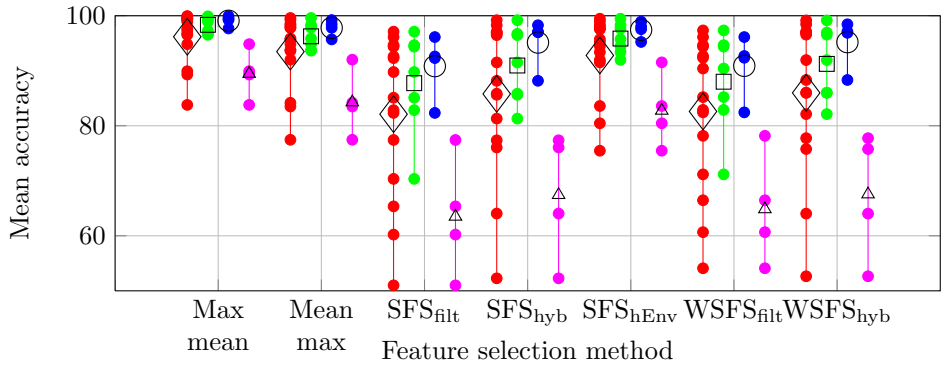
For both the DFT_{real} and DCT feature extraction methods, it can be seen there is no clear winner across the feature selection methods. Although not obtaining the best available mean accuracy as exhibited by the max mean result, they all perform comparable



(a) DFT$_{real}$ simple sUGC



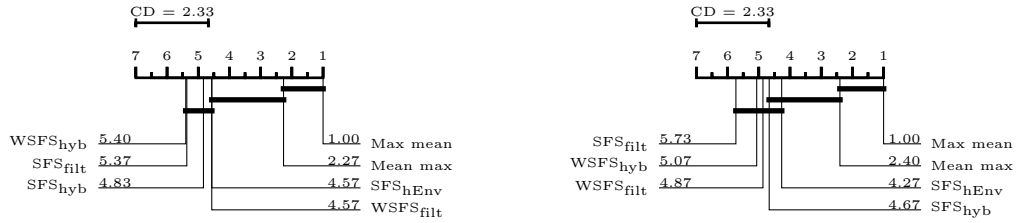
(b) DCT mRmRM sUGC



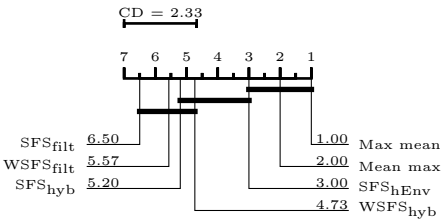
(c) PCA$_{pci}$ simple sUGC

Figure 5.19: Mean classification accuracy between the five feature selection methods alongside the mean max and max mean results for the DFT$_{real}$, DCT, PCA$_{pci}$ strategies. Results are shown for the following subject subset groupings, all subjects - red/diamond, custom well trained - green/square, generic well trained - blue/circle and generic minimally trained - purple/triangle. The mean max is the maximum classification accuracy across different feature subset sizes within each fold, averaged over all CV trials. The max mean is the maximum classification accuracy across different feature subset sizes after first averaging across all CV trials. For each method and subset, the dots joined by a vertical line, show the mean accuracy for each individual subject while the black symbol on the same line shows the pooled mean accuracy across all subjects in the subset.

to their mean max result using all feature selection schemes. On the other hand, the PCA_{pci} results show severely degraded performance when comparing the feature selection methodologies to their benchmarks, apart from the SFS_{hEnv} method. This indicates that these schemes were unable to select good feature subsets for this transform. To elucidate further on these results, the nonparametric Friedman test was applied to them. Across all subject data sets, analysis of just the feature selection methods (excluding the two benchmarks), yielded insignificant differences for the DFT_{real} and DCT transforms, however significant differences were found for the PCA_{pci} results (p = 3.9e-8). When including the benchmark results, the Friedman test rejected the hypothesis that all seven results had the same performance. Post-hoc Nemenyi tests were then performed for each feature extraction method across the seven results and are summarised in Fig. 5.20 using critical difference diagram. The p-value associated with the Friedman familywise significance is given in the captions.



(a) DFT_{real} sUG (p = 4.0e-10) (b) DCT sUG (p = 5.7e-10)



(c) PCA_{pci} (p = 9.4e-15)

Figure 5.20: Critical difference diagrams portraying the post-hoc Nemenyi tests between the five feature selection methods alongside the mean max and max mean results for the different subject subset groupings. The mean max is the maximum classification accuracy across different feature subset sizes within each fold, averaged over all folds. The max mean is the maximum classification accuracy across different feature subset sizes after first averaging across all folds. The familywise significance level, using the Friedman test, is given in brackets.

As expected, the critical difference diagrams indicate that there is no significant difference between the five feature selection methods for the DFT_{real} and DCT feature extraction schemes. In both cases, all feature selection methods are significantly worst then the max mean benchmark result. In the DFT_{real} case, the SFS_{hEnv} and $WSFS_{filt}$ are not significantly different than the mean max benchmark result but all other selection methods are. In

the DCT case, the SFS_{hEnv} and SFS_{hyb} are not significantly different from this benchmark while again all remaining selection methods are. For all the feature extraction methods, the SFS_{hEnv} is the best ranked feature selection. Having said this, it only makes sense to use this method in the case of the PCA_{pci} transform as it requires a priori knowledge based on these data sets and only makes a significant difference in this case. The reason that the other feature selection methods breakdown can be attributed to it being a data dependant transform and thus features extracted from the test data probably exhibit higher variability and are therefore less stable features. As the SFS_{hyb} doesn't rely on a priori knowledge, is ranked higher than its equivalent filter method and is relatively computationally efficient, it is preferred for feature selection when extracting DFT_{real} and DCT features from volitional TMEP data.

5.10.6 Custom vs. generic earpieces

When considering the TMEP concept specifically, it is useful to see if the type of earpiece used during signal capture has an effect on classification accuracy. The two data-sets compared are the seven well trained custom earpiece subjects (group A) against the best performing individual channel of the four well trained generic earpiece subjects (group B). The statistical tests used are the one and two-sided unpaired t-test as well as the non-parametric Wilcoxon rank-sum test. The one-sided unpaired t-test assumes that the custom earpiece will outperform the generic earpiece in terms of average classification accuracy across subjects. Table 5.3 shows the p-values for the three tests for each of the seven classification strategies from Section 5.10.4.

Statistical test	**eDC**	**MFC**	**AR_{burg}**	**DFC**	**DFT_{real}**	**DCT**	**PCA_{pci}**
Two-sided unpaired t-test	0.53	0.073	0.53	0.79	0.93	0.79	0.53
One-sided unpaired t-test	0.77	0.97	0.34	0.67	0.75	0.79	0.82
Wilcoxon rank-sum test	0.46	0.060	0.68	0.66	0.50	0.41	0.35

Table 5.3: Significance levels (p-values) comparing the mean accuracies of the custom and generic earpiece results for the seven classifier models. The subjects compared are all considered well-trained individuals. The tests performed are the parametric two-sided unpaired t-test, the parametric one-sided unpaired t-test (assuming the custom earpiece will yield a higher classification accuracy than the generic one) and the non-parametric Wilcoxon rank-sum test. All the tests show no significant differences between the two types of earpiece at the 5% level.

As can be seen none of the tests across all strategies show significance and indicates that, based purely on decoding power in the associated off-line conditions, there is no real advantage of using a custom earpiece over a generic one. A generic earpiece has the advantage that it is subject independent. This implies that it can be used 'off-the shelf', making the system cheaper and more cosmetically appealing.

5.10.7 Well vs. minimally trained subjects

In assisitve HMI systems, due to the unconventional nature of the input, there is generally a learning curve associated with operating the device. This is true for the TMEP concept whereby, the actions feel natural to execute but are not normally occurring. Therefore, it is of great interest to test if there is a significant difference between the well and minimally trained subjects in terms of their classification accuracy. The two data-sets compared are the four well trained custom earpiece subjects (group B) and the four minimally trained generic earpiece (group C). The other well trained subjects are not used due to the signals being recorded using a different earpiece type. Table 5.4 shows the p-values for the same three tests and seven classification strategies used during the earpiece comparison in the previous section.

Statistical test	**eDC**	**MFC**	**AR_{burg}**	**DFC**	**DFT_{real}**	**DCT**	**PCA_{pci}**
Two-sided unpaired ttest	**0.0088**	**0.018**	0.12	**0.011**	**0.0072**	**0.0029**	**0.0046**
One-sided unpaired ttest	**0.0044**	**0.0090**	0.059	**0.0054**	**0.0036**	**0.0015**	**0.0023**
Wilcoxon rank-sum test	**0.029**	**0.029**	0.20	**0.029**	**0.029**	**0.029**	**0.029**

Table 5.4: Significance levels (p-values) comparing the mean accuracies of the well trained and minimally trained subjects for the seven classifier models. The subjects compared all used generic earpieces. The tests performed are the parametric two-sided unpaired t-test, the parametric one-sided unpaired t-test (assuming the well trained subjects will yield better classification accuracy than the minimally trained ones) and the non-parametric Wilcoxon rank-sum test. The tests highlighted in bold indicate a significant difference at the 5% level.

Apart from the AR_{burg} classifier, all statistical tests show that there is a significant difference between well trained and minimally trained individuals. This corroborates the idea that the more the interface is used the better a subject is able to operate it. The fact that the tongue actions are not normally occurring, implies that they do require practise before an individual is able to use the system efficiently. The reason for AR_{burg} not showing significance can be attributed to it being the worst performing classification strategy. This implies that it will generally have large distributions in terms of accuracy, regardless of whether the subject is well or minimally trained.

5.10.8 Discussion

For true comparison of the proposed classification strategies for decoding volitional bioacoustic activity in assistive HMI applications, feature selection methods need to be included. This will give a true indication of the discrimination capacity associated with each decoding strategy. The model order of the AR_{burg} strategy is set at $p = 6$ as this was found to be the most effective number of features across all subjects. For the DFT_{real} strategy, sim-

ple feature ranking was implemented alongside SFS_{hyb} for feature selection and a sUGC for prediction. The DCT strategy used mRmRM for feature ranking alongside SFS_{hyb} and sUGC for prediction. The PCA_{pci} strategy used the simple feature ranking, sUGC for prediction and SFS_{hEnv} with an operating point set at 20/127 for feature selection. The eDC and MFC remain the same as in the initial comparison as they do not require features to be selected. The number of fusion decisions associated with the DFC was selected based on those which achieved 99% of the maximum accuracy found during training. Figure 5.21 highlights the results across the fifteen subjects.

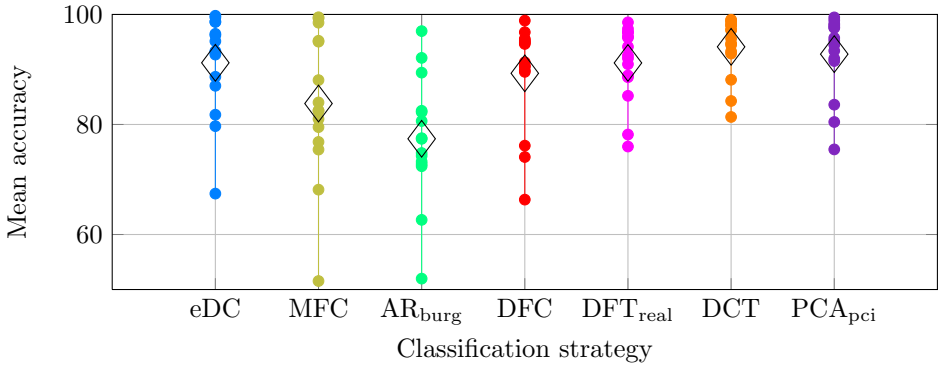


Figure 5.21: Mean classification accuracy between the seven classifier models indicated on the x-axis. Results are presented for all fifteen subjects with the mean classification accuracy calculated by averaging across all CV trials. For each method and subset, the dots joined by a vertical line, show the mean accuracy for each individual subject while the black diamond on the same line shows the pooled mean accuracy across all subjects in the subset

To further elucidate on these comparisons, the nonparametric Friedman test was performed and indicated that the classification strategies were significantly different. Post-hoc Nemenyi tests were subsequantly performed between the seven classification strategies and are summarised in Fig. 5.22 using a critical difference diagram. It can be seen that the DCT strategy still performs the most effectively in terms of classification accuracy and is significantly better than the AR_{burg}, MFC and DFC algorithms. The most notable change is the eDC strategy which is now ranked third and is significantly better than the AR_{burg} method. Conversely, the DFC only performs slightly better than the MFC algorithm. The increased performance of the eDC algorithm is attributed to its simplicity and thus has very few parameters to estimate during training. Its ability to provide effective prediction is based solely on a robust template being created with its effective performance attributed to the template creation process. On the other hand, the DFC algorithm although having the capacity for better classification, requires an $N \times m \times m$ likelihood matrix to be created thus potentially requiring a larger sample size to adequately estimate these multinomial probabilities [24]. The DCT is effective at representing frequency components on fewer

coefficients, having similar properties as the PCA transform [180]. Furthermore, it is a data independent transform which could explain its ability to generalise to unseen data better than the PCA transform.

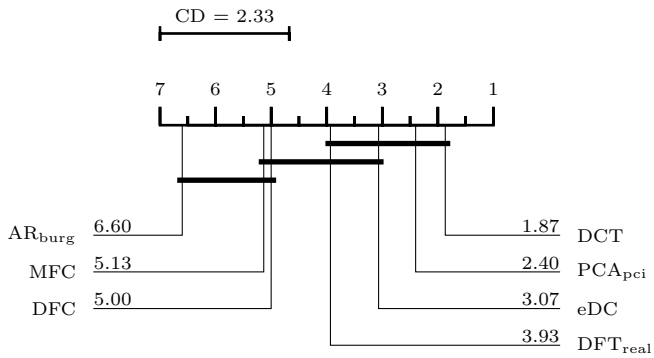


Figure 5.22: Critical difference diagram portraying the post-hoc Nemenyi tests between the seven classification models across all subjects. The familywise significance level using the Friedman test is p = 5.2e-010.

In summary, it has been shown that there are a variety of decoding strategies that can provide effective inter-action discrimination of volitional bioacoustic activity. This has been specifically demonstrated on four-action TMEP data. Although, no single strategy showed absolute dominance in terms of classification performance, the DCT-based strategy generally performed the most effectively while the AR model gave the poorest performance. The comparatively poor performance of the AR model may be attributed to it assuming the signal is (quasi-)stationary over the length of the segment which is probably untrue [96]. Therefore, for other biosignals, this assumption may hold truer and the AR model provide better performance. The effectiveness of the DCT can be attributed to it being data independent and therefore has better generalisation but with similar compaction properties to the PCA transform [180]. Compared to the DFT$_{real}$ transform, it provides double the feature resolution and is able to extract finer spectral content thus providing important additional discriminatory information. Another benefit of the selected DCT strategy is that the chosen feature selection method (SFS$_{hyb}$) requires no a priori knowledge about the data, for example the operating point used by the weighting envelope. Furthermore, in this case, the stability of the chosen feature set was not an issue, as neither the WSFS or SFS$_{hEnv}$ elevated the classification accuracy.

Specifically, in the case of the TMEP concept, it has been shown that a generic earpiece is preferable to a custom earpiece. This is due to there being no significant difference in performance across all the decoding strategies while the generic earpiece holds several advantages. These include subject independent design, cost and a more cosmetically discreet housing. The learning effect associated with performing the actions and thus operating the assistive HMI has also been evaluated. Apart from the AR model, all the decoding

strategies showed a significant difference in terms of discrimination accuracy between the well trained and minimally trained subjects. This is indicative of the fact that the more the interface is used, the more accustomed and adept an individual becomes at making the volitional tongue movements.

5.11 Ensemble framework

This section presents a framework for combining multiple channels and a set of heterogeneous discrimination strategies with the potential for combining their classification abilities. This is achieved using a meta-classifier known as an ensemble or multi-expert committee. Such meta-classifiers are based on the human philosophy of seeking multiple expert opinions before making an important decision. Not only can it increase generalisation performance but also provides a natural mechanism for interference rejection based on dissent among the members. By combining a set of heterogeneous models, diversity among the ensemble can be achieved without requiring data resampling procedures which generally require larger data sets combined with unstable classifiers [195]. Fusion is achieved through voting methods based on the predicted label preference from each member. This allows for any model type to be fused as well as any number of multiple heterogeneous data channels. A background to multi-expert committees is discussed next, followed by an introduction to some common preferential fusion methods and an overview of the heterogeneous implementation of the meta-classifier. Diversity among the members is highlighted and the enhanced performance attributes indicated on dual channel TMEP data.

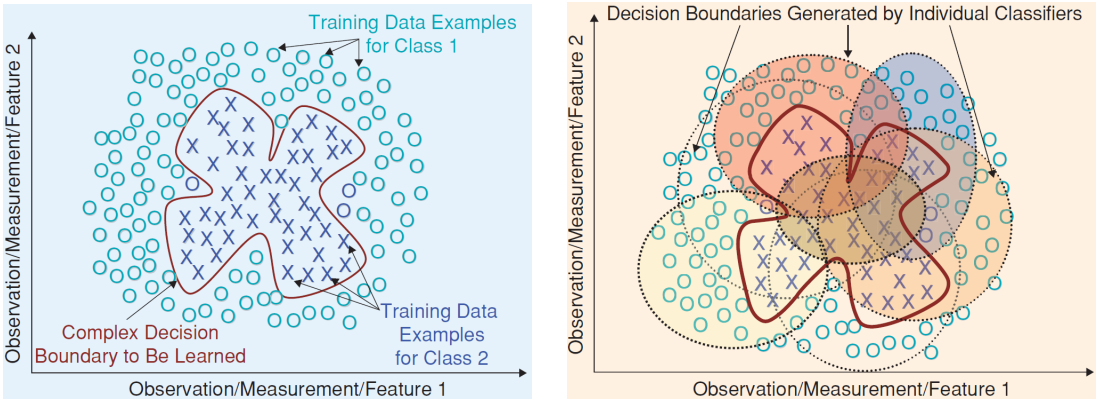
5.11.1 Background to multi-expert committees

The ultimate goal when designing pattern recognition systems is to achieve the best possible performance for the proposed task at hand [196]. This has therefore led to the development of a wide and diverse range of machine learning algorithms to solve a broad and general range of classification problems. Many of these classifiers require *a priori* assumptions which are either intrinsic to the specific model or necessary to make an algorithm tractable. Using such classification models can lead to *a priori* biases being placed on the classifier output thus leading to poor generalisation. Furthermore, it is well known in practise, that good performance on the training set does not generally predict good generalisation performance. This is especially true if the training set is small, as this implies the class distributions are under-represented and can only be poorly estimated [195]. From a human perspective, when an individual is required to make a decision, to ensure an effective choice is being made, they will generally ponder the situation for a long time. This involves gathering as much information as possible and processing all possible sce-

narios as well as evaluating risks associated with the various outcomes. This is analogous to presenting a classification model with a large volume of data and application of an appropriate learning algorithm. Moreover, if all possible scenarios are considered then an optimal decision can be made. On the other hand, if a human is required to make a rushed decision, without consideration of all the facts, then it is highly likely that a poor and regrettable decision will be made. In such cases, individuals will intuitively seek the opinions of others, especially if they can be considered an expert in the specified field. If the decision is an important one, individuals will seek help regardless. This approach has aided the effective decision making experienced in everyday life and therefore analogous ideas can be and are applied to the pattern recognition domain.

Ensemble-based classification systems also known as multi-expert committees, fuse the output from multiple classifiers producing a single reconciled decision [197]. Although such a meta-classifier may not necessarily produce better performance than the single best classifier in the ensemble, it reduces the overall risk of making poor selections during generalisation [195]. To obtain better performance from the ensemble, the individual outputs need to be diverse but in reality only need to perform slightly better than a random guesser. Ensemble systems can be both extremely effective for small training set problems and when the decision boundary is more complicated than can be modelled by a single simple classifier. Instead multiple classifiers can be fused enabling complex separating planes to be estimated. For example, Fig. 5.23 shows multiple quadratic learners being combined to learn a nonlinear boundary that would otherwise be impossible using a single simple classifier [195]. This is equivalent to creating a strong learner from a set of weak learners. Ensemble systems also allow for the combination of different types of data from different sources. Therefore, combining multiple classifiers leads to efficient and accurate learners, even when using simple classifiers [196]. Furthermore, depending upon the combination rule, they can naturally incorporate a reject option.

Multi-classifier systems can be broken down into two main types, namely, homogeneous and heterogenous ensembles. Homogeneous ensembles use a single base classifier model and create diversity amongst the individuals through either data resampling (as in bagging or boosting), using random feature subspaces (the random subspace method) or random initialisation of the model parameters (random initialisation of neural network weights for example) [167]. Conversely, within the heterogeneous framework, different base model types are incorporated into the committee with the idea being that different models can be both accurate and diverse [198]. When the training set is small and under-represented, enhancing the diversity of the committee in this way can be extremely beneficial. Additionally, if the training data is biased and only represents a small and local area of the population, such disparate mixtures can be more resistant to such data-derived traps [199]. Although the volume of research concerning homogeneous ensembles is



(a) Complex decision boundary that cannot be learned by linear or circular classifiers

(b) Ensemble of classifiers spanning the decision space

Figure 5.23: An ensemble of simple classifiers are able to learn complex decision boundaries ([195])

much greater, this work deals with small sample size training and therefore a heterogeneous ensemble framework is considered. As different feature extraction techniques such as time domain and frequency domain features have been explored, this is expected to further increase the diversity among the heterogeneous ensemble members.

5.11.2 Fusion by preference

Fusion between the individual class outputs can be achieved either through combination of the raw discriminant values or individual class labels. The continuous discriminant values are fused through either a linear combination or averaging, while class labels are combined using voting or label ranking strategies [196]. Within a heterogeneous framework, the individual discriminant values can be of disparate natures such as outputting continuous/discrete, numeric/nominal or probabilistic/score values. Therefore, when combining the raw discriminant values, normalisation to a common range is required and is commonly the probabilistic range that is [0,1). If a base model doesn't naturally output a probability, then it can be parametrically rescaled using a softmax normalisation. This requires the tuning of an additional parameter to control the linear mapping region of the sigmoid function [160]. On the other hand, rank-based preferential strategies enable fusion of disparate discrimination models as the ranking process ensures a nonparametric implementation [197]. A further and important benefit is label fusion naturally allows for a rejection output when ambiguity is experienced amongst the ensemble. As discussed previously, within a bioacoustic signal classification setting, the risk of interfering signals contaminating the recognition process is relatively high. Although within the real-time system, an additional interference rejection subsystem has been implemented, this predominantly only rejects higher frequency interference such as speech and misclassifies

lower frequency interference [27]. A secondary mechanism for rejection of lower frequency interfering signals is therefore highly desirable. This is assuming that for a given interference signal the class labels assigned by the ensemble members is sampled from a random distribution.

Within the ensemble literature, the most widely employed voting strategy for label fusion is the majority vote [195, 196]. This is due to its simplicity, tractability and it democratic decision. Other preferential strategies include plurality, alternative, Borda count and the Condorcet pairwise voting schemes and alongside the majority vote are described below [197].

Plurality Also referred to as first-past-the-post, plurality is a voting system in which the class with the most votes across the members is assigned as the overall ensemble output. Classification rejection will only occur if there is a tie for this top position.

Majority Similar but often confused with the plurality voting method, a majority vote gives a more democratic output. This is due to a winner only being selected if a majority verdict is reached. If less than 50% of votes are cast in favour of the winning class, then the classification is rejected. Therefore, this has higher rejection capacity than the aforementioned plurality scheme.

Alternative Also known as instant-runoff voting and is considered a more proportional representation of the voting procedure. It can be considered an extension to the majority vote. Following a majority vote, if a winner isn't assigned, the class with the minority is removed from the classification set and a majority vote takes place again. This is repeated until either a majority winner is found or the two remaining classes have an equal number of votes cast in their favour. This implies that classification will be rejected. Due to this iterative formulation the alternative voting scheme will generally also have a reduced rejection capacity.

Borda count This strategy was devised by the 18th-century French mathematician Jean-Charles de Borda. Each base classifier assigns a rank to each class based on the ordering of the discrimination values. These ranks are then converted directly to a score, ranging from $M-1, M-2, ...0$ for the rankings $1-M$. The scores are tallied over the entire ensemble and the class with the largest is declared winner. A tie for top position can occur and if so, the classification will be rejected.

Condorcet This method incorporates pairwise comparisons between each class, with a point tallied for each winner. It was developed and named after another 18th-century French mathematician Nicolas de Condorcet. The number of comparisons for each base model will be $M!/(2(M-2)!)$, with the class that wins the most head-to-head pairings declared overall ensemble winner. Due to the transitive nature of

this method, rejection of the classification can occur albeit when two or more classes are equal top scorers.

5.11.3 Heterogeneous implementation for multi-channel systems

Another benefit of ensemble learning systems is it provides a suitable framework for concurrent classification of multichannel inputs. In the case of bioacoustic signals, it is often useful to collect data from multiple signals positioned at unique locations across an individuals physiology. This can mitigate the effect of localised noise, increase the overall amount of useful information that is received, while differences between the channels can additionally provide useful discriminatory content. Figure 5.24 shows the implementation of a heterogeneous multi-channel classification system based on preferential voting fusion. It should be noted that this kind of approach naturally extends to heterogeneous signal sources, as individual channels do not need to contain information from a single data type.

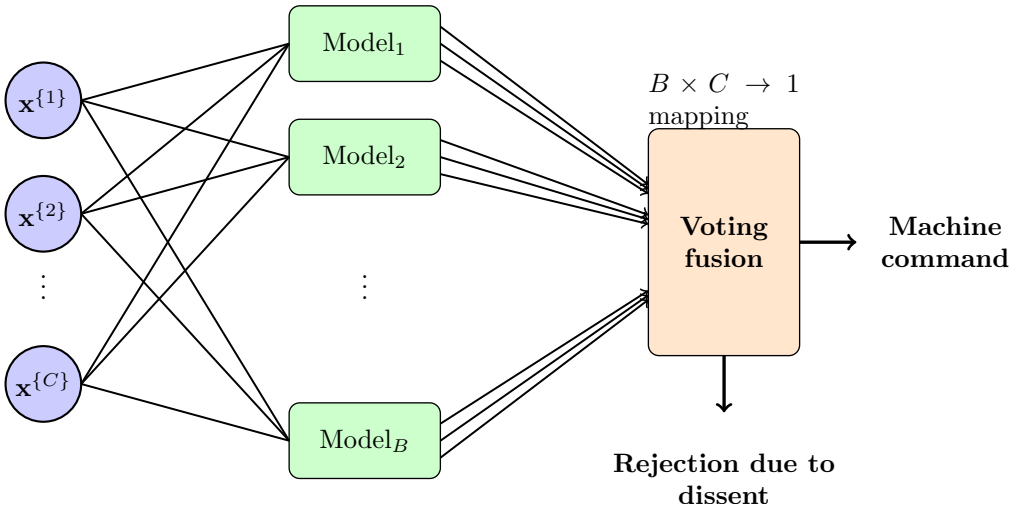


Figure 5.24: Multiple source ensemble classification based on preferential voting fusion. The output can either be a machine command or a classification rejection due to dissent amongst the individual ensemble members

5.11.4 Diversity measures

An ensemble is only needed if a single base model classifier performs imperfectly and makes erroneous classifications. It seeks to synergise multiple imperfect classifiers to enhance the overall performance of the learning machine. This enhancement will only occur if the errors made by the individual ensemble members occur on different pattern instances. If this is the case, strategic combination of these classifiers can lead to reduction

	v incorrect (0)	v correct (1)
u incorrect (0)	$D_{0,0}^{(u,v)}$	$D_{0,1}^{(u,v)}$
u correct (1)	$D_{1,0}^{(u,v)}$	$D_{1,1}^{(u,v)}$

Table 5.5: The 2×2 diversity matrix

of the overall error. Therefore, key to the success of the ensemble is the notion of diversity amongst its members [167]. Intuitively speaking, the ideal ensemble should contain members which provide as accurate predictions as possible and when inevitable errors are made, this should be on diverse instances across the pattern space.

Currently there is no unified method for measuring diversity between the classification outputs of a set of ensemble members. Instead, several measures have been defined [195]. These are commonly based on pairwise comparisons among the ensemble members with overall diversity estimated by averaging over these $T(T-1)/2$ comparisons. Pairwise comparisons are only considered in this work as they indicate which members are specifically diverse. The most commonly used of these measures is the Q-statistic, which between two ensemble members u and v is given by

$$Q^{(u,v)} = \frac{D_{0,0}^{(u,v)} D_{1,1}^{(u,v)} - D_{1,0}^{(u,v)} D_{0,1}^{(u,v)}}{D_{0,0}^{(u,v)} D_{1,1}^{(u,v)} + D_{1,0}^{(u,v)} D_{0,1}^{(u,v)}}, \qquad -1 \leq Q \leq 1 \tag{5.62}$$

with $D_{0,0}^{(u,v)}, D_{0,1}^{(u,v)}, D_{1,0}^{(u,v)}, D_{1,1}^{(u,v)}$ elements of a 2×2 diversity matrix described by Table 5.5. The diversity matrix is similar to a confusion matrix in the sense that the elements represent probabilities or counts. In the literature, the Q-statistic is often used for measuring diversity and indicates higher diversity as Q decreases. For uncorrelated classifiers $Q = 0$, as $D_{0,0}^{(u,v)} \approx D_{0,1}^{(u,v)} \approx D_{1,0}^{(u,v)} \approx D_{1,1}^{(u,v)}$. Other common pairwise metrics include correlation (Equation 5.63), disagreement (Equation 5.64) and double fault (Equation 5.65) measures and are given below.

$$\rho^{(u,v)} = \frac{D_{0,0}^{(u,v)} D_{1,1}^{(u,v)} - D_{1,0}^{(u,v)} D_{0,1}^{(u,v)}}{\sqrt{(D_{0,0}^{(u,v)} + D_{0,1}^{(u,v)})(D_{0,0}^{(u,v)} + D_{1,0}^{(u,v)})(D_{0,1}^{(u,v)} + D_{1,1}^{(u,v)})(D_{1,0}^{(u,v)} + D_{1,1}^{(u,v)})}}, \qquad -1 \leq \rho \leq 1 \tag{5.63}$$

$$DA^{(u,v)} = \frac{D_{0,1}^{(u,v)} + D_{1,0}^{(u,v)}}{D_{0,0}^{(u,v)} + D_{1,0}^{(u,v)} + D_{0,1}^{(u,v)} + D_{1,1}^{(u,v)}}, \qquad 0 \leq DA \leq 1 \tag{5.64}$$

$$DF^{(u,v)} = \frac{D_{0,0}^{(u,v)}}{D_{0,0}^{(u,v)} + D_{1,0}^{(u,v)} + D_{0,1}^{(u,v)} + D_{1,1}^{(u,v)}}, \qquad 0 \leq DF \leq 1 \tag{5.65}$$

The correlation measure has similar properties as the Q-statistic including maximal diversity at $\rho = -1$ as well as also being a symmetric measure. This symmetry implies that the values of $D_{0,0}^{(u,v)}$ and $D_{1,1}^{(u,v)}$ can be interchanged (and also $D_{0,1}^{(u,v)}$ with $D_{1,0}^{(u,v)}$) while still outputting the same diversity value. The disagreement measure is also considered symmetric but doesn't consider the instances where the outputs of u and v are the same. The larger the value of $DA^{(u,v)}$ the more diverse u is considered with respect to v and vice versa. The double fault measure is asymmetric as it is only relies on $D_{0,0}^{(u,v)}$ with higher diversity indicated by a smaller value.

A process has been developed for obtaining the 2×2 diversity matrix from the discriminant values associated with two classifiers. Figure 5.25 highlights this sequential process. For each pairwise comparison, the joint multinomial probabilities are first computed. Each element corresponds to the likelihood of class-m being classified to class-$c^{(u)}$ for ensemble member u, while for ensemble member v corresponding instances of class-m are classified as class-$c^{(v)}$. The green solid filled elements indicate where both members classified class-m correctly. The red (north-west) hatched elements indicate where member u classified class-m correctly but member v classified incorrectly to one of the $M-1$ remaining classes. Similarly for the blue (north-east) hatched elements but swapping u and v. The unfilled white elements indicate where both classifiers were incorrect and in terms of the input-output mapping could happen in $(M-1)^2$ ways for each class. These different regions can then be collected to form 2×2 diversity matrix for each class and pairwise comparison. They can be further collated across the individual classes to give an individual diversity matrix associated with each pairwise comparison. Finally, averaging over all pairwise comparisons will obtain a single diversity matrix associated with the whole ensemble.

Generally, diversity measures are used to assess and validate the potential of an ensemble system towards enhancing classification performance. Diversity driven ensemble learning although appealing, is generally not implemented in practise. This is due to there being a wide spectrum of diversity measures and notions and as such no single diversity measures consistently correlates with the actual ensemble performance [167]. The famous Adaboost algorithm is one of the only examples of a diversity driven meta-learners and is based on extracting a subset of classifiers from a larger library of base models that can consist of both homogeneous or heterogeneous members [200]. It uses an iterative strategy that weights the selection towards classifiers which can correctly predict those instances which were previously misclassified by the already selected ensemble members. It makes a final ensemble decision using a linear combination of the selected members based on these previously found weights. This continuous ensemble output implies that it doesn't have the natural rejection capabilities that preferential fusion strategies provide while additionally requiring an initially large selection of models to choose from.

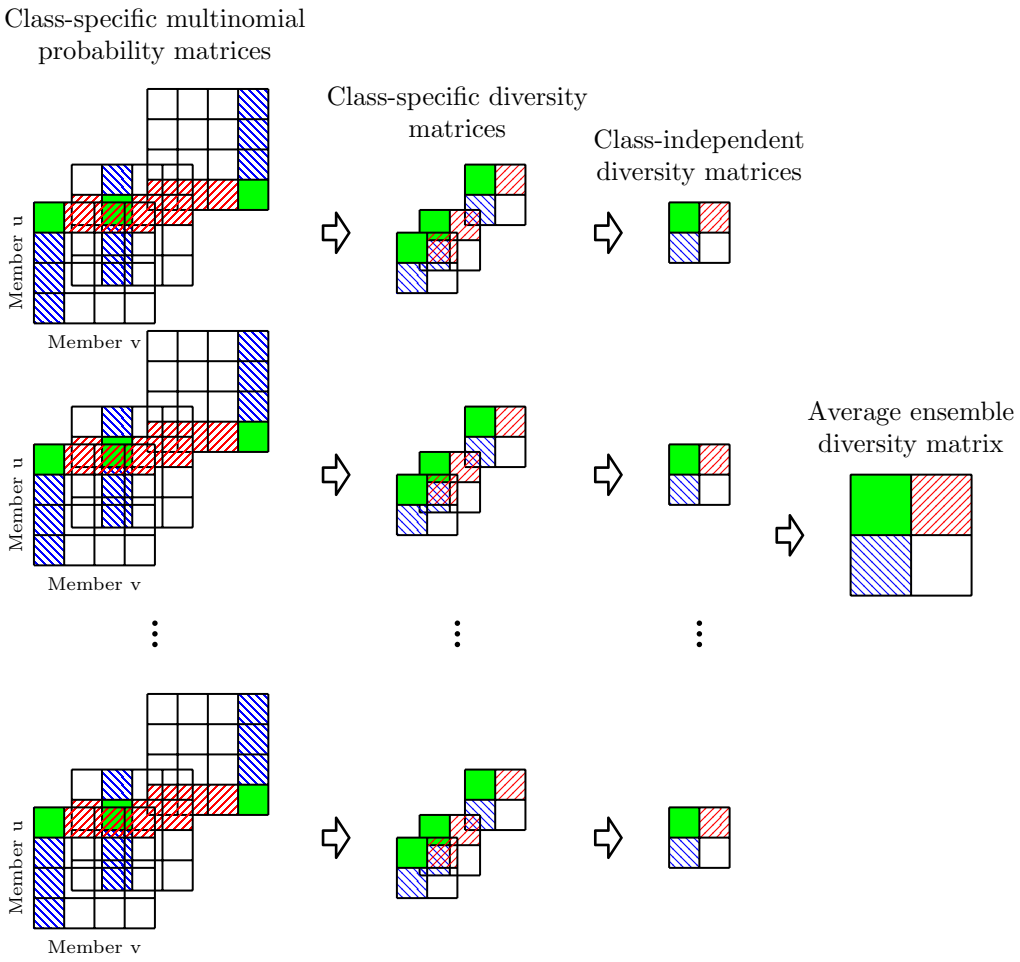


Figure 5.25: The stages associated with obtaining the average 2×2 diversity matrix. Each stage is described in the text. By breaking it down like this diversity can be measured between classes and across ensemble members

5.11.5 Study on the diversity of a dual-channel TMEP ensemble

Prior to ensemble classification testing, the potential of heterogeneous ensembles for decoding of multi-channel bioacoustic signals needs to be analysed. This initial validation is based on the diversity measures described in the previous section and is independent of the preferential fusion and ensemble meta-classifier. The proposed validation system is an extension to the single channel TMEP concept and instead utilises dual earpieces, one in each ear, in a stereo-recording manner. Figure 5.26 shows this dual-channel setup, which is expected to increase the amount of discriminatory information captured. It should therefore increase the classification power of the decoding algorithm as well as enhancing the overall interference rejection capabilities. This is at the expense of increasing the amount of computation necessary while potentially creating an 'over-safe' system due to elevating the FN^{-} rate. Furthermore, the user now has to use an earpiece in both ears simultaneously but although their ability to hear is somewhat reduced it is not considered by a significant amount. This is ratified by comparing it to using a personal stereo system but obviously without the obvious additional hearing block associated with music being played.

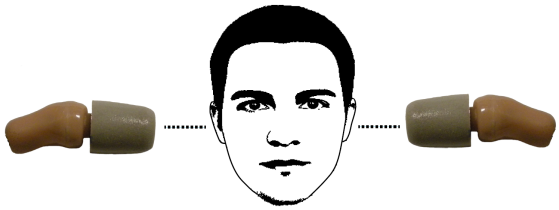


Figure 5.26: The dual-channel TMEP interface concept

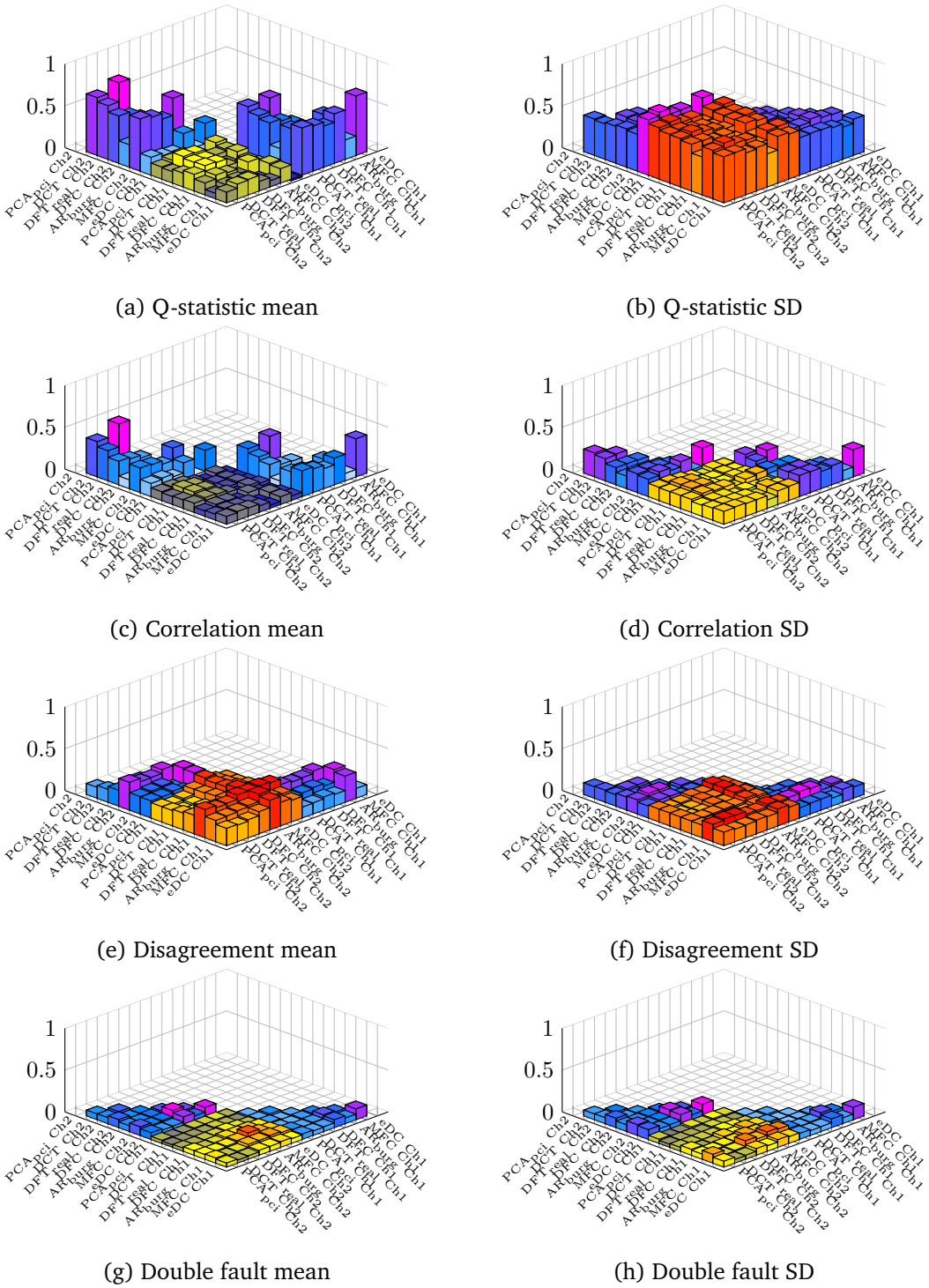
Dual-channel TMEP data was collected from eight subjects (four well trained and four minimally trained) as described in Section 3.5 with at least a hundred of each of the four actions per subject. The tongue movement events were detected and the signals segmented using the techniques outlined in Sections 4.1 and 4.2. The seven base classifiers ($B = 7$), namely, eDC, MFC, AR_{burg} sMG (p = 6), DFC, DFT_{real} sUG (SFS_{hyb}/simple ranking), DCT sUG (SFS_{hyb}/mRmRM ranking) and PCA_{pci} sUG (SFS_{hEnv}/simple ranking) were used as the base models in the ensemble. These are the same strategies compared previously in Section 5.10.8. For the two-channel system ($C = 2$), this implies a total ensemble size of fourteen members has been used in this study. The stratified K-CV testing procedure, was implemented to make efficient use of the available data, with twenty randomised trials and $K = 10$. For each trial, the data instances were used exactly once during the prediction phase, with storage of the discriminant values associated with each test instance for the ensemble classifications from both channels of data. Within each fold, the same data partitions were used as training and test sets, across all base model

classifiers and earpiece channels.

Class prediction labels were assigned based on the stored discriminant values using a decision output from a zero-one loss matrix. The pairwise multinomial probabilities between all the base model classifiers was computed and the 2×2 diversity matrices for each pairwise comparison was obtained for each subject and CV trial. For the four diversity measures, Fig. 5.27 shows the mean (left column) and standard deviation (right column) of the estimated diversity computed across all pooled subjects and trials. These are shown as 3D barcharts with each pairwise comparison represented at the XY intersections and the associated diversity value represented by the height of each bar. Only the lower half of each 14-by-14 barchart is displayed due to symmetry about the diagonal.

It can be seen that generally across all pairwise comparisons, both the Q-statistic and correlation measure assume positive values and are themselves correlated. This would tend to imply that diversity isn't very high across the ensemble. However, when comparing two strong classifiers, these diversity measures would tend to lean more positively, as the classifiers would have high $D_{1,1}^{(u,v)}$ values and thus be positively correlated. For highlighting when already reasonably strong base classifiers can potentially be enhanced through fusion, these diversity measures may not be the most suitable. Having said this, they do show that the inter-channel diversities, that is, comparisons between the separate earpiece channels indicated by the hot colour map are greater than the intra-channel diversities (comparisons on the same earpiece channel and indicated by the cool colour map). This indicates that utilising both channels of data can potentially increase the decoding performance by supplying more discriminatory information. This is also implied by the higher disagreement measure for comparisons between opposite earpiece channels. The high SD of the Q-statistic indicates that at times, the diversity was either very low or very high. When low it can probably be attributed to one or both classifiers performing nearly perfectly. When high diversity occurs it implies that, using the same two classifiers on a particular subset of training and test instances, may yield synergistically better performance. This could be, for example, when the user has had only minimal training and therefore the single model performance would be much lower.

A good indicator that the fusion between strong classifiers can be beneficial is having a high DA-to-DF ratio. As both DA-to-DF do not take into account $D_{1,1}^{(u,v)}$, they don't penalise for a single classifier performing well. However it is hard to quantify into a single measure as it is unbounded and therefore can assume infinite terms as either DF and/or DA can be zero valued. Instead, by comparing the DA and DF plots we can validate that on average, across all pairwise comparisons, disagreement between the pairs of classifiers occurs more frequently than double faults. Across all pairwise comparisons, subjects and trials, the average DA-to-DF ratio is larger than a factor of four. Furthermore, the double faults being made may differ across the classifier pairs. Although these different pairwise



(a) Q-statistic mean

(b) Q-statistic SD

(c) Correlation mean

(d) Correlation SD

(e) Disagreement mean

(f) Disagreement SD

(g) Double fault mean

(h) Double fault SD

Figure 5.27: Mean and SD for the four pairwise diversity measures across the fourteen ensemble members (seven classifiers $\times$ two channel sources). The measures are averaged over the eight subjects and twenty CV folds.

measures at times suggest different conclusions about the diversity of the ensemble, they are still useful measures. These discrepancies are in part due to there not being a clear definition of what diversity actually means [167]. However, they help in shedding different light on the pairwise interactions among the various members. However, regardless of the diversity measures used, there is not a direct relationship between better ensemble prediction and larger diversity amongst its members [195].

5.11.6 TMEP fusion results

To elucidate on the performance of the proposed heterogeneous ensemble framework, combinations of the seven base classification models and stereo channel data sets were explored. This was based on the same data as the pairwise diversity measures were calculated from. For each CV trial, all combinations of the seven base classifiers were fused, thus giving 127 different ensemble combinations. For ensemble sizes $B = 1, 2, \ldots, 7$ the number of associated combinations is $\{7,21,35,35,21,7,1\}$. The fusion of these results were tested for both channels individually and the two channels combined. In the stereo case, this led to even-valued ensemble sizes of $2, 4, ..., 14$. For each combination, the discriminant values corresponding to specific training and test set partitions were fused using the five preferential voting strategies outlined in Section 5.11.2. The output from this fusion process could be summarised by three measures, namely, (1) a correct prediction of the class associated with a particular test instance, (2) an incorrect labelling or (3) rejection of the decision due to ambiguity. This allows comparisons to be made between ensemble size, contributions from individual models, differences between the voting schemes and channel effects in terms of these performance measures.

To highlight the effect of ensemble size on classification performance, the mean classification performance can be computed across all the ensemble combinations of a particular size. Figure 5.28 gives results using a single channel setup while Fig. 5.29 gives results from the dual channel setup. As mentioned earlier, the size of the ensemble for the dual channel setup is double that of the single. The results are plotted for each subject, averaged over all CV trials and shown using the five fusion schemes. Also shown on the plots, is the pooled mean across all subjects as highlighted by the black symbol. Each symbol and colour is associated with one of the voting schemes as described in the figure caption. The individual plots are associated with the mean percentage of (1) correctly classified instances, (2) erroneously classified instances and (3) decision rejection rates.

The selected channel used in the single channel analysis is the best performing channel in terms of maximising the correct-to-error rate ratio for each subject averaged across all CV trials and individual models. Although this would not be known a priori, it was felt this gives a fairer comparison than just simply selecting the left or right channel. When an ensemble size of one is defined (leftmost grouping in Fig. 5.28) this is equivalent to

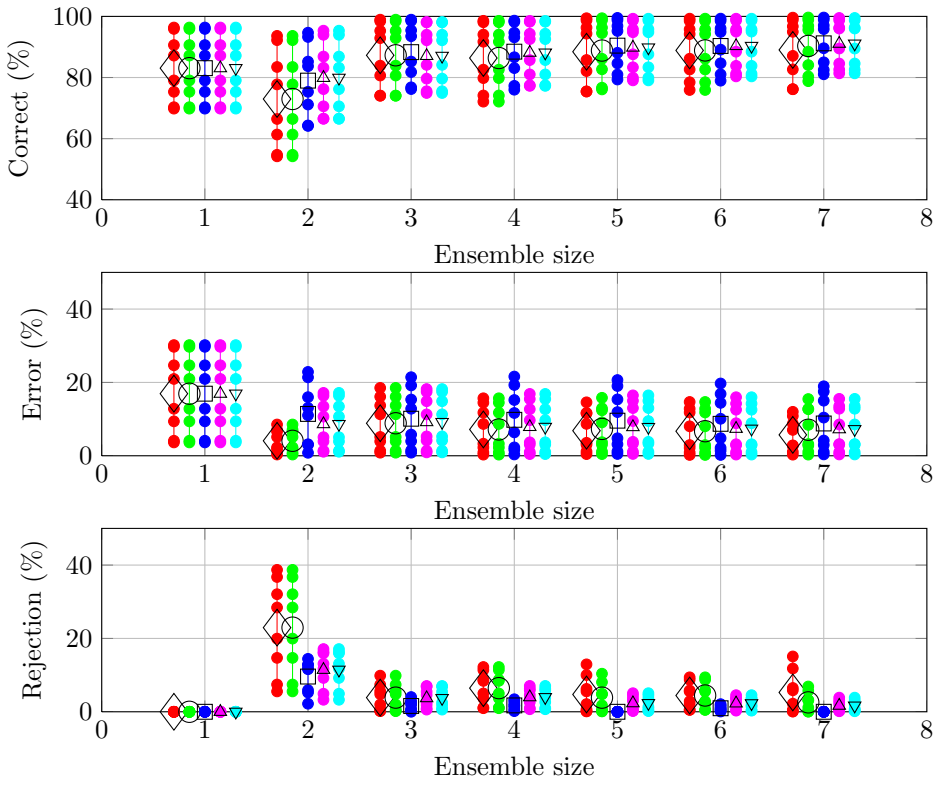


Figure 5.28: Average correct classification, error and rejection rates for different ensemble sizes from 8 subjects using a single earpiece channel. Results are presented for 5 different fusion strategies represented by different colors and symbols. For each subject, the results are averaged over all the potential combinations out of the 127 available as well as the 20 CV trials. Each dot represents a subject while the symbol is the pooled mean over all subjects. Key: majority - red/diamond, plurality - green/circle, alternative - blue/square, borda count - purple/up triangle and condorcet - turquoise/down triangle

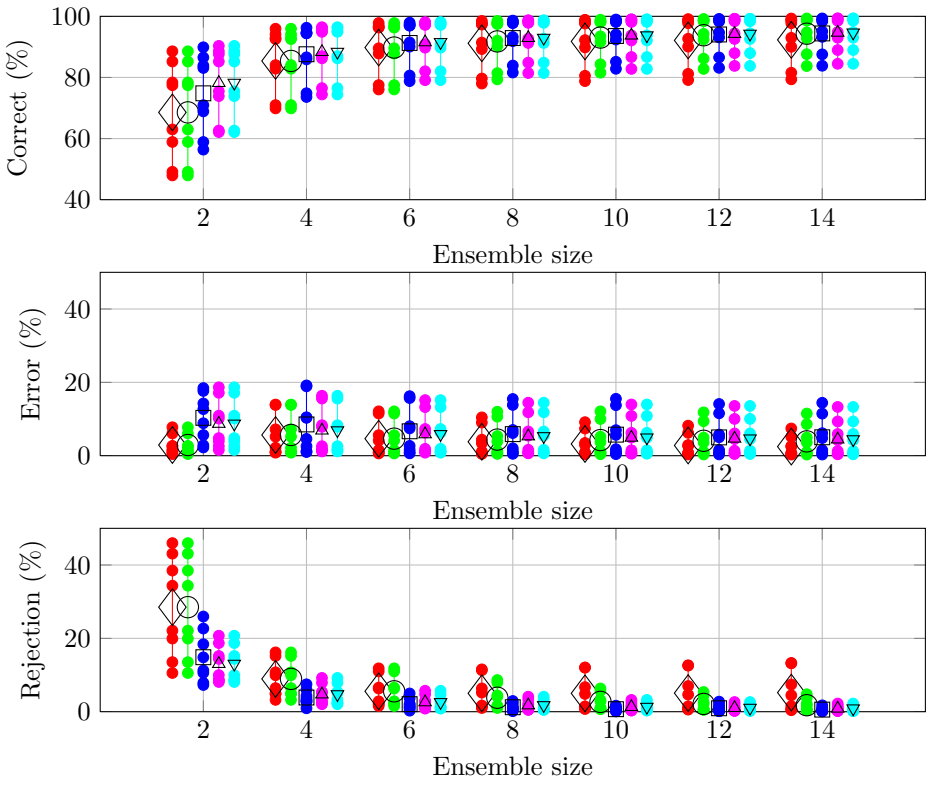
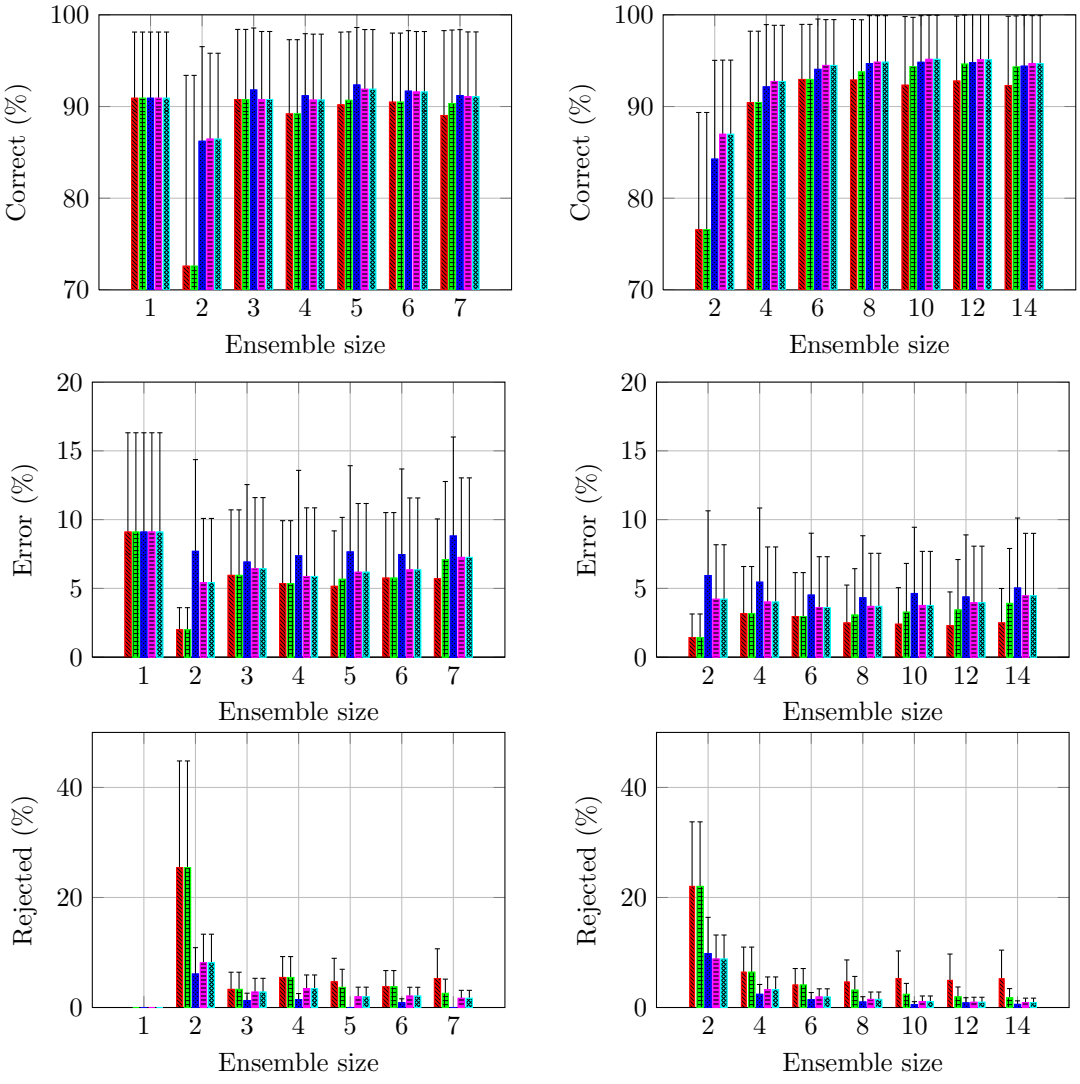


Figure 5.29: Average correct classification, error and rejection rates for different ensemble sizes for 8 subjects using both earpiece channels. The number of ensemble sizes are now obviously double that of the single channel results. Results are presented for 5 different fusion strategies represented by different colors and symbols. For each subject, the results are averaged over all the potential combinations out of the 7 available as well as the 20 CV trials. Each dot represents a subject while the symbol is the pooled mean over all subjects. Key: majority - red/diamond, plurality - green/circle, alternative - blue/square, borda count - purple/up triangle and condorcet - turquoise/down triangle

the average performance across all the individual classifier outputs and as such, has zero rejection capacity. The fusion methods have no effect on this result and therefore have the same output, with this result considered the baseline. An ensemble size of two gives the highest rejection rate and is to be expected due to a higher risk of discrepancies. In this case, the majority and plurality voting schemes have particular high rejection rates due to disagreements in predictions leading to rejection. This has the effect of significantly reducing both the correct classification and error rates simultaneously. In general, the majority voting strategy has the biggest effect in terms of reducing the error rate. This can be attributed to a higher rejection capacity, although it is at the expense of a slight reduction in correct classification. On the other hand, the alternative voting strategy has the smallest rejection capacity thus giving the largest error rate alongside slightly higher correct classification performance. The Borda count and Condorcet schemes have the same performance and in terms of the error rate lie between the two previously mentioned cases. As the ensemble size is increased further, there is a trend of increasing correct classification and decreasing error rates. In the case of the single channel setup, this tails off after the ensemble size reaches approximately five members and for the dual channel setup, when the ensemble reaches approximately eight members.

To truly motivate the effects of using a heterogenous ensemble in terms of increasing inter-action classification performance, specific ensemble combinations need to be analysed. Although averaging over the various combinations presented previously elucidates general trends in performance associated with different ensemble sizes and channel configurations, they do not indicate when and if the ensemble outperforms all single classifiers. Reduction in the error rate is considered the most pertinent attribute for a learning machine and therefore results are now presented in this regard. For each subject, the average correct classification, error and rejection rates are computed across all CV trials, for every ensemble combination. By grouping ensembles of the same size, the combination which provides the minimum error rate can be selected within each of these subsets for each subject. For both the single (left column) and dual channel (right column) configurations, Fig. 5.30 highlights these results, giving the pooled mean $\pm$ 1 SD across all subjects for each voting strategy. The results for the ensemble size equal to one, are now for the best performing single base model, namely, the DCT-based classifier which has the following performance: $90.9 \pm 7.2\%, 9.1 \pm 7.2\%, 0 \pm 0\%$ (correct$\pm$1SD, error$\pm$1SD, rejected$\pm$1SD). For the rest of this section, results will be presented as triplets following this convention.

For the single channel setup, there is similar decrease in error rate across all ensemble sizes, however this is due to both an increasing rejection rate and decreasing correct classification rate. In many cases, the correct classification rate decreases below that of the best performing single classifier average. For example, using all seven base classification models and the majority vote fusion gives $89.0 \pm 9.3\%, 5.7 \pm 4.3\%, 5.3 \pm 5.4\%$. This still signifies



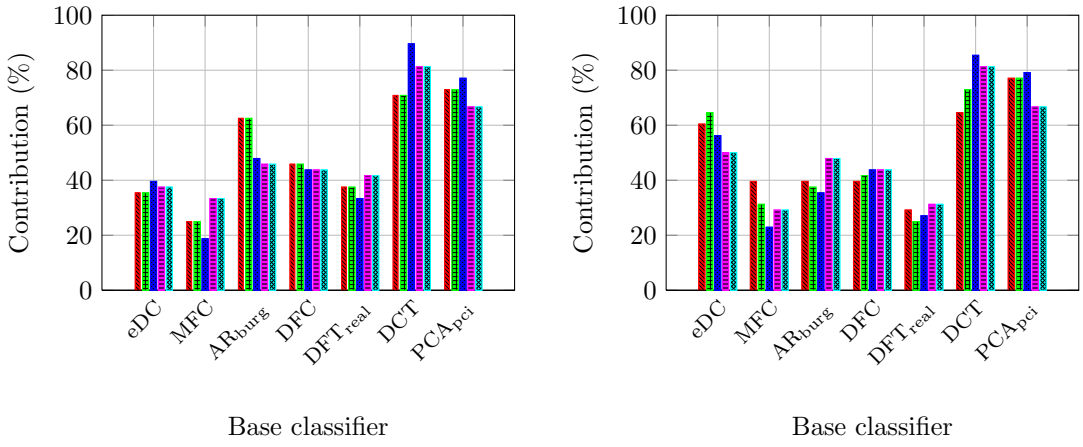
(a) Best single channel

(b) Dual channels

Figure 5.30: Mean correct, error and rejection rates for the minimum error combination of ensemble members for each subject across different ensemble sizes. Results are shown for both the single and dual channel setup as well as the 5 different fusion methods. The bars indicate the pooled mean across 20 CV trials and the 8 subjects while the error bars indicate the corresponding standard deviation. The intra-bar grouping (from left-to-right) indicate the majority (red), plurality (green), alternative (blue), borda count (purple) and condorcet (turquoise) fusion strategies.

an improvement, as the drop in classification accuracy is only 1.9% while the decrease in error rate is more than double this at 3.4%. In the dual-channel configuration, for the same ensemble and voting method, the performance achieves $92.3 \pm 7.5\%, 2.5 \pm 2.5\%, 5.2 \pm 5.2\%$ which indicates both an increase by 1.4% in correct classification and a decrease by 6.6% in the error rate when compared to the best single classifier result. Similarly to Fig. 5.29, there is a general trend of increasing correct classification and decreasing error rates as the dual-channel ensemble increases.

Figure 5.31 shows the relative contributions of the individual base classifiers as a ratio of how many times they were actually chosen to the overall number of times they could have been selected. This has been scaled to be represented as a percentage and the contributions from the full ensemble removed as this naturally selects all members. This gives a relative indication over different ensemble sizes, of which classifiers contributed to the ensemble in terms of lowering the error rate. Similar plots can also be made for maximising the correct classification rate, for instance.



(a) Best single channel

(b) Dual channels

Figure 5.31: Relative contributions from each base classifier to the reduction in the error rate associated with the test data and the initial pool of ensemble members. This is shown for the five preferential fusion methods for both the best single channel and dual channel setup.

This plot indicates which classifiers may be more suitable for use as a base models within an ensemble framework. The contribution can potentially be used to rank the classifiers based on the initial pool of classifiers and the given data set. It should be noted that the ensemble members that are shown to contribute more are not necessarily the individually better performing classifiers. Take, for example, the AR_{burg} classification model and the single channel ensemble setup, it contributes over 60% (ranking it 3rd) even though it has the worst individual classification performance out of the seven base

classifiers. This indicates that the AR_{burg} adds diversity to the ensemble which is generally the key to elevating its performance [195].

5.11.7 Discussion

In this section we have introduced a heterogeneous ensemble framework for enhancing the decoding performance associated with volitional bioacoustic activity. This meta-learner can consolidate the label outputs from heterogeneous source channels using heterogeneous base discrimination models. To allow for fusion of any type of discrimination model, the individual members are fused using a preferential voting strategy. This has the powerful benefit of allowing for classification rejection based on ambiguity in the ensemble and leads to a reduction in the associated generalisation error. This was demonstrated on a dual channel TMEP setup, whereby when fusing all seven models from both channels using a majority vote the average performance was $92.3 \pm 7.5\%, 2.5 \pm 2.5\%, 5.2 \pm 5.2\%$. The single best model performance was only $90.9 \pm 7.2\%, 9.1 \pm 7.2\%, 0 \pm 0\%$ indicating a clear increase in terms of both the classification accuracy and error rate.

The potential of this enhanced performance was initially validated through the use of pairwise diversity measures. Diversity is considered the cornerstone of ensemble methods and indicates whether combining pairs of base models will be likely to reduce the error. Having said this, no single diversity measure consistently correlates with the actual ensemble performance and therefore it is more useful as a design or initial validation tool [167]. One of the main advantages that the heterogeneous ensemble alongside a preferential voting strategy provide, is the potential to increase the rejection capabilities associated with the overall bioacoustic decoding. Based on this evidence, that is, enhanced inter-action classification performance and interference rejection capabilities, the dual channel heterogeneous ensemble using the seven base classifier models consolidated through a majority vote is selected for real-time TMEP decoding. To this end, results are presented in the following chapter during real-time inter-action evaluation and challenging interference rejection testing of the bioacoustic processing framework alongside the chosen dual channel heterogeneous ensemble.

5.12 Chapter summary

The most important aspect of an assistive HMI is the ability to effectively interpret the intention of the user based on the sensory input. In terms of using volitional bioacoustic activity as input, it allows for simple, imperceptible and noninvasive hardware, at the expense of a challenging decoding task. This is due to the signals being highly non-stationary, subject-specific and exhibiting significant intra-class variance. Therefore, this

chapter has presented various decoding techniques enabling effective interpretation of multi-class bioacoustic signals.

Initially, a general classification model was introduced for single channel multi-class discrimination. The foundations of this generalised model was the alignment process whereby unseen test instances were aligned to M templates, prior to decoding them. This implied that only class-specific classification models could be utilised. The remainder of the model utilised tools from the field of pattern recognition and incorporated three main steps, namely, feature extraction, feature selection and classification. For each of these constituents, various methods were assembled and introduced. This led to the proposal of various bioacoustic discrimination strategies which were evaluated using single channel TMEP data.

The chapter culminated with the introduction of a heterogeneous ensemble meta-learner. This framework allowed the decision outputs of various bioacoustic discrimination strategies and multiple channels to be consolidated. This fusion was based on preferential voting which not only allows different types of discriminant values to be consolidated but also adds rejection capabilities to the framework based on ambiguity between ensemble members. This ability to reject decisions not only implies a reduction in the generalisation error but can additionally reject interfering bioacoustic signals generated through non-voluntary activity. Based on the elevated inter-action classification performance, that is, increased classification accuracy and reduced error rate, alongside the interference rejection capabilities, a heterogeneous ensemble based on two bioacoustic input channels and seven individual discrimination models has been selected for real-time TMEP decoding. The chosen fusion method utilises a majority vote, as this was found to reduce the error (at the expense of slightly decreasing the correct classification rate) based on increased rejection of ambiguous instances. Similarly, this fusion strategy will provide the highest interference rejection capacity, with rejection results, based on this dual channel heterogeneous ensemble framework, demonstrated in Section 6.3.4.

Chapter 6

Real-time bioacoustic HMI based on TMEP signals

To fully validate the proposed bioacoustic processing architecture, a real-time implementation is required. This chapter demonstrates the effectiveness of the proposed strategies for online extraction and decoding of impulsive bioacoustic signals using a dual-channel TMEP setup. The bioacoustic activity is created through prescribed tongue motions and recorded bilaterally from the ear canals of the user. The reader is directed to Chapter 3 for a full description of TMEP signals. Within the real-time system, a specific interference rejection subsystem was implemented based on work carried out by research collaborators [27]. This was previously developed for off-line analysis and has been implemented into the real-time system. It can be considered a dichotomous classifier, which uses generic interference and the subject-specific TMEP signals as training data. An overview of this processing block is described for completeness. Although TMEP signals provide asynchronous control of peripheries, a stimulus-based strategy was devised enabling real-time evaluation. This allowed the actions to be stipulated by the machine and therefore tracking of the associated ground truth. The inter-class heterogeneous ensemble framework, previously introduced in Section 5.11, has been evaluated for real-time translation of tongue commands using both three and four action subsets. This discrimination strategy was trained using *a priori* collected subject-specific training data associated with each tongue motion. Due to the wide-ranging nature of non-voluntary activity that can corrupt the system, interference is generally considered subject-independent and generic. Therefore, the interference capabilities of the proposed decoding architecture was also assessed, using off-line generic interference data. A broad and challenging spectrum of interference was tested, in the form of three interference subsets of varying difficulty. The final subset contained lower frequency interference and although many of the signals are considered unlikely to occur, it simulates the most challenging interference that the system is likely

to encounter. The final evaluation is performed using the previously defined measures of sensitivity, specificity and overall classification accuracy, see Section 5.3.1. This indicates the potential of the proposed bioacoustic processing framework for real-time translation of volitional TMEP activity.

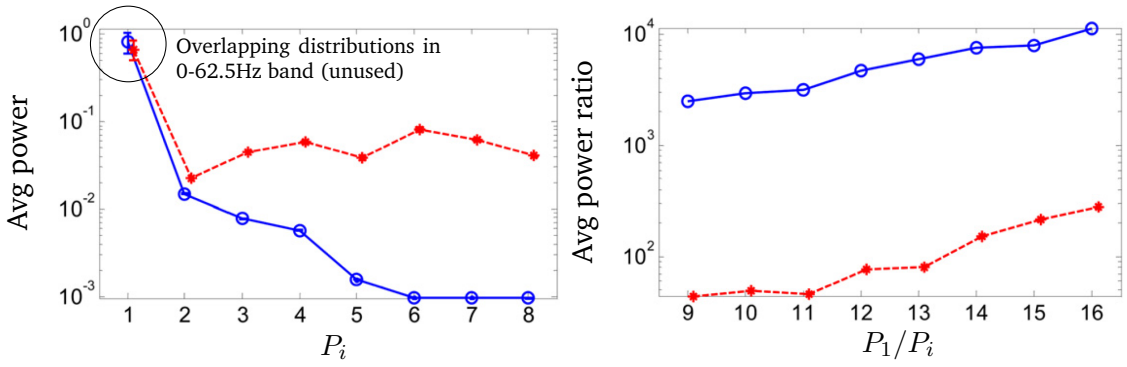
6.1 Interference rejection subsystem

The IR subsystem was developed to improve the robust realtime identification of tongue movement commands in the presence of internal and external interferences [27, 97]. It was developed concurrently by research collaborators and has shown effective rejection of higher frequency interference, such as speech and coughing. The subsystem is implemented directly following activity detection, preceding inter-action discrimination, and provides immediate defense against misclassification of false positives. The WPT (see Appendix A) can be used to extract transient changes in TMEP segments within different frequency bands. Therefore, the WPT bands were selected according to statistical distributions of the wavelet packet coefficients with the aim of maximising the separability between the tongue commands and interference signals. The interfering signals cover a broad and diverse spectrum including internally generated artifacts such as speech, swallowing, coughing, eating, drinking and external artifacts such as the individual's heart beat, remote muscular activity, bodily movement, limb tremor and environmental sounds. Such environments are common when extracting volitional activity from bioacoustic signals especially as the sensors are positioned remotely from the source of the activity. In the case of TMEP activity, the propagation of the acoustic wave between tongue and ear canal increases the risk of interfering activity merging with the captured signal in real-time. Having said this, by placing the microphone away from the oral cavity allows for an unobtrusive and noninvasive HMI system.

The IR subsystem is based on feature extraction, feature computation and selection and dichotomous discrimination between the controlled and non-controlled activity. The system is trained on subject-specific tongue movement signals and a subset of generalised interfering signals collected from three independent subjects. As the interference training data can only be a limited subset of the potentially occurring signals, a representative set of interfering activities were chosen including speech, coughing and a drinking signal. The speech activity included utterances of the words, '0', '1', ..., '9', 'on', 'off', 'start', 'stop', 'open' and 'close'. To extract features, segmented TMEP signals were decomposed in the WPT domain using a Sym7 wavelet filter at scale four [27]. At a sampling frequency of 2 kHZ, this provides sixteen channels of WPT coefficients, each with a frequency bandwidth of 62.5 Hz. It has been observed that tongue generated TMEP signals have the majority of their energy located in the low frequency bands (0-62.5, 62.5-125 Hz) whereas the energy

associated with interfering signals is distributed at both low and/or high frequency. For each channel, fifteen discriminatory features were computed and based on combinations of the absolute power and low-to-high frequency power ratios among the sixteen WPT bands. If P_i represents the power associated with a WPT band $i (i = 1, 2, \ldots, 16)$ calculated from the variance of the wavelet coefficients, then fifteen features per audio channel are computed as follows. The first seven features ($x_1 = P_2, x_2 = P_3, ..., x_7 = P_8$) are computed as the variance of the wavelet coefficients in WPT bands 2 - 8 (frequency range 62.5 - 500 Hz). The remaining eight features ($x_8 = P_1/P_9, x_9 = P_1/P_{10}, ..., x_15 = P_1/P_{16}$) are computed as the power ratio between band 1 and the remaining 9 - 16 bands. Previously, band 1 was also used as a feature but due to the highly overlapping class distributions, has been ignored in the real-time implementation.

Figure 6.1 shows typical average values of these power and power ratio features from a single subject. It can be seen, that the distribution of the energy associated with the interfering signals has large variations across all frequency bands. Conversely, the distribution of the energy associated with tongue generated TMEP signals occur mainly within the low frequency 0-62.5 Hz band. As this has a similar distribution to the interference data set this feature isn't considered during real-time feature computation.



(a) Average power (P_i) of WPT bands 1 - 8

(b) Average power ratio (P_1/P_i) of WPT bands 9 - 16

Figure 6.1: Average power and power ratio in different WPT bands for the controlled tongue action (circle, blue) and non-controlled interference (red, star) data sets for a single subject (adapted from [27])

Based on these fifteen features, two classifiers were tested, namely, a SVM and sMGC [27]. Although, the SVM performed better during offline testing, it was found that the sUGC had much better predictive accuracy when classifying in real-time. The degradation in performance when using interference data not specific to the subject, that is generalised interference, was found to be minimal. Specifically, for the sMGC the average accuracy across ten subjects was $97.8\% \pm 2.1\%$ (mean $\pm$ 1 SD) in the subject-specific case, compared to $96.4\% \pm 3.8\%$ in the generalised case. This implies that the subject is not required

to collect a prohibitively large interference training set alongside collection of their own personal tongue movement data. Instead, data previously collected from three subjects performing three of each of the eighteen speech, drinking and coughing actions have been used during IR training. The tongue action training data is created by pooling each of the training instances associated with each action into a single data set. For dual channel classification, the fifteen features from each channel are concatenated to form a single feature vector of thirty features. Based on the two training data sets, these thirty features can then be used to train the sMGC during real-time IR classification.

The efficacy of the IR subsystem can be attributed to its ability to discriminate higher frequency interference from lower frequency tongue signals. However, the vast plethora of potential interference implies that some interfering signals will have a similar frequency response as the tongue actions. This implies that, based on the current IR subsystem alone, many LF interfering signals would be incorrectly classified as tongue actions. The heterogeneous ensemble framework employed during inter-action classification provides an additional mechanism for rejecting interfering signals independent of their frequency content.

6.2 Testing strategy

During real-time evaluation of the TMEP system, the testing strategy shown in Fig. 6.2 was utilised. The main and obvious difference between real-time and off-line validation is that the test set is unknown prior to real-time prediction. Conversely, in off-line validation the data sets are collected prior to testing, can be labelled and thus their ground truth is known. Furthermore, error estimation can be performed using resampling techniques, such as cross-validation allowing the off-line data sets to be used in an efficient manner. To automatically and reliably evaluate the system in real-time, a visual stimulus based approach has been employed and allows the ground truth of the incoming test instances to be dictated by the computer. The drawback to using such a procedure is that it increases the cognitive load required by the user, as rather then just thinking of and performing a preferred action, an additional visual feedback step is required.

The testing strategy involves two distinct phases, namely a TMEP action phase and an interference phase. The TMEP action phase is randomised and involves a visual cue, consisting of a direction arrow, indicating to the participant one of either the left, bottom, right or top action to perform. To ensure that only one action is recorded during a TMEP action phase, when a TMEP event is detected and classified, it switches instantly to the interference phase. If no event is detected within the four seconds, then it is assumed that there was an attempt at a tongue movement but was rejected by the detection criteria thus leading to a missed action being recorded (FN^{-}). The interference phase lasts for

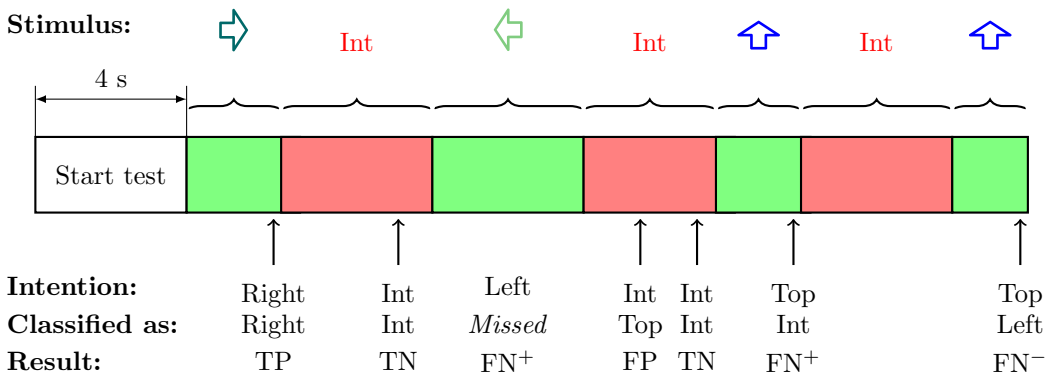


Figure 6.2: Overview and example of the testing strategy used during real-time evaluation of the TMEP system as represented by a timeline. Green blocks represent a TMEP action stimulus phase (left, bottom, right or top action) while red blocks indicate an interfering signal stimulus phase. During a TMEP action phase (green block) the participant is expected to perform only one action with the block switching after either a classification is made or 4 seconds has passed. An interference phase always lasts for 4 seconds with multiple decisions able to be made during this time. The vertical arrows give a typical sequence of TMEP events which could potentially occur, with the associated intention, classification output and result labelled below it.

four seconds regardless of how many TMEP events occur which stops the phases from constantly switching, thus preventing subject confusion and the dictated intention of the user being misinterpreted. A general interference stimulus is presented to the participant (consisting of the word 'interference') and indicates to them that any event occurring during this period will be regarded by the system as an interfering signal. During this time they are not obliged to perform an interfering signal but should they, for example, need to clear their throat then this would be the time to execute such an event.

To evaluate the real-time performance, the following collection procedure has been proposed. Four randomised test runs per individual with each test run consisting of presentation of ten of each TMEP action stimulus phase (lasting approximately six minutes per test run). The test runs are considered in two pairs, with each pair associated with a particular training set. The training sets are collected prior to the test runs and consist of thirty of each of the actions used. The interference training sets incorporated 162 instances of generalised data collected previously from three independent subjects. This consisted of the eighteen types of interference described previously. To ensure that the subject is accustomed to the test strategy, 1-2 practise runs are performed prior to the actual testing. This ensures that the participant understands how the test strategy works thus inhibiting the effect that the stimulus based approach has on the results. During this time, the detection parameters can be modified by a test instructor to ensure that the majority of TMEP actions are being detected. Having said this, the detection parameters are generally fixed

across both test runs and subjects and were set within the following ranges: $W_E = 0.2$s (1600 samples), $T_L = 3 - 5$, $T_U = 1000$, $D_L = 0.1 - 0.2$s, $D_U = 0.6 - 0.8$s. All data was collected using two earpiece channels with the subject sat comfortably in front of a screen displaying the stimuli in a standard office environment. A single recording session took approximately between 1-1.5 hours including short breaks between trials taken at the participants discretion. The sampling frequency was initially set at 8 kHz and when a TMEP event was detected, the segmented signal was digitally downsampled to 2 kHz using an anti-aliasing lowpass FIR filter. This meant that frequencies between 1-4 kHz wouldn't alias the signal. A standard and cost effective 'off-the-shelf' soundcard was used for data recording (see Section 3.4.1).

6.3 Results

6.3.1 Participants

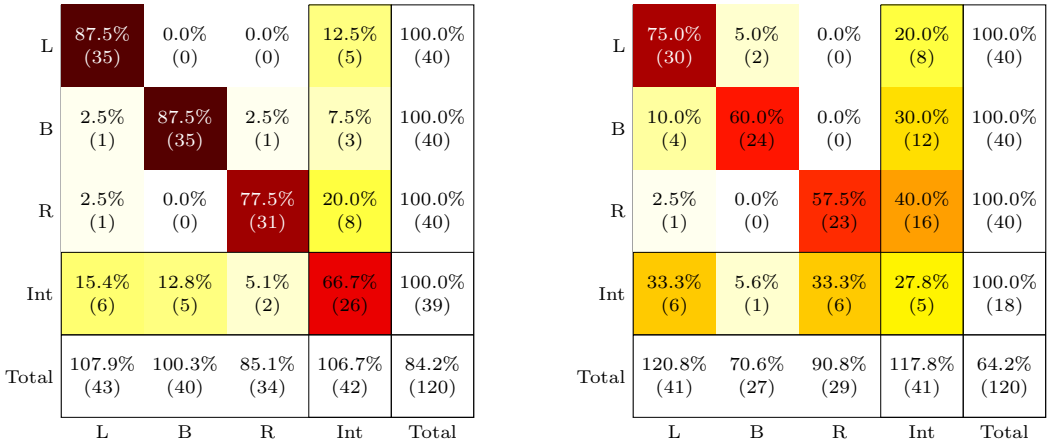
For real-time validation of the TMEP system, six healthy subjects (all male) ranging in age from 17 - 32 were recruited for testing. The experiment was approved by the local research ethics committee. Participants gave their informed consent before taking part in the study. Compared to off-line testing, performance of the classification system is expected to drop when used in a real-time setting. The pressure of having to perform tongue actions in situ and without any prior knowledge of the incoming test signals can have a significant effect in this regard. The addition of the IR rejection subsystem can also lead to dismissal of TMEP actions especially if they have high intra-class variability. A possible reason for this is due to the subjects being less familiar with that specific tongue movement. Learning three instead of four actions, considerably reduces the time required by the subject of becoming sufficiently trained in performing the actions. Furthermore, the expected performance of the system should increase when classifying between only three actions. This is at the obvious cost of a reduction in the size of the instruction set. However, three distinct outputs can still give a high degree of control especially when compared to asynchronous BCI systems, for example [1, 201]. Out of the six subjects, all were trained to perform three actions of their choice while four subjects went on to learn all four actions. This allowed for comparison between three and four action classification performance whilst also reducing the overall practise time required. S1 and S2 were instructed to perform three actions to a similar level to the minimally trained subjects used during off-line data collection while S3-S6 were instructed to perform all four actions in a similar manner to the well trained subjects (see Section 3.5.2).

6.3.2 Three-action classification

Figure 6.3 presents confusion matrices for three action real-time discrimination for the six subjects. They follow the same layout as the example confusion matrix given in Fig. 5.2 with $M = 3$. The row space indicates what action the subject intended to perform while the column space indicates what class a particular signal was assigned to. Therefore, each element of the matrix indicates both by percentage and number of instances (given in parentheses), the probability that a signal from class-m was classified to class-c. Additionally, the bottom row highlights the average percentage of instances classified to a particular class with the sum given in brackets. The far right column is associated with the total number of test instances collected from each class and highlights the number of test instances that equate to 100% of the class-specific data set. The bottom right element shows the mean classification accuracy across the tongue actions with the number of TPs given in parentheses. This percentage is therefore equivalent to SN^{+} (Equation 5.2). Each element of the confusion matrix is coloured according to the given percentage, with darker colours indicative of higher classification accuracies and therefore better performance. The actions used by each subject are indicated by the labels of the confusion matrices (L - left, B - bottom, R - right, T - top) while the interference class is indicated by Int and the border.

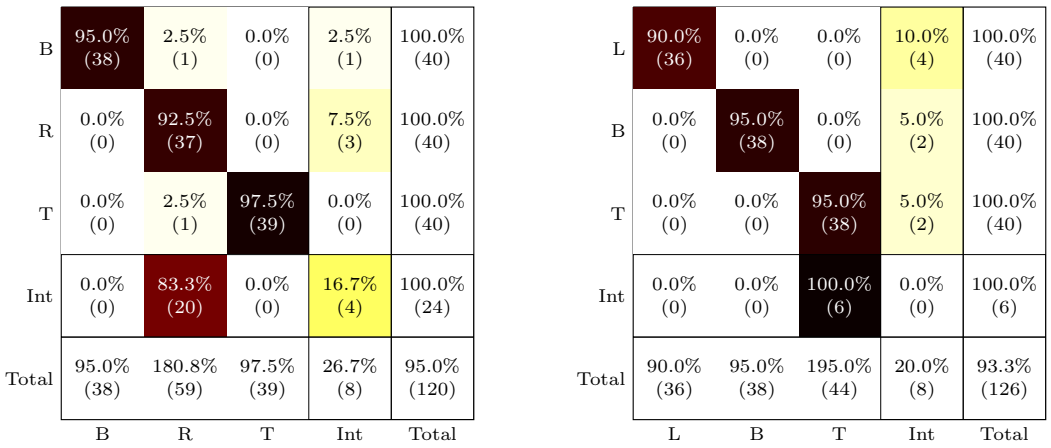
Subjects S1 and S2 had the least amount of practise in making the actions prior to testing and is evidence when comparing them to the other subjects. S1 achieved $SN^{\pm} = 84.2\%$ while S2 achieved only 64.2% in this respect. The remaining subjects all achieved over or close to $SN^{\pm} \geq 90\%$. The main reason for this weaker performance from S1 and S2 is due to an elevated rejection rate rather than misclassification, as corroborated by SN^{+} of 97.1% and 91.7%. The SN^{+} results for the remaining subjects (S2,S3,S4,S5) are 98.3%, 100%, 99.15% and 100%, respectively. This shows that in the case of S4 and S6, no tongue actions were misclassified incorrectly as another action, while two or less tongue actions were misclassified for S3 and S5. Therefore, in environments where no response from the system is not a critical issue, for example, controlling domotic applications, these accuracies highlight the potential for effective control using only three tongue movements.

During real-time inter-action testing, the rejection of interfering signals was generally poor and below 50% for four of the six individuals. The number of interfering signals varied considerately subject-to-subject from between 3-39 instances. The poor performance is attributed to two main factors, the type of signal executed and the testing procedure itself. The type of interference signals were generally movement artifacts, swallowing or gulping like signals and are associated with the subject moving their head in a particular manner or removing excess saliva from the oral cavity. These type of signals are low frequency in nature and have a similar TMEP representation as the tongue movements. This makes them extremely hard to discriminate from the movements themselves and as such they



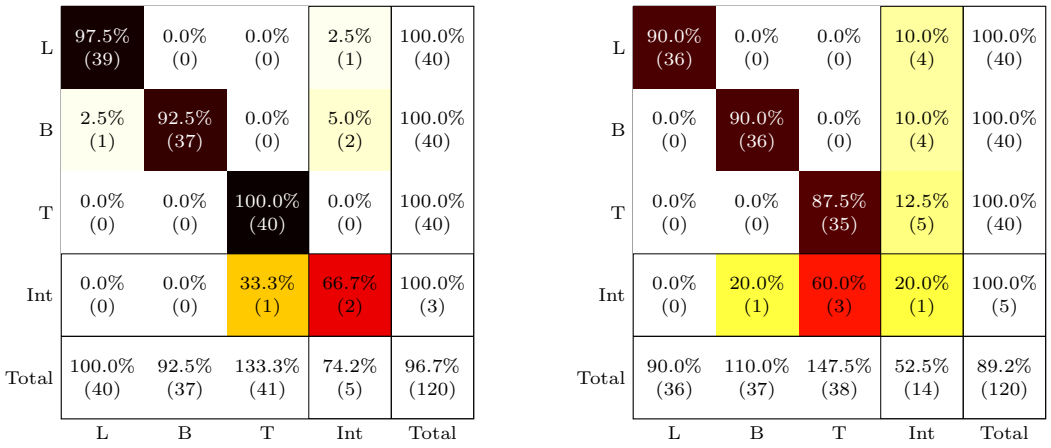
(a) Subject S1

(b) Subject S2



(c) Subject S3

(d) Subject S4



(e) Subject S5

(f) Subject S6

Figure 6.3: Real-time confusion matrices for three class inter-action discrimination. Confusion matrices based on Fig. 5.2 and described in Section 5.3.1

can often be misinterpreted. Additionally, it was observed that the subjects would often accidently perform a tongue action during the interference phase of the testing procedure. This was either due to the subject getting confused or the delay between the individual thinking about making the movement and the movement being classified. It often occurred while the stimulus switched from the action to interference phase. This could also occur during the interference-to-action switch but was generally more infrequent, as the interference phase was a set duration of four seconds. For the reasons stated, these results are not useful indicators of the IR rejection capabilities of the system so instead full interference testing based on challenging off-line data-sets is presented for the same subjects in Section 6.3.4.

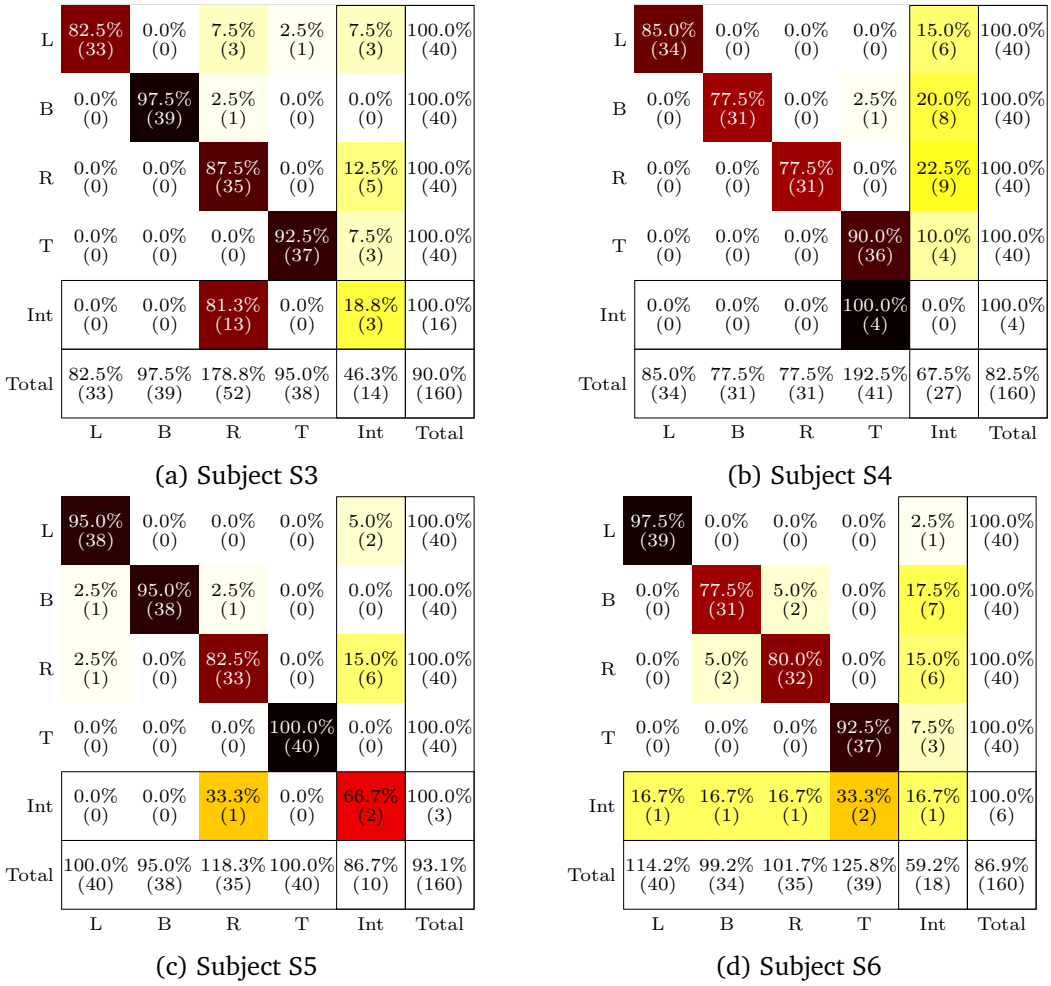
6.3.3 Four-action classification

The confusion matrices for four action real-time discrimination for the four trained subjects are presented in Fig. 6.4 and follow the same layout as previously but with $M = 4$.

As expected, the classification performance when discriminating between four actions dropped but still achieved $\text{SN}^{\pm} \geq 80\%$ across all subjects and over 90% for two of the subjects, namely S3 and S5. Again this is mainly due to an elevated rejection rate, with $\text{SN}^{+} =$ 96.6%, 99.2%, 98.0% and 97.2% for subjects S3 - S6. This shows that very few tongue actions were actually mislabeled in real-time to one of the three remaining incorrect classes. The reduction in performance when discriminating between four rather than three actions can be attributed to the increase in ambiguity due to the addition of a weaker action. This implies that the intra-class variance of this additional action is higher than the other actions and as such is more likely to be confused. In three out of the four subjects, the additional action was the worst performing action. There is also an elevated risk of the other actions getting misclassified to it, as highlighted by subject S6. Regardless of this, these results indicate that a practised individual can classify between four exclusive TMEP actions for control of assistive or robotic peripheries. This effectiveness is maximised if the application is not critical to FN^{-} responses, whereby it achieves over 95% sensitivity across the four subjects.

6.3.4 Interference rejection testing

Although the proposed testing strategy permits the possibility of executing and therefore classifying interference signals, the user was not obliged to perform any. The resulting real-time interference set were both poorly represented and negatively biased due to subject confusion. During a test run, the type of interference that a user performs would generally be of a similar type, for instance, swallowing. Therefore, forcing the user to perform interfering signal of their choice would result in similar types of signals and not



(a) Subject S3

(b) Subject S4

(c) Subject S5

(d) Subject S6

Figure 6.4: Real-time confusion matrices for four class inter-action discrimination. Confusion matrices based on Fig. 5.2 and described in Section 5.3.1.

be representative of all real-world situations. Additionally, given the expansive and diverse nature of potentially interfering signals, real-time collection of interference signals through a stimulus based approach would be inhibitive to perform across all subjects. As mentioned previously, the WPT features associated with interference have similar distributions across subjects indicating that generalised subject independent interference training sets can be used [27]. Furthermore, it can be considered infeasible to expect every subject to produce a large and diverse set of interfering signals during real-time TMEP action testing. Therefore the use of generalised subject independent interference signals is proposed for both testing and training the IR capabilities of the TMEP system. Data-sets based on different levels of complexities have been collected previously and are described alongside their content in Table 6.1 [27]. The third data set represents a new and very challenging subset of interfering signals due to their frequency content and morphology closely resembling that of the TMEP actions.

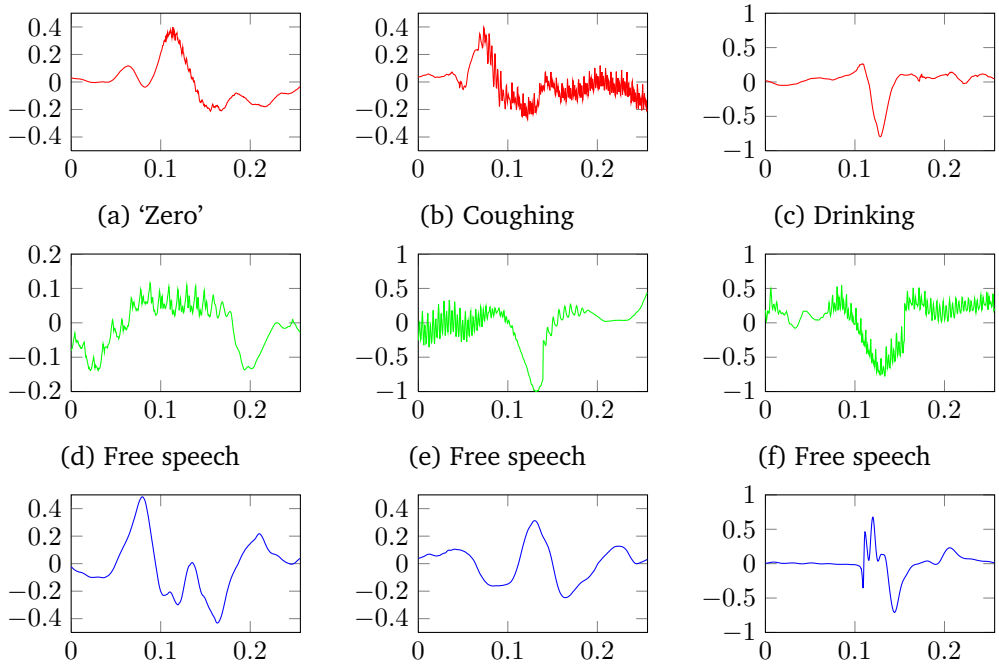
IR data set	Description	Nr. of subjects	Nr. of instances	Content
IR1	Standard interference	5	900	'0'-'9', 'open', 'close', 'on', 'off', 'start', 'stop', drinking, coughing
IR2	Free speech interference	3	540	1 minute conversation, 5-min newspaper reading, swallowing, drinking, coughing
IR3	Lower frequency interference	2	520	Heavy breathing, moving lips, up-down/left-right head movements, sucking, tapping on forehead , left/right temple taps, chin taps, sternum taps, chewing, swallowing, drinking

Table 6.1: Description and content of the three levels of interference tested. Words enclosed in apostrophes indicate that it is spoken otherwise the action is performed

Figure 6.5 shows nine segmented examples of typical signals from each of the IR data sets. The IR1 data set is the standard interference type as it represents the same type of signals that the IR subsystem is trained on. In all cases, the training and testing data-sets are from different subjects. Although predominantly composed of higher frequency interference, such as speech and coughing, a drinking type signal has also been included. Its inclusion during the training process increased the ability of the IR subsystem to reject similar low frequency interferences, assuming they show significant dissimilarity from the tongue signals in term of the WPT statistical features. During testing, these are the signals which have the highest probability of confusing the IR subsystem. The free speech signals (IR2), are predominantly unknown segmented speech signals collected from eighteen minutes of continual recording from three subjects. The subjects were also permitted

to drink, swallow and cough during this recording although this activity was generally rare. This data set provides a comprehensive subset of high-frequency interfering signals that can occur during real-time continual TMEP recording. The final data set is an attempt to capture a subset of signals that represents the most challenging interference the TMEP system would encounter in real-world situations. It comprises different types of low frequency interference, including internal interferences, movement artifacts and external interferences. The most common cause of internal interferences which were found to corrupt the TMEP signals were generally generated from within or through the oral cavity. These include signals associated with breathing, swallowing, drinking and eating. Head motions can also cause signal interference with impulsive signals generated due to vibrations of the connecting wires. The external interferences were all created by tapping various parts of the human anatomy, that is the forehead, temples, chin and sternum. This was done using a single finger which was tapped once against the aforementioned anatomy. Although, these types of signal are unlikely to occur as frequently as they are represented within the LF data set, they exhibit very similar morphologies as the TMEP due to there impulsive generation and therefore are highly likely to be confused as intentional behaviour. It should be noted that the detection parameters were relaxed to ensure each instance was segmented. This implies that many of these signals would actually be rejected by the detection thresholds prior to any classification being required.

The IR subsystem, as described previously in Section 6.1, was trained using the same standard interference type as utilised during real-time testing. An additional two inter-action training sets were collected from each subject so that there were a total of four training sets per subject and action subset. The training was performed in exactly the same manner as during the real-time testing procedure. Each of the data sets, namely, standard, free and LF consist of two corresponding channels of TMEP interference segments. Each data set was initially passed into the IR subsystem associated with a particular subject and inter-action training set and the number of rejected instances recorded. The remaining instances were then passed through the corresponding inter-action ensemble classifier and the number of rejected instances at this stage of the processing was also recorded. The remaining instances could then be considered FPs and would be classified to one of the M classes. Figure 6.6 shows results for the six subjects during three-action discrimination while Fig. 6.7 gives results for the four subjects during four-action discrimination. The results are presented as separate bar charts associated with each of the subjects and indicate mean percentages of the IR subsystem rejection rate and the inter-action ensemble rejection rates of the remaining instances. The percentage of instances that are not rejected by either mechanism are therefore considered the FP rate. Each of the interference data types (standard, free and LF) used during testing are shown as separate bar groupings with the left bar indicating the IR subsystem rejection rate (blue bars with diagonal lines),



(a) 'Zero' (b) Coughing (c) Drinking

(d) Free speech (e) Free speech (f) Free speech

(g) Swallowing (h) Chewing (i) Forehead tap

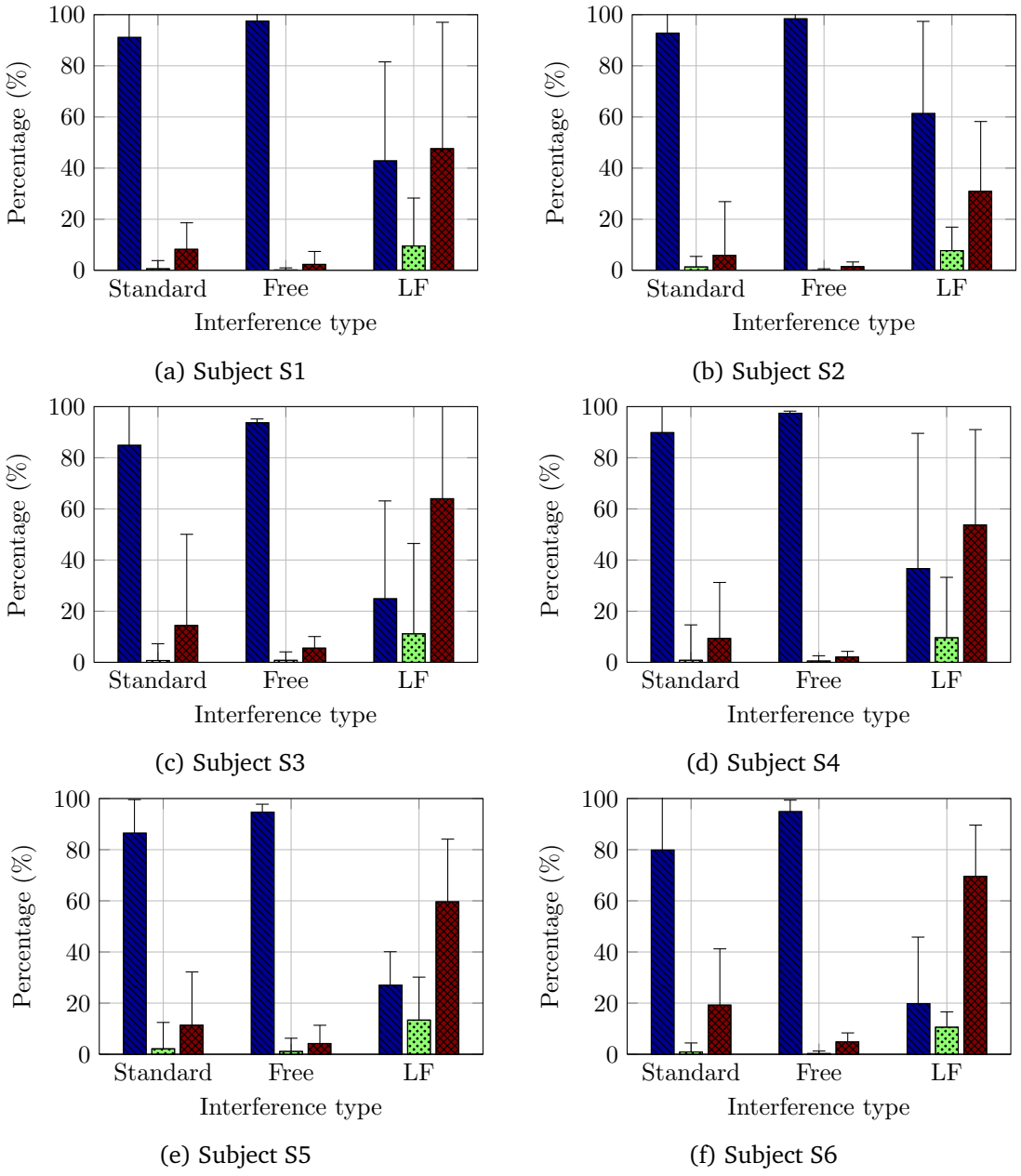
Figure 6.5: Examples of segmented interference signals from each interference type. Top row (red) - standard interference, Middle row (green) - free speech and bottom row (blue) - low frequency interference.

the middle bar indicates the inter-action ensemble rejection rate (green dotted bars) while the FP rate is indicated by the right bars (red crosshatched bars). The standard deviation across the four training sets associated with each result are shown by the error bars.

Across all subjects and for both three and four action discrimination, rejection of the free interference data is the greatest. The average false positive rate of the free speech interference is $3.4 \pm 1.7\%$ and $3.0 \pm 0.8\%$ during three and four action discrimination, respectively. The majority of this interference is rejected by the IR subsystem which has an average rejection rate of $96.3 \pm 1.9\%$ and $96.0 \pm 1.8\%$. Having said this, the inter-action ensemble does play a role in rejection, with a rate of $0.5 \pm 0.6\%$ and $1.2 \pm 1.2\%$ which equates to $13.6 \pm 12.2\%$ and $24.3 \pm 15.3\%$ of the remaining instances that were misclassified by the IR subsystem. The effective rejection by the IR subsystem in the presence of standard and free interference is due to the higher frequency content that is present in speech and normal conversation. This indicates that it is unlikely that the TMEP system will respond unintentionally when used in situations where the user is required to speak and/or converse. This is an important property of any bioacoustic HMI concept, as speech is inherent to daily human living and therefore one of the most common sources of signal corruption, similar in nature to the 'midas-touch' issue.

Rejection of the standard interference data, although still demonstrating good performance is lower than the free interference. On average, the IR subsystem rejected $87.5 \pm 5.1\%$ and $85.0 \pm 6.1\%$, the inter-action ensemble rejected $1.1 \pm 1.0\%$ and $1.8 \pm 2.0\%$ (equivalent to $8.8 \pm 7.3\%$ and $12 \pm 6.1\%$ of signals misclassified by IR subsystem) and a false positive rate of $11.4 \pm 5.0\%$ and $13.2 \pm 5.0\%$ for both three and four action discrimination. In the case of subject S5, the ensemble system rejected over half of the instances that the IR subsystem misclassified during four class discrimination. Although this data represents the same subset of instances that the IR subsystem was trained on, the presence of the LF drinking signal tends to cause confusion due to its similarity to the tongue actions. The subject-to-subject differences within the inter-action training sets imply that some subjects perform better with regard to the ensemble rejection. Ensemble rejection is based on the assumption that the features extracted from the data will be assigned random classes across the various classification strategies and channel inputs. Additionally when more actions are being classified, the diversity of the labels being assigned by the individual ensemble members also increases leading to higher ensemble rejection rates for four as opposed to three action discrimination, and is the case for all three interference types.

For rejection of LF interference, the performance across all rejection mechanisms is reduced, with the IR subsystem achieving average rejection rates of $35.4 \pm 15.5\%$ and $27.8 \pm 11.6\%$, ensemble rejections rates of $10.3 \pm 3.9\%$ and $22.1 \pm 8.9\%$ and FP rates of $54.2 \pm 14.2\%$ and $53.1 \pm 7.3\%$ for three and four action classification. The ensemble framework exhibited



(a) Subject S1

(b) Subject S2

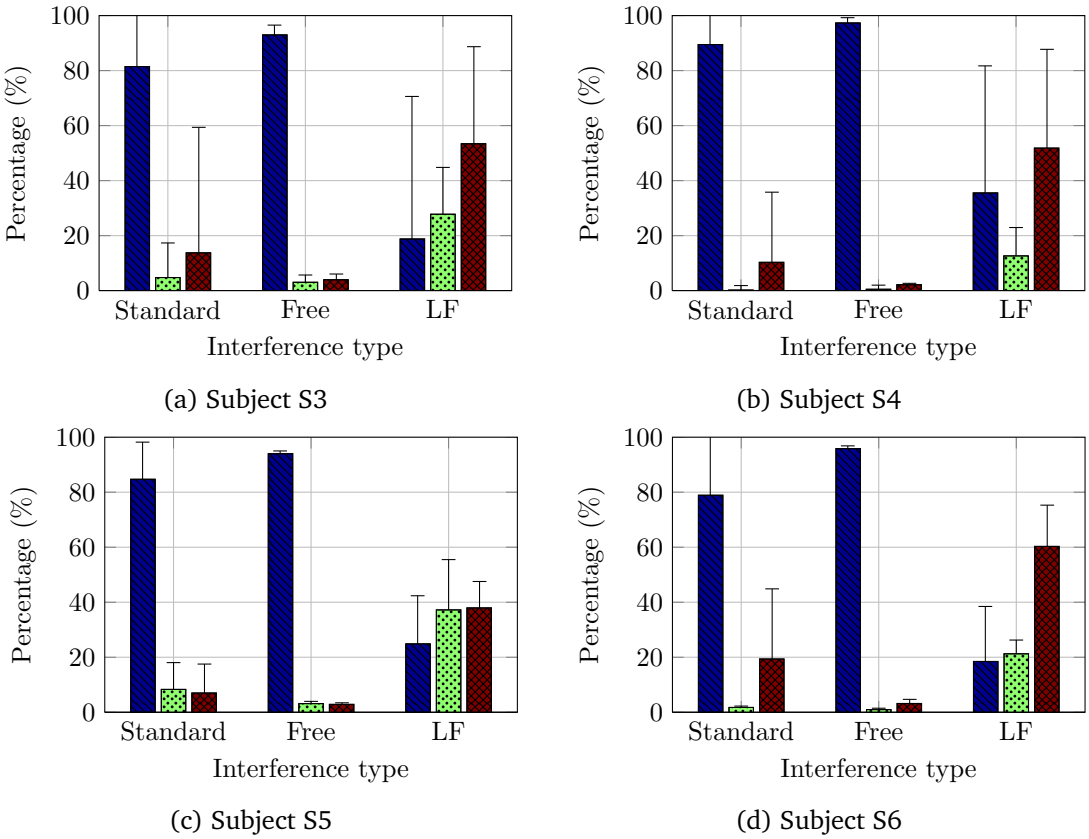
(c) Subject S3

(d) Subject S4

(e) Subject S5

(f) Subject S6

Figure 6.6: Interference testing using real-time three action training data and off-line data sets for six subjects. The blue bars with diagonal lines indicate the average percentage across training sets for each type of interference that was rejected by the IR subsystem. The green dotted bars indicate the percentage of interference rejected by the heterogeneous ensemble classifer. The red crosshatched bars are the remaining percentage of actions incorrectly classified as tongue actions (FP rate). The error bars indicate the SD across the four trials.



(a) Subject S3

(b) Subject S4

(c) Subject S5

(d) Subject S6

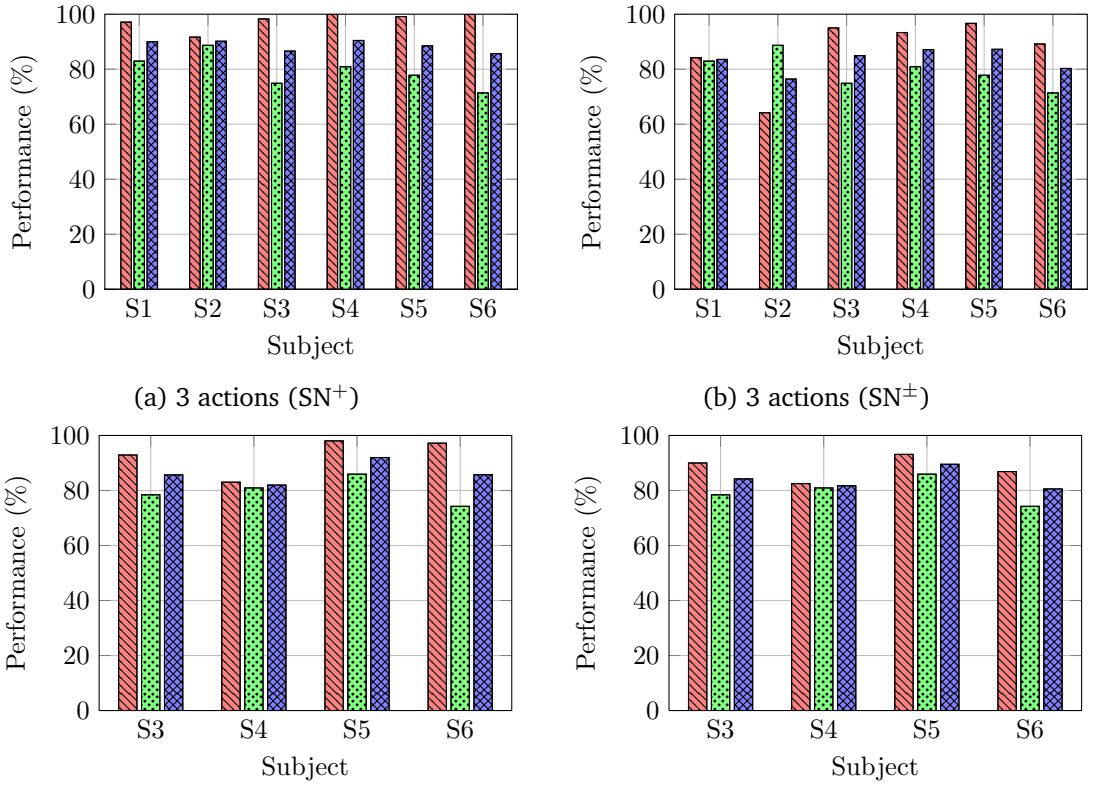
Figure 6.7: Interference testing using real-time four action training data and off-line data sets for four subjects. The blue bars with diagonal lines indicate the average percentage across training sets for each type of interference that was rejected by the IR subsystem. The green dotted bars indicate the percentage of interference rejected by the heterogeneous ensemble classifer. The red crosshatched bars are the remaining percentage of actions incorrectly classified as tongue actions (FP rate). The error bars indicate the SD across the four trials.

increased rejection performance especially in the case of four action discrimination with an average rejection rate of $16.4 \pm 5.5\%$ and $26.0 \pm 6.4\%$ of the remaining LF instances misclassified by the IR subsystem. The large increase in the FP rate is due to the LF interference posing a much more challenging classification problem. In reality many of these signals wouldn't be encountered due to initial rejection from the detection block (based on tighter detection thresholds) as well as infrequent overall occurrence of these challenging signals. Having said this, even under these difficult conditions the system is able to reject $45.8 \pm 14.2\%$ and $47.0 \pm 7.3\%$ of the LF interference. Furthermore, without the rejection capabilities of the ensemble, the rejection rate would only be equal to the IR subsystem rejection rate, indicating that the ensemble framework significantly increases the overall rejection capacity.

6.4 Discussion of results

The overall performance of the real-time TMEP system can be evaluated by combining the previous results. This can be expressed using the statistical performance measures, namely, sensitivity, specificity and accuracy, as outlined in Section 5.3.1. It is useful to present results pertaining to both SN^{+} and $SN^{\pm}$ as the relevance of each measure is dependent upon the application. A non-critical application can be considered one where a rejected command doesn't impact negatively in terms of the response of the peripheral. Instead, if the command is rejected, it can be repeated with the only consequence being a delay in the response of the device. An example of such a system would be navigating a computer menu using up, down and select commands. On the other hand, in safety critical applications, a rejected classification (when a command is issued) can lead to negative consequences in the state of the peripheral. For example, if controlling a wheelchair and a stop command is issued, it is generally imperative that the wheelchair brakes at the time this first command was issued. This can be thought of in terms of whether the issuing of a command is time critical which is generally associated with the control of dynamic systems, such as wheelchairs. Therefore, the SN^{+} measure, ignores incorrect classifications to the negative interference class while $SN^{\pm}$ accounts for them. Figure 6.8 highlights collaborative results based on results from the previous section for the six subjects for both three and four action discrimination. These are presented in terms of SN^{+}, SP, CA^{+} (left plots) and $SN^{\pm}$, SP, $CA^{\pm}$ (right plots). Results relating to three action classification are presented on the top row while four action classification results are given along the bottom row. Classification accuracy has been calculated based on the assumption that interfering signals occur as frequently as intentional actions and therefore positive and negative cases both have a prior probability of 0.5.

Apart from subject S2, all subjects show high sensitivity (above 80%) for both three



(a) 3 actions (SN^{+})

(b) 3 actions ($SN^{\pm}$)

(c) 4 actions (SN^{+})

(d) 4 actions ($SN^{\pm}$)

Figure 6.8: Real-time sensitivity (red with diagonal lines), specificity (green with dots) and accuracy (blue with crosshatch pattern) results for three and four action classification for each individual subject.

and four action classification. The main cause of subject S2 having a lower $SN^{\pm} = 64.2\%$ is from rejection rather than misclassification as highlighted by a much increased $SN^{+} = 91.7\%$. This can be attributed to subject S2 being inexperienced in making the tongue actions and therefore the signals having higher intra-class variance. All other subjects during three action discrimination exhibit similar sensitivity differences with the average $SN^{+} = 98.9 \pm 1.2\%$ across the remaining subjects (S1,S3-S6) compared to $SN^{\pm} = 91.7 \pm 5.0\%$. This indicates the role that the ensemble classifier plays during real-time inter-action classification. Instead of potentially misclassifying ambiguous instances as a single classifier would, it rejects uncorroborated decisions. In non-critical applications, such a mechanism is extremely beneficial and prevents the interface from being an ineffective communication tool due to too many misclassifications occurring. As expected, when comparing three against four action discrimination, there is a drop in sensitivity across the four subjects S3 - S6. Furthermore, the differences between SN^{+} and $SN^{\pm}$ are less pronounced indicating that inter-action misclassification contributes more heavily to this drop in performance. The highest specificity (SP) is exhibited by subject S2 and is due to increased rejection of the LF interference by the IR subsystem as indicated earlier in Fig. 6.6. This is attributed to the subject being the least proficient in making the tongue movements, which makes them less defined and therefore generally lower frequency in nature. The benefit is that the drop in their nominal frequency implies that the IR subsystem is more likely to reject lower frequency interfering signals.

To highlight these differences further, Fig. 6.9 shows mean $\pm$1SD results for both three and four action classification from the four experienced subjects. This is shown in terms of both SN^{+} and $SN^{\pm}$ with only subjects S3-S6 considered due to participation in both three and four action testing. The specificity for three and four action discrimination are $SP = 76.2 \pm 4.1\%$ and $SP = 79.9 \pm 4.9\%$. The increase in specificity during four action recognition is attributed to increased ensemble rejection as the additional action increases the probability of interference being classified differently across the ensemble members. This would have the effect of increasing the risk of misclassifying negative instances although the effect is somewhat negligible. The sensitivities for three action recognition are $SN^{+} = 99.4 \pm 3.2\%$ and $SN^{\pm} = 93.5 \pm 0.8\%$ while for four action recognition $SN^{+} = 92.8 \pm 6.9\%$ and $SN^{\pm} = 88.1 \pm 4.5\%$ respectively. This highlights that very effective classification between three and four tongue actions can be achieved in real-time. Furthermore, the majority of positive class errors during three action classification is nearly all from rejection rather than misclassification. For four action recognition, high sensitivity is still achieved with the additional action increasing the instruction set associated with the HMI concept. The drawback is that this additional action also decreases the inter-class separability, leading to increased misclassification among the positive instances. Assuming equal prior probabilities of volitional and interference actions occurring, the average accu-

racies for three action real-time classification are $CA^{+} = 87.8 \pm 2.1\%$ and $CA^{\pm} = 84.9 \pm 3.3\%$ while for four action discrimination $CA^{+} = 86.3 \pm 4.1\%$ and $CA^{\pm} = 84.0 \pm 4.0\%$.

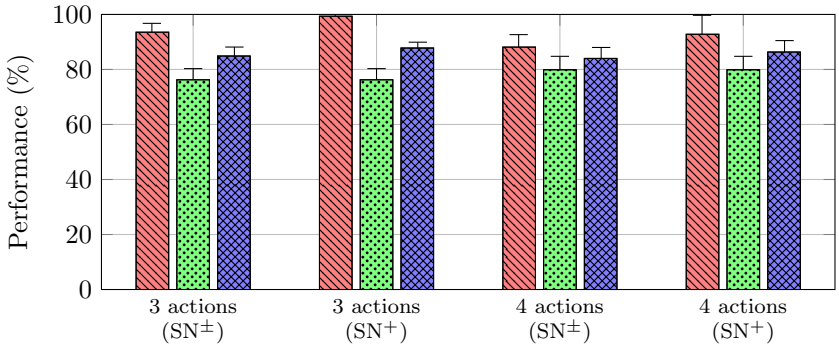


Figure 6.9: Mean real-time results for subjects S3-S6 in terms of sensitivity (red diagonal lines), specificity (green dots) and accuracy (blue crosshatch) results for three and four action classification across the four experienced subjects.

6.5 Comparison to alternative interfaces

Based on this real-time implementation, it is useful to compare the bioacoustic TMEP concept to alternative assistive HMI systems (described previously in Chapter 2). This can be done in terms of the *key drivers* introduced in Section 2.3. Figure 6.10 shows the relative position of the TMEP concept, as determined by the author, in terms of the ten *key drivers* and indicated by a black star overlaid on to the previous HMI visualisations (Figs. 2.13 - 2.17). For visual clarity, only the input method groupings are shown, that is, body, brain, biofeedback, gaze, oral, tongue based systems. The justifications for the relative positioning of the TMEP concept are now discussed in terms of each of these *key drivers*.

The ITR is a measure of the channel capacity associated with a communication system and can be estimated using Wolpaw's definition as follows [93]

$$\text{ITR} = \frac{1}{T}\left(\log_2 M + P\log_2 P + (1-P)\log_2\left(\frac{(1-P)}{(M-1)}\right)\right) \tag{6.1}$$

where M is the number of states (or classes), P is the probability that the desired command will be selected and T is the response time in terms of the command intention-generation latency. During real-time testing, the time between a stimulus appearing and a machine command being issued, was stored. Therefore, based on the four action results using $CA^{\pm}$, the average response time over the four subjects was $T = 1.53 \pm 0.47$ seconds. With $P = 0.84$ and $M = 4$, this gives an ITR$= 0.73$ bits/sec. The cost of the system is currently <£100 and could be reduced significantly if the system was commercialised. Based on these quantities, the TMEP can be positioned relatively to the alternative HMI concepts

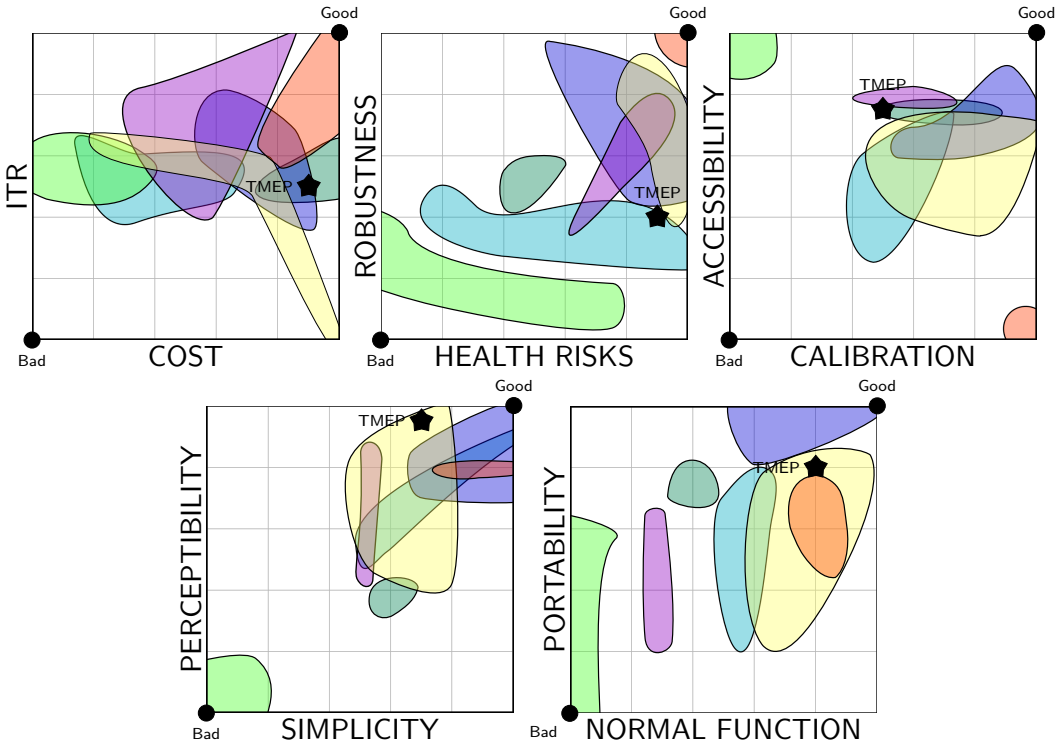


Figure 6.10: HMI comparison space showing the TMEP concept against the current state of art in assistive HMI described previously in chapter 2. The reasoning behind the positioning is described in the text

and as can be seen, performs comparably to the other tongue-based interfaces in terms of ITR and cost.

The remaining *key drivers* are positioned at the author's discretion and are justified as follows. Due to the system being completely noninvasive, there are no health or hygiene risks associated with donning the device. The robustness of the system has yet to be fully tested and therefore has been positioned relatively low in this respect, located lower than the other established tongue-based inputs. Subject-specific training data is required by the system and may need to be routinely updated, this implies there is an associated calibration period. Therefore, the TMEP concept has been located near similar interfaces with respect to this driver. The accessibility of the TMEP concept is the same as other tongue-based interfaces. Although there is a learning period associated with performing the actions, the donning of the device is very simple. Furthermore, the earpieces can be worn for an extended period of time whether or not the device is in use. This implies that once the actions are learnt, the device is simple to use, as well as being highly imperceptible to the user and those around them. As no hardware is placed within the oral cavity, the system doesn't affect an individuals ability to eat or drink. The system is also extremely effective at rejecting speech and therefore the system can be considered to not heavily interfere with the normal functioning of a person. The portability of the system is generally high, as it is easy to transport from place-to-place and although cannot be directly used when on the move due to movement artifacts, commands could be issued when the user became stationary again.

These visualisations help to highlight some of the many trade-offs that a user must deal with when choosing an assistive interface. The diverse nature and the wide-ranging abilities of the physically challenged imply that no single HMI system will fully address all or even the majority of their needs. This is without even considering factors such as personal preference and user-specific needs. Therefore, increasing the choice of interfaces available to this diverse user group is of upmost important. This is through not only leveraging alternative communication pathways but by also broadening the range and attributes of interfaces already available to the user.

6.6 Chapter summary

This chapter has principally been concerned with the real-time evaluation of the bioacoustic processing framework. The processing architecture has shown effective discrimination of both three and four action TMEP signals, using a dual-channel setup. The rejection capabilities of the system were further highlighted using three subject-independent off-line data sets of varying difficulty. An IR subsystem (not part of this research) was introduced into the real-time system and was able to reject high frequency interference with high

specificity [27]. For rejection of low frequency interference it is considerably less effective due to the signals having a similar frequency response as the tongue actions. However, the additional rejection capabilities of the heterogeneous ensemble were able to significantly increase the specificity in this case. This implies that not only does the ensemble framework provide better inter-action discrimination but also increases the rejection capabilities of the real-time system.

Chapter 7

Conclusions and future work

7.1 Thesis summary

The goal of this thesis was developing extraction and decoding techniques that enable volitional bioacoustic signals to be used as command to assistive HMI systems. Volitional bioacoutic signals have been highlighted as a novel method for providing communication links through the user (human-sensor interaction) and offer significant advantages over current methods based on neurophysiological signals. It epitomises the idea of 'disruption-free' signal capture, whereby the transducers can be positioned away from the source of the volitional activity. This is possible due to the penetrative and translational propagation of mechanical oscillations across the human body. The notion of 'disruption-free' sensor placement, allows the sensory hardware to be located in discreet anatomical locations, implying that a highly imperceptible and non-invasive HMI system can potentially be achieved. Such separation prevents the sensory hardware from interfering with the normal functioning of the chosen access pathway, for instance, the oral cavity. The bioacoustic concept is further endorsed, by enabling freedom in terms of the definition of the generating actions, with the only constraints being that an adequate SNR can be achieved and the actions themselves are distinct. Utilising impulsive commands, which have detectable and unique signatures in both the time and frequency domains, allows for rapid command of peripheries.

Based on this idea, the main focus has been on the systemisation of a processing architecture for translation of impulsive bioacoustic signals generated through volitional activity. This has included the formulation of novel techniques for online extraction and decoding of these biosignals into effective machine commands. Extraction is concerned with the isolation and exemplification of the intentional commands from continuous stream(s) of data. This has included subsystem processes of detection and segmentation to extract tractable segments of volitional activity and formulation of a robust template creation procedure. Development of a robust template creation allowed the temporal morphology

associated with each action type to be accurately estimated from a set of class-specific patterns.

Bioacoustic decoding has dealt with recognition between the various predefined actions, allowing user intention to be elucidated. This incorporated concepts of feature extraction, feature selection and class-specific classification, combined to form various discrimination strategies. Effective decoding of the extracted bioacoustic segments was achieved based on these strategies. The culmination of the decoding architecture (Chapter 5), was the formulation and implementation of a multi-channel heterogeneous ensemble, consolidated using preferential voting methods and rank-based fusion. This not only improved the inter-action decoding performance but provided an additional mechanism for natural rejection of interfering signals.

The proposed framework was implemented on a novel assistive HMI concept based on TMEP signals [24]. These bioacoustic signals are initiated from four prescribed tongue flicks, namely a left, bottom, right and top action, creating impulsive and distinctive activity. Signals are captured within the ear canals using simple microphones with associated housings, creating a completely noninvasive and imperceptible interface. The efficacy of the system has been evaluated in real-time, on six healthy subjects for both three and four action discrimination. A drawback of using bioacoustic activity is the relatively high risk of interference, such as speech, to merge with the signals in real world situations. Therefore, for bioacoustic signals to be a viable input for HMI applications, sufficient IR capabilities are required. To this end, a specific IR subsystem based on work in collaboration with Mamun et al. [27], has been implemented into the real-time system. Although effective at rejecting higher frequency interference, such as coughing, it is less effective at rejecting interference which have similar frequency responses as the tongue actions. When faced with such challenging interference, the IR capabilities of the heterogeneous ensemble significantly improved the rejection performance of the system. This real-time implementation demonstrated that volitional bioacoustic signals can be a useful control signal for assistive HMI applications using specially developed extraction and decoding algorithms. Furthermore, many of the developed techniques can be readily applied to other general biosignals, including for example, discrimination of EEG, MMG or EMG activity.

7.2 Achievements

Throughout the process of developing a translation framework for volitional bioacoustic signals, various novel extraction and decoding techniques have been formulated. In this section, the overall processes relating to these techniques are summarised including the challenges encountered. Particular attention is made in highlighting the novel details of the processing architecture with full details of the extraction and decoding techniques

found in chapters 4 and 5 respectively.

The first stage of bioacoustic extraction is based on a detection subsystem to flag when interesting activity occurs. Implementation of this subsystem is performed using the notion of parsimonious computation to ensure efficient real-time monitoring of the continuous waveform. This involves thresholding of the short-term energy contour associated with a signal channel and has previously been implemented within ASR detection subsystems [150]. Although, already a relatively efficient calculation, the computation required to calculate the short-term energy can be reduced significantly, through the use of a new online update law (Equation 4.6). This reduces the computation from $2W_E - 1$ to only two operations at each update. Following this detection subsystem, signal segmentation is performed which generates a tractable finite signal for further processing. To exemplify the time domain waveform associated with each volitional action, a new robust template creation process using randomised PPCCAT has been introduced. This is considered an important and general step during any of the proposed learning phases, as alignment of the waveforms to the templates prior to discrimination reduces the intra-class variance, based on the distinct morphologies associated with the impulsive activity. This template can be considered a central tendency estimate associated with each sample conditioned across the entire waveform. The robustness of this template creation process was assessed in terms of the jackknife standard error and absolute bias indicating increased stability of the produced templates compared to previous methods [24].

Bioacoustic decoding deals with interpretation of user intention expressed through predefined actions. Due to the complex and unknown nature of the physiological processes involved, this task is approached from a black-box perspective using tools extrapolated from the pattern recognition domain. Information pertaining to each of the actions is introduced to the decoding algorithms using labelled training data and as such, can be considered a supervised learning problem. Impulsive bioacoustic signals are highly non-stationary in nature and can exhibit large intra-class variance, especially if the user has had minimal training in performing the actions. Therefore, effective decoding strategies are required to reliably discriminate between the various signal classes. The subject-specific nature of the actions also implies that collection of a large personal training database is infeasible. As such, only small samples of data are available to the learning algorithms. This is further exacerbated by the need for large initial feature spaces, necessary to adequately represent the class-specific information associated with the individual actions. To this end, a general single-channel classifier model has been proposed. Central to this generalised model is the alignment process, whereby unseen test instances are initially aligned to class-specific templates. This implied that only class-specific classification models could be utilised. The remainder of the model was based on tools from the pattern recognition domain and incorporated three main steps, namely, feature extraction, feature selection

and class-specific classification.

A variety of feature extraction methods have been explored and include AR, DFT, DCT and PCA transforms. They extract discriminatory information relating to underlying general model parameters, frequency components and a representative de-correlated subspace. From these the most effective features were selected using feature ranking and subset evaluation strategies. This included implementation of various non-parametric ranking filters, such as the mRmRM. The subset of features are selected using either a wrapper or a simple filter metric based on the mean F-score. Stability of the feature set can also be an issue and as such, two stability mechanisms were introduced, namely, a weighting envelope and weighted sequential feature selection [97]. Class-specific classification models proposed, including a simple Euclidean distance classifier, matched filter, decision fusion classifier and naïve Gaussian Bayes classifier. The decision fusion classifier has been previously introduced and updated in this work with an additional Laplace correction term [24]. The Gaussian Bayes classifier is a widely used generative classification model implemented under univariate conditions thus alleviating issues associated with high dimensionality and small samples. The parameters of this model are estimated from either standard statistics including the mean and standard deviation, or robust statistics, such as the median, median absolute deviation and interquartile range. Comparison across the discrimination strategies was evaluated using four action off-line TMEP datasets collected from fifteen subjects. The Friedman test was used to assess the familywise significance of the mean cross-validation accuracy, across the fifteen subjects. Post-hoc comparisons were performed using the Nemenyi test and results visualised using critical difference diagrams [193].

By consolidating decision outputs from a variety of discrimination models, an overall increase in the classification performance can be achieved. Even if the ensemble does not perform better than the best individual classifier it reduces the risk of making poor selections during generalisation [195]. A further benefit of using such a meta-learner is its ability for fusion of multi-channel inputs and any number of base classifier models to be incorporated into the ensemble. To this end, a heterogeneous ensemble classification framework was formulated and implemented, allowing consolidation of the output labels from a variety of base classifier models. Within this general framework, there is no restriction on the type of base classifier model that can be incorporated and the number of heterogeneous source channels that can be included. Diversity amongst the members is considered key to the efficacy of the ensemble [195]. The heterogeneity of the ensemble provided a mechanism for achieving diversity and meant that a smaller training sample could be employed, as diversity wasn't solely data-driven. Pairwise diversity measures were used to highlight the potential of combining the decision outputs across the proposed base models and are based on 2×2 diversity matrices. The process of obtaining such a

matrix, from the pairwise class-specific joint multinomial probability matrices was also illustrated. During real-world use, there is a relatively high probability of interference, both internal and external to the user, merging with the bioacoustic signals in real-time. This unintentional activity can cause confusion by incurring a significant number of false positives within the system. By employing voting methods, such as the majority vote to fuse the decisions of the individual ensemble members it provides an additional natural mechanism for interference rejection, thus increasing the specificity associated with real-time performance.

The efficacy of the proposed heterogeneous ensemble and general translational methods was demonstrated by implementing them into a real-time assistive HMI system. This was based on bioacoustic TMEP signals, which present a completely noninvasive and imperceptible method for capturing volitional tongue movements. A specific IR subsystem based on work in collaboration with Mamun et al. [27], was also implemented into the real-time system and provides a 'first line of defense' for rejection of non-volitional action segments. This is based on frequency based features, specifically, statistical properties of wavelet packet coefficients which are effective at discriminating higher frequency interference from the lower frequency tongue signals. A stimulus based real-time testing strategy was devised, retaining the asynchronous nature of the control strategy, while simultaneously attempting to maintain real-life conditions as much as possible. Three off-line data-sets, each portraying varying levels of difficulty were utilised to fully examine the interference rejection capabilities of the framework. The final data set represented a particularly challenging subset of interference due to low frequency signatures and similar morphologies to the volitional tongue activity. Across six subjects, three-action discrimination achieved sensitivities of $\text{SN}^{+} = 99.4 \pm 3.2\%$ and $\text{SN}^{\pm} = 93.5 \pm 0.8\%$ and interference rejection specificity of $\text{SP} = 76.2 \pm 4.1\%$. For four-action discrimination, $\text{SN}^{+} = 92.8 \pm 6.9\%$ and $\text{SN}^{\pm} = 88.1 \pm 4.5\%$ while $\text{SP} = 79.9 \pm 4.9\%$ across four experienced subjects. Furthermore, during four-action discrimination the ensemble framework rejected $22.1 \pm 8.9\%$ of the LF interference incorrectly classified by the specific IR subsystem. This highlights that the additional IR mechanism provided by the ensemble framework, substantially reduces the number of false positives that these challenging LF interfering signals would otherwise produce.

To the best of the author's knowledge, this is the first tongue-based assistive HMI that requires no sensory hardware to be placed within the oral cavity. This is possible due to the capture of bioacoustic signals generated through impulsive flicks of the tongue. These distinct mechanical oscillations propagate throughout the maxillofacial and Eustachian tube anatomies, allowing them to be sensed with a simple microphone positioned discreetly in the ear canal. This leads to a highly imperceptible and completely noninvasive HMI concept with additional benefits of portability and low cost. For effective extraction

and decoding of the volitional signals from normal activity, several challenges had to be surmounted. Firstly, impulsive bioacoustic signals are nonstationary in nature and have a certain level of intra-class variability. They also contain redundant discriminatory information and the initiating actions generally require subject training prior to operating the system. For less trained individuals, the discriminatory information can be weaker and thus hidden within the signals. Another challenging aspect is the size of the training set, which is limited, to increase the useability of the concept. In normal use, there is a high chance of interfering signals merging and masquerading as volitional activity. Such interference can be produced from both internal and external sources and come in a variety of guises. Considering these characteristics associated with volitional bioacoustic activity, a processing framework has been proposed in this thesis enabling effective translation of this information into useful machine commands. Furthermore, the proposed extraction and decoding architecture enable accurate discrimination between impulsive actions and effective rejection of a wide range of challenging interference. Although erroneous behaviour cannot be fully eliminated, for non-critical applications, real-time command and control of assistive devices using the notion of 'disruption-free' bioacoustic activity can be achieved with a high degree of accuracy.

7.3 Conclusions

7.3.1 Summary of contributions

The summary of contributions, highlighted at the beginning of this thesis, are reiterated below and their location within the body of text indicated.

- The systematisation and development of an architecture for extraction and decoding of volitional bioacoustic activity with applications in the field of assistive HMI. This is the general theme of the thesis and therefore features throughout. The translation has been split into two key concepts, namely, bioacoustic extraction (Chapter 4) and bioacoustic decoding (Chapter 5).

- Introduction of a new robust template creation process based on randomisation of a partial pairwise cross-correlation averaging tree (PPCCAT). This is a generalised extraction process used as a precursor to and during classification of volitional activity in HMI systems. It is specifically implemented in the context of a bioacoustic TMEP application. This is described and evaluated in Chapter 4.

- Formulation and implementation of a multichannel ensemble framework based on heterogeneous classifier models. This enhances classification performance while enabling natural interference rejection through preferential fusion strategies. The proposed implementation is described and evaluated in Chapter 5.

- Real-time implementation of the proposed bioacoustic architecture and evaluation on an imperceptible HMI system using completely noninvasive TMEP signals. This included both three and four action discrimination and evaluation of the interference rejection capabilities under challenging conditions. This is the focus of Chapter 6.

7.3.2 Suggested future directions

While the ability to discriminate volitional bioacoustic activity, specifically TMEP signals, has been demonstrated, there are many extensions to the proposed bioacoustic processing framework and research area in general. This can lead to enhanced performance in the proposed domain and wider applicability of the framework to a broader range of biosignals.

The main focus of the decoding algorithms has been towards effective inter-action discrimination. The separate IR subsystem developed concurrently to this research is effective at rejecting higher frequency interference. To help combat the merging of lower frequency interference, the proposed ensemble methodology has an intrinsic mechanism for rejection, based on dissent amongst its members. Although this substantially increased the rejection capabilities of the system, especially under the most challenging of real-world interference, false positives still occurred. Therefore, further work on the rejection task to reduce the false positive rate would be extremely beneficial and would increase the overall efficacy of the framework, especially in application to TMEP signals. One potential rejection method, is to include a garbage or world class into the various inter-action classification schemes, making interference rejection not only intrinsic to the heterogeneous ensemble but the base classifier models themselves. Another option is to impose reject bounds on the class-specific posterior probabilities [160]. This implies that a decision is only made at a certain level of confidence. A major issue that would need to be addressed is the trade-off in terms of increasing the specificity at the expense of decreasing the sensitivity. This could be explored using receiver operating characteristic (ROC) surfaces.

To fully validate the useability of the framework for decoding volitional action in the context of HMI, it should be integrated into a real-time assisitve control system. Work towards this end has been initiated, although due to external time constraints and factors beyond the control of this project, it has yet to be fully realised. Specifically, tongue control of a prosthetic hand has been envisaged for restoring upper extremity grasp functionality in below elbow amputees. Current control methods employ antagonistic sEMG signals from the forearm to initiate the opening and closing function of the hand. A further manual (hand-operated) switch is used to toggle between grip patterns [202]. This toggle functionality could easily be controlled using the TMEP or bioacoustic HMI systems, enabling completely 'hands-free' control of the prosthesis. As this only requires one tongue action, the open and close control signals could also be mapped onto additional tongue ac-

tions. Although current EMG control provides proportional signals, it is hypothesised that a similar level of control can be achieved with tongue actions by toggling a discrete signal on and off. Due to the impulsive nature of the actions, fine control resolution would be possible with rapid response in the order of 200ms after initiation of the action. Currently, this TMEP control scheme has been implemented onto a Bebionic prosthetic hand with full evaluation yet to be performed. This evaluation could be done based on the SHAP methodology, for instance [203]. Additionally, a full comparison against EMG-based control and as a complementary control input would also need to be assessed.

In terms of the bioacoustic TMEP concept, the hardware could be improved by making it a completely wireless interface. This would increase the portability of the system and alleviate the effect of wire vibration based interference. By adding accelerometer channels to the sensors, head motion artifacts could also be counteracted. Additionally, these sensors could also be used to capture a proportional control signal based on head rotations (similar to the IMU concepts described in Section 2.4.1). Having said this, it has been highlighted that head movement concepts provides a less accessible route for the physically impaired. Although the four tongue actions defined in this research readily map to the standard control strategy of left, down, right and up, there is the scope for specifying additional actions thus augmenting the instruction set. The main challenge would be to define them so that they are repeatable, have a reasonable SNR and sufficient inter class-separability from the remaining actions. Furthermore, tongue motions which provide proportional control signals, for example tongue sweeps, could also be explored. However, it may prove difficult to obtain repeatable signals based on the current system setup. To fully validate the TMEP concept as a communication tool for the physically challenged, a full study with physically impaired subjects would need to be conducted. Although tongue control is often independent of the impairment, especially in regards to damage of the upper extremities, such a study would fully validate the potential of the concept as means of providing a truly assistive communication link.

The processing framework has been developed as a general method for translating volitional bioacoustic activity into useful machine commands. Although, demonstration of the methodologies on real-time TMEP data does validate its effectiveness, applying it to different bioacoustic signals would further highlight its generality. Different volitional bioacoustic activity could be produced by modifying the initiating action, its physiological location, the type of sensors used and/or the placement of these sensors. For example, MMG or bone conduction microphones could potentially be used to capture tongue movements or tooth clicks by sensor placement over the mandibular condyles. If a reliable signal could be recorded, this would have the additional benefit of not interfering with an individual's hearing. Furthermore, many of the proposed methods have much wider applicability within the biosignal domain in general. Various parts of the framework could

be used for analysing various types of neurophysiological data, including EEG, EMG and LFP signals, for both off-line and online scenarios. The generality of the ensemble framework could be implicated by not only consolidating different base classifier models but also different types of data from heterogeneous sources. For instance, concurrent work has been carried out into the analysis of multichannel LFP activity in conjunction with motion tracking data associated with large-scale movement behaviours in Parkinson's patients (see Appendix C for submitted paper). These gross movements produce neural signals that not only show frequency band synchronisation/desynchronisation but concurrent high amplitude temporal bursting. By using a combination of time and frequency domain processing, in conjunction with a heterogenous classification framework, there is potential for discrimination between the various complex behavioural states. However, one of the biggest challenges when analysing this type of data is the collection of adequate and balanced data sets which would allow various recognition strategies to be evaluated. Once such issues can be surmounted, the proposed decoding methods could be evaluated, with a possible clinical application being a fall prediction and early warning system for advanced Parkinson patients.

7.4 Final comments

This work has investigated using volitional bioacoustic activity as command input for assistive human-machine interfaces. The predominant focus has been developing extraction and decoding techniques for robust translation of user intention expressed through impulsive bioacoustic signals. Using bioacoustic signals enables an imperceptible and completely noninvasive HMI system based on the concept of 'disruption-free' signal capture. This is possible due to the penetrative nature of these mechanical oscillations, allowing the transducers to be positioned away from the source of the volitional activity. Significant contributions have included a robust template creation process enabling stable class-specific signal exemplars to be created and formulation of a heterogeneous ensemble framework. This framework not only increases the effectiveness of the inter-action decoding performance but introduces an additional interference rejection mechanism based on preferential fusion. Demonstration of the processing architecture was performed on bioacoustic TMEP signals and indicated that a real-time tongue-based interface could be developed which required no hardware to be placed within the oral cavity. The proposed architecture has been developed in a generalised manner and so should be applicable for decoding a diverse range of bioacoustic and bio-signal activity.

Appendix A

A note on the wavelet packet transform

This is a brief note regarding the wavelet packet transform (WPT) and its redundant counterpart, the stationary wavelet packet transform (sWPT). It is included for completeness, due to these transforms being utilised in subsequent chapters. Wavelets provide a method for extracting a joint time-frequency representation of a signal using a compactly supported basis [170]. It is because of this compact support that wavelets are considered more efficient at representing non-stationary and transient signals than standard Fourier transforms. This is equivalent to wavelets being, not only localised in frequency, but also in the time domain [204]. This means that wavelet transforms often give better signal representation through multiresolution analysis, providing balanced resolution at any time and frequency. During analysis of non-stationary bioacoustic signals it can provide an efficient method of extracting both time and frequency information simultaneously. Just as the discrete Fourier transform (DFT) provides a realisable and efficient implementation of its continuous counterpart, the discrete wavelet transform (DWT) does this for the CWT. Furthermore, the WPT extends the DWT to provide both high as well as low frequency subband signals. The discrete WPT represents a generalisation of multi-resolution analysis, decomposing a signal into sub-bands and presents both approximation and detail spaces in a binary tree as highlighted by Fig. A.1 [27, 204]. The ideal frequency response associated with each subband output is indicated below the tree and indicates that the transform can be used as a bandpass filter.

The wavelet packet coefficients at one scale are recursively decomposed into the coefficients at the next scale using a low-pass and high-pass analysing filter. To compute the WPT coefficients of scale levels $j = 1, \ldots, J$, it filters the WPT coefficients recursively at the previous stage. Let $W^{\{j,p\}}(k)$ represent the WPT coefficients at level j, then the following two wavelet packet orthogonal base equations are used to compute the wavelet

packet coefficients

$$W^{\{j,2p\}}(k) = \sum_{l=0}^{L-1} h(l) W^{\{j-1,p\}} \left(2k + 1 - l \mod (N_{p-1})\right) \tag{A.1}$$

$$W^{\{j,2p+1\}}(k) = \sum_{l=0}^{L-1} g(l) W^{\{j-1,p\}} \left(2k + 1 - l \mod (N_{p-1})\right) \tag{A.2}$$

where $k = 1, \ldots, N$ and $N_p = N/2p$ while $h(l)$ and $g(l)$ are the impulse responses of the scaling and wavelet filters, which represent low-pass and high-pass filters respectively. Furthermore, they are quadrature mirror filters and have only finite non-zero filter coefficients, which results in an efficient way to compute the WPT coefficients.

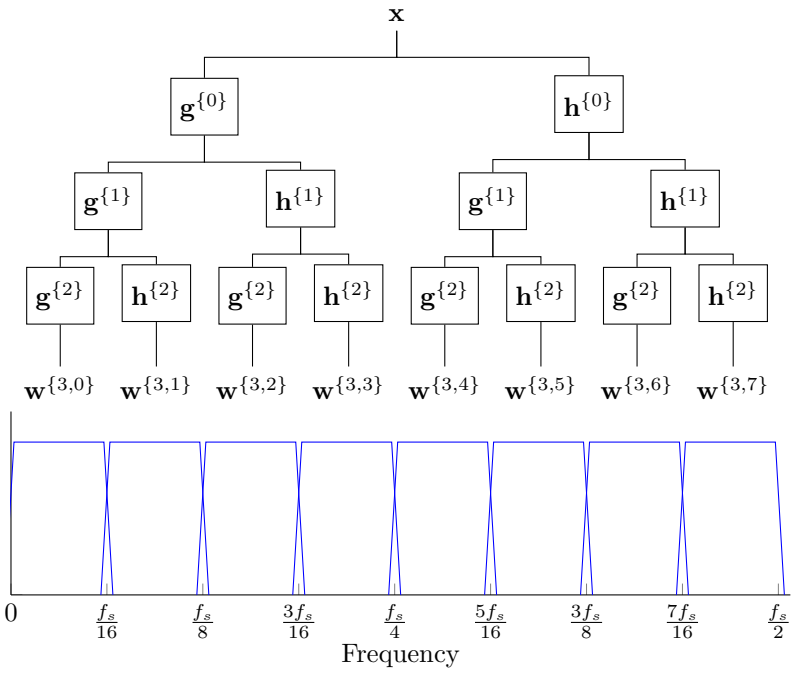


Figure A.1: Generalised filter bank representation of the standard and stationary wavelet packet decomposition for $J = 3$. During implementation of the standard WPT $\mathbf{h} = \mathbf{h}^{\{j\}}, \mathbf{g} = \mathbf{g}^{\{j\}}, \forall j$ and includes a downsampling of the wavelet packet coefficients by a factor of 2 at each level. For implementation of the sWPT no downsampling is performed instead the filters themselves are upsampled by a factor of 2 at each level. The frequency content associated with each packet is indicated by the bottom plot .

Additionally, the sWPT is an adaption of the WPT and provides a redundant method for performing multi-resolution analysis [97, 204]. It differs from the WPT by removing the downsampling between decomposition levels and introducing an upsampling of the filters at each level. This implies that the two new signal components at each level have the same length as the original signal. It has been introduced to overcome the limitations

experienced by the DWT and WPT, by improving the time resolution of these wavelet decompositions. There are many different wavelets that can be selected based on their characteristics and analysing performance, although the 'Sym7' wavelet has been previously shown to be effective in the analysis of bioacoustic signals for HMI applications [27, 97].

Appendix B

Comparison of discrimination strategies

This appendix highlights initial results associated with the four feature extraction methods, namely, AR, DFT, DCT and PCA features. For each feature extraction method, various discrimination strategies are compared based on various combinations of the specific type of feature extraction transform used, the seven feature ranking methods (Section 5.6.2) and the sUGC classifier model. In the case of DCT features, the $rUGC_{mad}$ and $rUGC_{iqr}$ are also tested. These were found to perform less effectively than the sUGC and so were not tested on the remaining feature extraction strategies. Due to the smaller and user specified nature of its feature set, the AR features were also classified using a sMGC. The results are initially compared visually and then statistical tests, in the form of nonparametric Friedman and post-hoc Nemenyi tests are performed. This allowed a specific discrimination strategy associated with each feature extraction method to selected for further comparison in the main body of text (Section 5.10.4).

B.1 AR results

For the discrimination strategies based on AR features, the following feature extraction and classifier combinations were evaluated. These included feature extraction based on both Yule and Burg's method and classification using either the sUG or sMG classifier models. This implies that a total of four discrimination strategies were tested. Figure B.1 shows the mean classification accuracies across the fifteen subjects for the four discrimination strategies.

These results highlight that the sMG is considerable more effective at classifying AR features than the sUG across both the Yule and Burg feature extraction methods. Additionally, Burg's method shows elevated classification accuracy compared to the Yule method

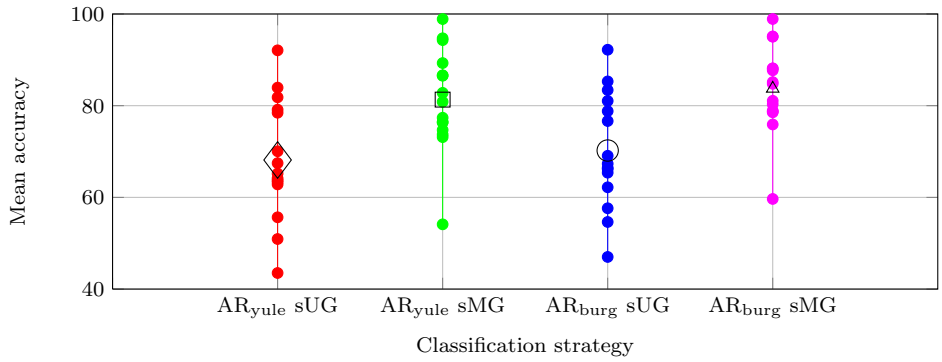


Figure B.1: Mean classification accuracy using the AR transform for feature extraction. The classification accuracy in each fold is taken to be the maximum achieved across the different subsets of features used. Results are presented for each of the four different AR classification strategies as indicated on the x-axis. For each method, the fifteen dots, joined by a vertical line, show the mean accuracy for each individual subject. The black symbol on the same line shows the pooled mean accuracy across all subjects.

when analysing each classifier in turn, although the difference is less pronounced. To highlight these results more rigorously the nonparametric Friedman test was applied to them and indicated that there were significant differences between them at the 5% level. Post-hoc Nemenyi tests were additionally performed with the CD diagrams shown in Fig. B.2. This highlights that the sMG is significantly better than sUG at classifying AR features. Furthermore, Burg's method is close to being significantly better than the corresponding Yule method when used in conjunction with the sMG classifier. The better performance of the multivariate classifier can be attributed to the AR coefficients being based on the autocorrelation function and are therefore jointly distributed. The more effective discrimination associated with Burg's method can be attributed to it being a more accurate estimate of the model coefficients, especially for short signal segments [178]. Based on these initial findings and out of the strategies tested, for extraction of AR features the AR$_{burg}$ sMG is considered the most effective for discrimination of inter-action TMEP signals.

B.2 DFT results

For the discrimination strategies based on DFT features, the following feature extraction and feature ranking combinations were evaluated. These included feature extraction based on the magnitude, phase, real and imaginary of the DFT transform and feature ranking using the seven feature weighting methods described in Section 5.6.2. Classification was based on the sUGC model with a total of 28 discrimination strategies tested. Figure B.5 shows the mean classification accuracies across the fifteen subjects and the 28 discrimination strategies.

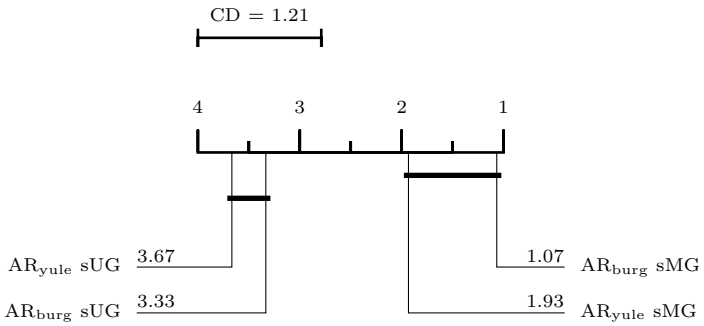


Figure B.2: Critical difference diagrams portraying the post-hoc Nemenyi tests between the four AR classification strategies. The familywise significance level using the Friedman test was p = 1.13e-008.

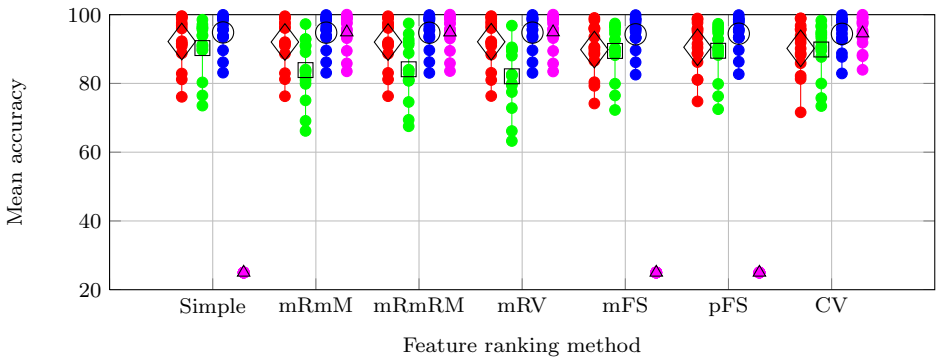


Figure B.3: Mean classification accuracy using the four DFT transform variants (within groupings) and seven feature ranking methods (across groupings). The classification accuracy in each fold is taken to be the maximum achieved across the different subsets of features used. The seven different feature ranking methods are indicated by the x-axis while the four DFT variants are indicated by different colours. For each method, the fifteen dots, joined by a vertical line, show the mean accuracy for each individual subject. The black symbol on the same line shows the pooled mean accuracy across all subjects with DFT$_{mag}$ - diamond (red dots), DFT$_{phase}$ - square (green dots), DFT$_{real}$ - circle (blue dots) and DFT$_{imag}$ - triangle (pink dots).

The results indicate that across all feature ranking method, the phase provided the least discriminative information of the feature extraction strategies. The DFT_{real} and DFT_{imag} had similar performance across most feature ranking methods and generally provided the best discrimination. However, during the simple, mFS and pFS ranking, the features selected from the DFT_{imag} transform continually classified to a single class and as such, their performance was no better than a random guesser. Due to the more stable behaviour of DFT_{real} it has selected for feature extraction of inter-action TMEP signals. The nonparametric Friedman test was applied to the DFT_{real} results and indicated that there were significant differences between them at the 5% level. Post-hoc Nemenyi tests were additionally performed with the CD diagrams shown in Fig. B.4. Although no feature ranking method performed significantly better, the simple filter gave the highest ranking. Based on these initial findings and out of the strategies tested, for extraction of DFT features, the DFT_{real} simple sUG is considered the most effective for discrimination of inter-action TMEP signals.

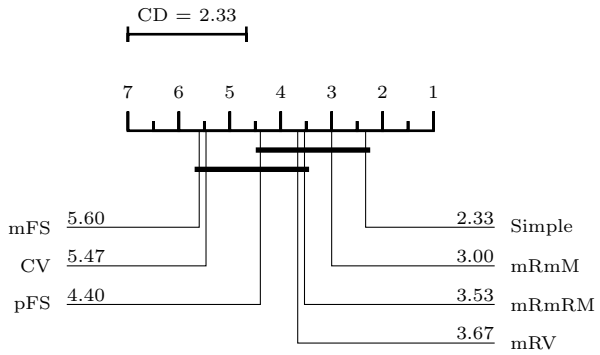


Figure B.4: Critical difference diagrams portraying the post-hoc Nemenyi tests between the seven feature ranking methods. The familywise significance level using the Friedman test was p = 6.47e-5.

B.3 DCT results

For the discrimination strategies based on DCT features, the following feature ranking and classifier combinations were evaluated. These included feature ranking using the seven feature weighting methods described in Section 5.6.2 and classification using the sUGC, $rUGC_{mad}$ or $rUGC_{iqr}$ models. This implies that a total of 21 discrimination strategies were tested. Figure B.5 shows the mean classification accuracies across the fifteen subjects for the 21 discrimination strategies.

The results highlight that across all feature ranking methods, the sUGC classification model provided the most effective discrimination of the TMEP signals across all fifteen

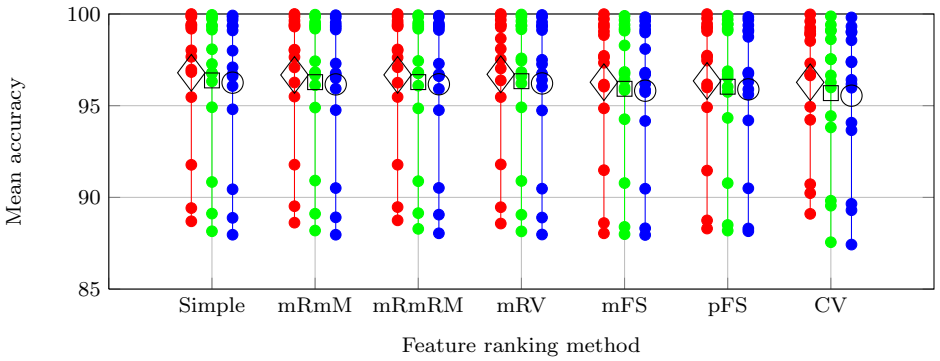


Figure B.5: Mean classification accuracy using the DCT transform with different UG classifiers (within groupings) and seven feature ranking methods (across groupings). The classification accuracy in each fold is taken to be the maximum achieved across the different subsets of features used. The seven different feature ranking methods are indicated by the x-axis while the three classifier variants are indicated by different colours. For each method, the fifteen dots, joined by a vertical line, show the mean accuracy for each individual subject. The black symbol on the same line shows the pooled mean accuracy across all subjects with sUGC - diamond (red dots), rUGC_{iqr} - square (green dots) and rUGC_{mad} - circle (blue dots).

subjects. Both rUGC models performed comparably and worst than the sUGC and is especially noticeable in the case of the CV ranking method. This indicates that out of the feature subsets selected, the training set was well represented with minimal outliers present in the subsets. Further analysis of the sUGC model highlights that the difference across the feature rankings is much less distinct, with the simple, mRmM, mRmRM and mRV all performing similarly. The nonparametric Friedman test was applied to the sUGC results and indicated that there were significant differences between them at the 5% level. Post-hoc Nemenyi tests were additionally performed with the CD diagrams shown in Fig. B.6. Out of the simple, mRmM, mRmRM and mRV feature ranking methods none performed significantly better than the others, although the mRmRM ranking had the highest ranking fractionally. Based on these initial findings and out of the strategies tested, for extraction of DCT features, the DCT mRmRM sUG is considered the most effective for discrimination of inter-action TMEP signals.

B.4 PCA results

For the discrimination strategies based on PCA features, the following feature extraction and feature ranking combinations were evaluated. These included feature extraction based on the PCA_{ucs}, PCA_{pcs} and PCA_{pci}, and feature ranking using the seven feature weighting methods described in Section 5.6.2. Classification was based on the sUGC model with a total of 21 discrimination strategies tested. Figure B.7 shows the mean

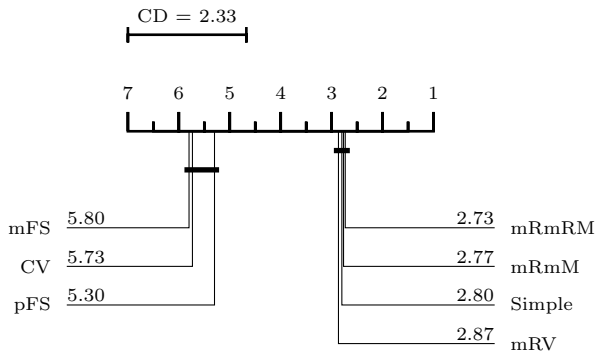


Figure B.6: Critical difference diagrams portraying the post-hoc Nemenyi tests between the seven feature ranking methods. The familywise significance level using the Friedman test was p = 5.69e-8.

classification accuracies across the fifteen subjects for the 21 discrimination strategies.

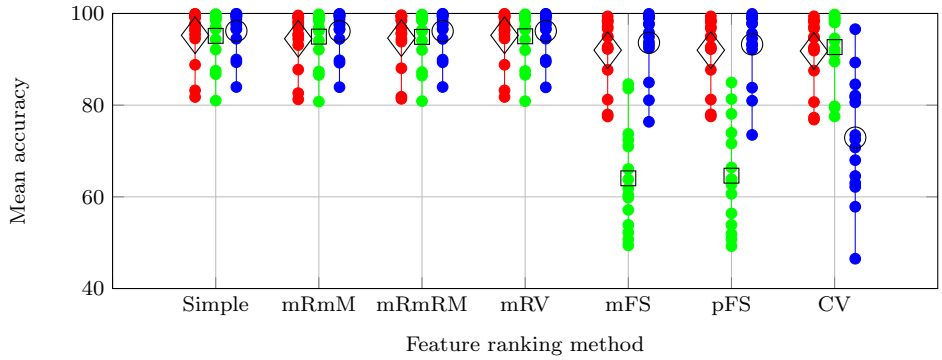


Figure B.7: Mean classification accuracy using the three PCA transform variants (within groupings) and seven feature ranking methods (across groupings). The classification accuracy in each fold is taken to be the maximum achieved across the different subsets of features used. The seven different feature ranking methods are indicated by the x-axis while the three PCA variants are indicated by different colours. For each method, the fifteen dots, joined by a vertical line, show the mean accuracy for each individual subject. The black symbol on the same line shows the pooled mean accuracy across all subjects with PCA_{ucs} - diamond (red dots), PCA_{pcs} - square (green dots) and PCA_{pci} - circle (blue dots).

The results indicate that across the majority of the feature ranking methods, the PCA_{pcs} generally had the lowest classification accuracy of the feature extraction strategies. For the simple, mRmM, mRmRM and mRV, the PCA_{pci} had the best performance and is therefore considered the most effective feature extraction method based on PCA features. The non-parametric Friedman test was applied to the PCA_{pci} results and indicated that there was significant differences between them at the 5% level. Post-hoc Nemenyi tests were additionally performed, with the CD diagrams shown in Fig. B.8. Although no feature ranking

method performed significantly better, the simple method was ranked highest. Based on these initial findings and out of the strategies tested, for extraction of PCA features, the PCA_{pci} simple sUG is considered the most effective for discrimination of inter-action TMEP signals.

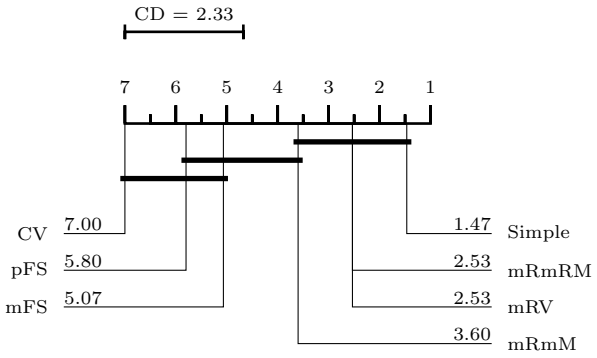


Figure B.8: Critical difference diagrams portraying the post-hoc Nemenyi tests between the seven feature ranking methods. The familywise significance level using the Friedman test was p = 9.37e-15.

Appendix C

Paper submitted to journal of neuroscience methods

This is a paper based on work carried out by the author, during a six month period, performed concurrently to the main research discussed in this thesis. The paper has been submitted to the journal of neuroscience methods and is currently under review. It introduces a method for detecting events associated with complex or large-scale behaviours from neurological activity. Local field potentials were recorded, following surgical implantation, from three patients with Parkinson's disease as they participated in a large movement-based paradigm. The proposed methodology provides a basis for extracting event-related potentials from these signals based on complex motions in unconstrained experiments. For full details of this research, the reader is directed to the adjoining paper. Although not directly related to bioacoustic signal processing, it does share some commonalities with the main research focus of this thesis. The notion of extracting complex behaviours from neurological signals based on temporal features bears similarities with the detection of volitional activity in bioacoustic signals. In fact, both the underlying detection strategies are based on estimation of the short-term energy and thresholds. Although in the research described in this Appendix, the threshold was adaptive and parameters associated with the detection model were optimised using a multi-objective genetic algorithm. Many of the bioacoustic decoding strategies, especially the heterogeneous ensemble framework, are readily applicable to the analysis of multichannel neural signals. However, a major limitation with invasive neural recording is the amount of data that can be collected from a patient and often implies that there is an insufficient amount for both training and testing a classification strategy.

An automated approach towards detecting complex behaviours from deep brain structures

Michael Mace, Nada Yousif, Mohammad Naushahi, Khondaker Abdullah-Al-Mamun, Shouyan Wang, Dipankar Nandi, Ravi Vaidyanathan

Abstract

Extracting event-related potentials (ERP) from neurological rhythms is of fundamental importance in neuroscience research and future visions of brain machine interfaces. Standard ERP techniques, however, require associated ERP waveforms to have low variance, be shape and latency invariance, and rely heavily on many repeated trials. Utility is therefore limited to the analysis of simple behaviours in tightly constrained experiments. In this paper, we introduce a method for automatically detecting events associated with complex or large-scale behaviours (e.g. gross movements) from neurological activity, where the ERP is not constrained by the aforementioned requirements. The algorithm is based on the calculation of a detection contour and adaptive threshold. These are combined using logical operations to produce a binary signal indicating the presence (or absence) of an event. The associated detection parameters are automatically optimized using a multi-objective genetic algorithm. To validate the proposed methodology, deep brain signals were recorded following surgical electrode implantation from three patients with Parkinson's disease, as they participated in a large movement-based behavioural paradigm. The experiment involved bilateral recordings of local field potentials from the sub-thalamic nucleus and pedunculopontine nucleus. The subjects were asked to search and orientate towards one of eight random light sources positioned uniformly around them. Due to the complex motor control loops and additional cognitive processes utilised during such a task, the neural signals contain a vast amount of integrated information. Even with this large amount of intrinsic (and extrinsic) noise, the algorithm was able to correctly detect events with [87.5±6.5, 76.7±12.8, 90.0±4.1] and [92.6±6.3, 86.0±9.0, 29.8±12.3] average sensitivity and specificity (±1sd) across the four brain sites. Thus, despite these challenging intrinsic conditions, strong correlations between the onset of the large-scale movement behaviours and the basal ganglia-brainstem local field potential activity were found. Furthermore, the methodology can be utilised in real-time as only a single trial ERP is required, with the detection algorithm calculated causally. We believe these results provide a basis for extracting ERPs from complex motions in unconstrained experiments, and show future promise for ERP recording in non-clinical environs.

1. Introduction

Detection of neuronal events relating to modulations in brain states is of fundamental importance in the field of neuroscience. These generally follow from a change in environmental stimuli, which can be of varying complexity. Such events can be detected within a neural signal or rhythm, highlighting stereotypical responses and correlations between the neurological phenomenon, physical or pathological response and spatiotemporal neuroanatomical relationships. This modulation in the brain's state generally manifests as an event-related potential (ERP) within a specific neurological rhythm [1, 2]. By examining these evoked potentials in terms of event-related synchronisation (ERS) and/or de-synchronisation (ERD), useful insights about the nature and timing of these events and how they sub-serve sensorimotor, perceptual or pathological processes can be elucidated

[3][4][5][6]. The majority of research regarding ERP's relates to movement or visual responses triggered using cued, self-paced or imagined stimuli [6][7][9]. As such, their associated potentials can be sub-categorised as movement-related potentials (MRP) or visually-evoked potentials (VEP) and occur within specific cortical-subcortical neural circuits. For instance the Bereitschaftspotential (BP) also known as the pre-motor potential is a measure of the neural activity exhibited in the motor cortex preceding a volitional muscle contraction. An example of a pathological event related potential is the change in epileptiform activity due to the clinical onset of seizures [10][11].

The majority of ERP research is performed using electroencephalographic (EEG) scalp electrodes on healthy individuals [2][4][12]. As such, the ERP is generally deeply buried in noise, with a very low signal-to-noise ratio (SNR). This is mainly attributed to the fact that the EEG signal represents an amalgamation of far-field potentials which are heavily attenuated due to transmission through the cranium and scalp. To therefore extract the ERP, many trials need to be conducted so that the resulting waveforms can be averaged together allowing the random brain activity to be averaged out and the signal of interest exposed. This is under the assumption that the signal of interest is event-locked with invariant latency and shape, and that the remainder of the signal constitutes noise which can be approximated by a zero-mean independent Gaussian process.

More recently, MRPs captured within deep brain structures have been explored [7][8][13] [14]. This is for the most part, due to recent developments in functional neurosurgery, specifically, clinical implantation of deep brain stimulation (DBS) electrodes, for example, within the Basal Ganglia structure for therapeutic relief of Parkinsonian symptoms [15][16][17][18][19]. The same DBS electrodes can also be used to record local field potentials (LFP) alongside action potentials thus bringing much higher spatial resolution than corresponding EEG whilst researchers explore various motor and behavioural paradigms. Such studies have included defined finger, wrist, foot and pointing movements, to name a few and have helped to highlight the role of these sub-cortical structures within a variety of sensorimotor control loops [20][21].

The majority of motor paradigms tested have involved small and discrete bodily movements, corresponding to fine motor skills which allow for high subject and thus inter-trial repeatability. This increases the chance of the associated waveform being event-locked to the stimulus and physical response, with invariable latency and shape. However the ERP will generally have a much smaller SNR and thus inter-trial averaging becomes essential. While the importance of untethered and more open experiments has been recognized, most analysis techniques are designed around more constrained and controlled experimental protocols [23]. In recent years the exploration of more complex movement and behavioural paradigms have become of interest [24][25][26]. These include the study of gross movement behaviours such as gait, locomotion and postural control, where multiple muscle synergies and neural control loops are engaged. Examples of such movements include walking, turning and balancing (involving vestibular functions) and standing up / sitting down. This leads to a much higher SNR of the evoked response due to an increased rate in neural firing within the specific cortical-subcortical network. Further to this, if enough excitatory neural pathways are activated then the ERP signal will no longer be buried in noise. However, the shape and latency invariance of the ERP is generally compromised due to both intrinsic and extrinsic neural variability. While it is still of great interest to extract these gross ERPs, a different approach is required.

In this paper we outline such an approach whereby events-relating to gross movements and behaviours are able to be extracted from neural signals in an automated and systematic fashion. Here we define a gross neural response as one which produces an ERP not buried in intrinsic or extrinsic noise. This will be at the cost of low inter-trial variability thus preventing standard ERP methods to be employed. In the next section, the automated detection methodology is introduced which incorporates a decision envelope and adaptive threshold. The detection algorithm is validated on a data-set collected from patients with Parkinson's disease (PD) who have undergone bilateral DBS electrode placement within both the sub-thalamic nucleus (STN) and pedunculopontine nucleus (PPN) substructures. This allows for LFP signals to be recorded while they took part in a post-operative turning and balance experiment. The parameters of the detection algorithm are optimised using a multi-objective genetic algorithm in response to motion data collected from the head and trunk of the subject. Finally, experimental results are presented for three such subjects and the potentiality and usability of the methodology discussed.

2. Material and methods

a. Large-scale behaviour signal model

For clarity, the following simplified signal model is proposed to represent large-scale neuronal behaviours

$$s(t) = f(t) + I(t) + E(t)$$

Where $s(t)$ is the measured neural signal, $f(t)$ is an ERP or signal of interest, while $I(t)$ and $E(t)$ are intrinsic and extrinsic noise signals, respectively. It is assumed that $f(t)$, $I(t)$ are non-stationary and therefore $s(t)$ is also a non-stationary signal from an unknown distribution, while $E(t)$ is assumed to be sampled from a zero-mean Gaussian process. The intrinsic noise component constitutes internal near-field volumetric activity which is generated from simultaneously occurring brain responses and to some extent may be correlated to the ERP. The extrinsic noise component comprises measurement errors and far-field volumetric effects (such as those experienced in EEG signals) and therefore can be assumed Gaussian. When the inter-trial variability is high, the ability to detect an ERP has to be on an individual basis and is only possible if the signal energy associated with an ERP is significantly higher than the noise i.e.

$$\sum_t |f(t)|^2 \gg \sum_t |I(t) + E(t)|^2$$

There are still many challenges associated with the detection of such evoked neuronal events. These include but are not limited to temporally changing signal energies and stationary conditions, limited data sets to train and optimise the detection algorithm and also disparate neuroanatomical locations of ERP occurrence.

b. Detection algorithm

Well established techniques used for the prediction of epileptic episodes have been adapted and generalised for the purpose of this paper [11][27][10]. Epileptiform activity subscribes well to the proposed signal model outlined previously. This is due to a relatively quick transition from a quasi-stationary pre-ictal baseline to a highly non-stationary and energetic ictal state whilst continually being interspersed with activity associated with the individual's general cognitive processes. The

proposed detection algorithm can be broken down into two main parts, namely, computation of a detection contour and the implementation of an adaptive threshold. Combining these two constituents through logical operations produces a binary detection output signal.

Detection contour: The detection contour is based on the short-term energy (STE) of the measured neural signal which provides an efficient and reliable contour for detection. The standard way to calculate the STE is equivalent to convoluting the square of the signal with a moving average filter (MAF). However, the arithmetic mean is known not to be a robust measure of the central tendency of a signal. This can be especially problematic when analysing neural signals as the potential for outliers is high, due to action potentials from individual neurons being superimposed onto the neural activity. This is further compounded by the squaring of the sample value. Instead, a more statistically robust and general filter is based on order statistics (OS) and is derived as follows.

Given a set of observations, $S = [s(1), s(2), \dots, s(n)]^T$ with $s(n)$ being the most recent signal sample, the order statistics in the interval $[k, k+n]$ are obtained by sorting the individual observations in ascending order. This produces $\hat{S}$ satisfying

$$\hat{s}(k) < \hat{s}(k+1) < \cdots < \hat{s}(k+n)$$

If $\alpha_i, i = 1,2, \dots, N$ are the filter coefficients, the order statistics filter (OSF) at $n = k + N$ is

$$\hat{y}(n) = \alpha_1 \hat{s}(k) + \alpha_2 \hat{s}(k+1) + \cdots + \alpha_N \hat{s}(n)$$

This is a generalised filter with the moving average, trimmed mean and median filters specific instances of the OS filter. The robustness of the filter is exercised when the tails of the filter coefficients are set to zero, as is the case with the trimmed mean and median filters. The median filter provides a 50% statistical breakdown point.

A robust estimate of the STE is therefore given by the following

$$E(n) = \alpha_1 \hat{s}(k)^2 + \alpha_2 \hat{s}(k+1)^2 + \cdots + \alpha_N \hat{s}(n)^2$$

With the final detection contour based on the instantaneous amplitude of the STE signal. This is found using the Hilbert transform $H\{\}$ and provides a smoother envelope for detection

$$\tilde{E}(n) = |E(n) + iH\{E(n)\}|$$

Where $|\dots|$ is the magnitude and $i = \sqrt{-1}$.

Adaptive threshold: An adaptive threshold is required to minimise the false detection of events due to the intrinsic noise within the measured signal. By allowing the threshold to follow the detection contour, only when relatively large changes in the STE value occur will a true detection be flagged. The basic idea of the threshold being, when it is exceeded and then receded within certain time constraints a detection will be made. The adaption process involves downsampling of the detection

contour, OS filtering and an exponential forgetting factor with the threshold remaining at that value till the next update. The exponential forgetting is calculated using the following equation

$$y(n+1) = (1-\lambda)x(n) + \lambda y(n)$$

Where $\lambda \in [0,1]$ is the forgetting factor and dictates how closely $y(n)$ follows $x(n)$.

The adaption process alongside the overall detection algorithm is highlighted in . To prevent spurious activity from being detected, $\tilde{E}(n)$ also needs to exceed the threshold for a minimum time period t_m. The final binary detection output is therefore given by the following

$$D(n) = \begin{cases} 1 & if\ \tilde{E}(n) > \beta T(n)\ and\ (b-a) > t_m \\ 0 & otherwise \end{cases}$$

Where $T(n)$ is the adaptive threshold, β is a scaling parameter and $[\tilde{E}(a), \ldots \tilde{E}(b)]$ is the set of consecutive samples which satisfy the first condition.

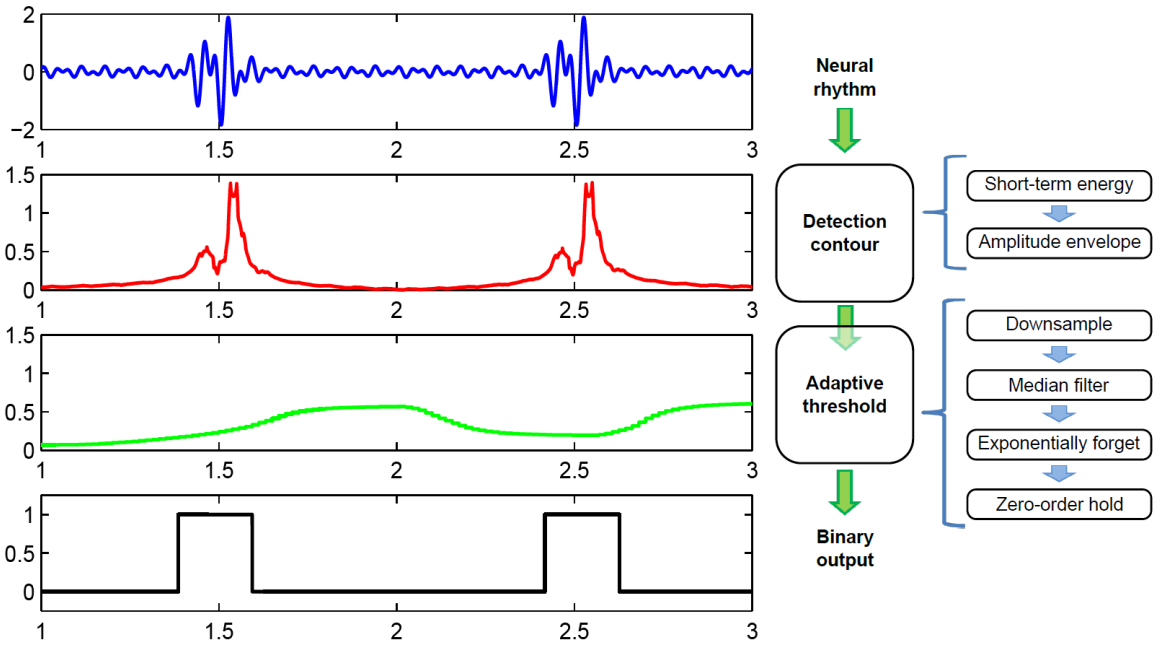


Fig. 1. Overview of the detection algorithm: The left shows artificial plots highlighting the output at each stage of the methodology including the neural rhythm, detection contour, adaptive threshold and binary output, respectively. The right shows the steps required to compute each of these outputs

c. Movement and recording paradigm

The basal ganglia-brainstem system (BG-BS) plays an important role in the control of various types of behavioural expression [22][21]. It plays an important role in both automatic and voluntary control processes including rhythmic limb movements and adjustment of postural tone during locomotion. Central to this neural architecture is the PPN which provides dense efferent and afferent connections between the two subsystems [28][29]. The PPN modulates its activity in response to locomotion and voluntary arm motions [13]. Dysfunction of the PPN structure has been highlighted as a possible cause of a variety of clinical PD symptoms, including gait freezing and a decrease in

vestibular control [26][24][30]. Having said this, the PPN has only recently been explored with regards to PD, DBS and LFP signal analysis [6], [13], [24]. Furthermore, research pertaining to both the basal ganglia and PPN structures and linking this to movement, balance and gait function has so far been minimal. The majority of research regarding LFPs and deep brain recordings has generally been on individual brain structures usually from electrodes placed in the STN or internal segment of GP as these are established targets for therapeutic application of functional DBS for Parkinson or dystonia [17][31][32][33].

Three patients with Parkinson's disease who were selected for bilateral implantation of DBS electrodes in both the STN and PPN participated in the experiment. They were part of a Medical Research Council funded peer-reviewed clinical trial, exploring the relative benefits of STN and PPN DBS in PD. All participants took part with informed consent and ethical approval was obtained from the Local Research Ethics Committee. The DBS electrodes (models 3387 (PPN) and 3389 (STN), Medtronic Neurological Division, Minnesota, USA) were bilaterally implanted in the STN and PPN. The electrodes have four platinum–iridium cylindrical contacts (1.27 mm diameter and 1.5 mm length) and a contact-to-contact separation of 0.5mm (model 3389) or 1.5mm (model 3387). LFPs were recorded via the externalised electrode leads from the STN and PPN during the week immediately post-implantation, before the pulse generator was implanted and the DBS system internalised. Final electrode contact positions were being selected where optimal decrease in Parkinsonian symptoms occurred during intra-operative electrical stimulation. The implantation site was further confirmed by examining the fused images of pre-implantation MRI with post-implantation CT. All patients also had a delayed post-implantation MRI brain scan to establish electrode position. The electrodes are utilised in a bipolar recording configuration with the contact pairs chosen for analysis, involving contacts 1-2 for the STN and contacts 0-3 within the PPN. The four electrode sites are therefore labelled STN L 12, STN R 12, PPN L 03 and PPN R 03, respectively. These pairs were chosen based on anatomical considerations and where they produced better therapeutic effects.

The experimental paradigm was devised so as to instigate large scale behavioural responses in each of the subjects. These included gross motor function, vestibular control, regulation of postural muscle tone and complex co-ordinations between distinct efferent pathways. The experimental procedure was as follows, a subject stood surrounded by a circle of 8 light sources. The light sources were distributed uniformly around the individual at a distance of 0.6m. shows an overview of the experimental paradigm including electrode placement alongside a schematic and photo of the experimental setup. Initially the reference light was on, with the subject instructed that when this light turns off another random light source will turn on. Their task was to locate this light and turn to face it so that their body was positioned normally to it. The light remained on for 15 seconds and then switched back to the reference point, this time for a shorter duration of 10 seconds. There was only ever one light source on, with the process repeating (approximately ten times) until the experiment terminated.

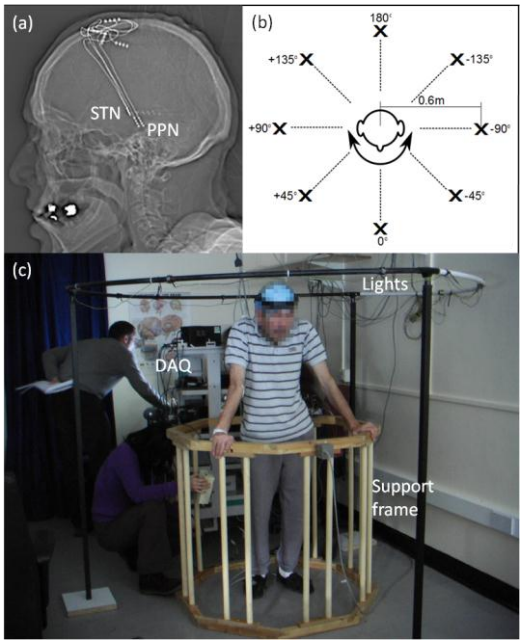


Fig. 2. Overview of the experimental paradigm and setup: (a) X-ray showing electrodes placement within the STN and PPN. (b) Schematic showing bird's eye view of the experimental setup for the proposed turning and large-scale movement paradigm. The crosses indicate light sources which are positioned uniformly radially around the subject at a distance of 0.6m. (c) Photo of the experimental procedure prior to recording.

During the experiment, LFP signals from the aforementioned sites were recorded. At the same time motion tracking, electromyography (EMG) and DC electrooculography (EoG) signals were also recorded. The motion data was measured with a Polhemus Fastrak system and tracked the angular displacement of the anteroposterior axis about the yaw plane of the head, trunk, left and right feet within the global frame. Due to the experimental setup, tracking the anteroposterior angular displacement within the global frame was found to adequately map the overall motion of the various body parts. This was due to the individual generally orientating to face the light source. Additional sensors included an EMG electrode placed on both the left and right tibialis anterior muscles while the EoG recorded the electrical activity across both eyes. The electrode-based signals were amplified using CED 1902 amplifiers and digitised using CED Micro 1401 at a rate of 2000 Hz for both the LFP and EMG channels and 250 Hz for the EoG channel. The motion tracking data was sampled at a rate of 100 Hz. The signals were displayed online and saved onto a hard disk using a custom written program in Spike 2 (Cambridge Electronic Design, CED, Cambridge, UK). To minimise the stress an experiment of this nature imposes on the PD subjects, the number of trials had to be kept to a minimum. To warrant DBS surgery with electrode placement in both the STN and PPN, the subject needed to exhibit significant dysfunction in postural control, gait and locomotion in addition to the usual motor symptoms of tremor, rigidity and bradykinesia. There was also a limited amount of time between the operation and experiment (1-2 days), with these factors contributing to the length of time a subject could be expected to perform. The time of each trial was relatively long (5-6 minutes), which was necessary to allow the subjects to exhibit these large-scale neuronal responses. Under such circumstances there will always be challenges with the amount and consistency of the data that can be collected and therefore designing an analysis method that doesn't rely on a large number of repeated trials is of paramount importance.

d. Detecting the neural response onset

Ideally a detection process would like to minimise both the number of false negatives (N_{FN}) and number of false positives (N_{FP}). A false negative is a missed event given that one occurred while conversely, a true positive is the correct identification of an event. A false positive is the incorrect flagging of an event given that no event occurred while a true negative is correctly not flagging an event when no event occurred. A neural response onset occurs when the detection contour changes from low to high indicating a change from inactivity to a period of activity. This neural onset is indicative of substantial neural firing and can be linked to the physical response onset. The physical response can be a stimulus, behavioural, physical movement or pathological response for example and is defined as the response to which the neural activity is being correlated with. As these onsets represent single points in time, with the neural response either leading or lagging the physical response, a time window surrounding the physical onset has to be defined to allow a neural onset occurring within this period to be flagged as a true positive. If no neural onset is detected within this window then a false negative is recorded. False positives occur in periods when no physical response is indicated but a neural onset occurs, while a true negative occurs when a period of no physical response contains no neural onsets within it.

e. Optimisation of the detection parameters

The proposed detection algorithm (as described in section 2b.), requires a certain number of parameters to be selected with these outlined and described in below. Additionally, the actual OS filter coefficients (and not just the number of them) can also be adapted so as to maximise the efficacy of the detection process. However, this significantly increases the search space and therefore its complexity, as each coefficient would individually need to be optimised. For this reason a median filter is selected with only the length of the filter requiring to be tuned. The median is chosen as this has the largest statistical breakdown point and therefore is the most robust to outliers. To further reduce the search space, L_t has been made dependent on DF (by a factor of 0.25) as they both ultimately affect the update rate associated with the adaptive threshold.

Table 1: The names and description of the tuneable parameters associated with the detection algorithm. Also given are the pre-defined grid values used during the gridded search optimisation approach and the lower and upper bounds defined during the multi-objective GA approach.

Design parameter	Description	Grid search space	GA lower bound	GA upper bound
$\boldsymbol{L}$	Number of OS filter coefficients during STE contour calculation	[32,64,128,256]	16	512
$\boldsymbol{DF}$	Downsample factor during threshold adaptation	[128,256,512,1024]	64	2048
$\boldsymbol{L_t}$	Number of OS filter coefficients during threshold adaptation	$DF/4$	16	512
$\boldsymbol{\lambda}$	Exponential forgetting factor during threshold adaptation	[0.1,0.2,…,0.9]	0.01	0.99
$\boldsymbol{t_m}$	Minimum time frame to prevent spurious activity from being detected	[32,64,128,256]	16	512
$\boldsymbol{\beta}$	Threshold scaling factor	[0.5,1,1.5,…,40]	0.1	100

The selection of the detection parameters can be cast into a multi-objective optimisation problem. This design set can be tuned based on the minimisation of certain objectives, which in the case of a

detection problem are N_{FN} and N_{FP} respectively. As with most multi-objective optimisation problems, there is a trade-off between minimisation of the various objective criteria. When the detection parameters are relaxed there is a reduction in the number of false negatives while the number of false positives will increase; the converse being true when the detection parameters are tightened. This means that it is highly unlikely that there will be a unique solution to this type of problem. Instead, the available solution space can be represented using a Pareto optimal front whereby the set of non-dominated solutions is retained, while the dominated solutions are discarded. The corresponding parameter set for this non-dominated front is called the Pareto optimal set. If $\boldsymbol{\xi}^i = [\xi_1^i, \xi_2^i, \dots, \xi_p^i]$ is a specific instance of the parameter set and $\boldsymbol{J}(\xi^i) = [J_1(\boldsymbol{\xi}^i), J_2(\boldsymbol{\xi}^i), \dots, J_q(\boldsymbol{\xi}^i)]$ is the set of all objective functions evaluated on this parameter set instance, then a solution $\boldsymbol{\xi}^a$ is said to dominate $\boldsymbol{\xi}^b$ if the following holds

$$\{J_j(\boldsymbol{\xi}^a) \leq J_j(\boldsymbol{\xi}^b), \forall j \in [1,2,\dots,q]\} \wedge \exists \{J_k(\boldsymbol{\xi}^a) \leq J_k(\boldsymbol{\xi}^b), \forall k \in [1,2,\dots,q]\}$$

This can be summarised via the following, for any two instances, the dominant solution performs better in at least one objective and worse in none. This holds true between all pairwise combinations of the Pareto solutions and the discarded solutions. Further to this, the Pareto optimal front and set can be represented graphically using a method based on level diagrams [34]. The level diagrams tool is based on the classification of the Pareto front approximation according to the proximity to a hypothetical ideal point, with this point characterised by having a minima in all the objectives. By calculating vector norms on $\boldsymbol{J}(\xi^i)$, the proximity of each solution to the hypothetical minimum can be found and each point in the Pareto front and set can be plotted against this value, thus giving an indication of its relative performance. A commonly used norm is the Euclidean norm which gives a measure of distance of a particular solution to the ideal point. Level diagrams are a useful tool for a designer to explore the parameter space comprising potential solutions without having to make a priori decisions about the relative importance of the individual objective functions.

The simplest way to tune the detection parameters is to define a multi-dimensional grid of predetermined design values and evaluate each solution in turn. Those solutions which dominate can be retained while the rest are discarded. However, if the solution space is highly discontinuous then a small change in the design parameters can lead to a relatively large change in the objective functions. Specifically, the perturbation of one detection parameter by a small amount can lead to a large change in the number of false positives and/or negatives being recorded. Unless the grid is defined with a high enough resolution then dominant solutions can be easily missed. Besides having to determine the grid a priori there may also be certain regions of the search space which yield no dominant solutions and therefore it is a waste of resources to search within such areas. The advantage of using a grid search is it's easily implemented without the requirement of any additional algorithmic parameters to be set.

To alleviate the issues of setting up an a priori grid with redundant search regions, a directed search methodology can be employed. Due to the highly discontinuous nature of the design space, a gradient search cannot be implemented and therefore a stochastic algorithm needs to be used. In this work, a multi-objective genetic algorithm (GA) has been chosen due to its ability to search the

space in an efficient and directed manner [35][36]. GA's are based on evolutionary theory with the concept of natural selection used to guide the process. Through reproduction and the mixing of chromosomes, a population of chromosomes is evolved through generations, with stronger chromosomes more likely to pass on their genes than weaker ones. Chromosomes are made up of discrete units called genes, with each gene encoding a specific instance of the design set. By effective use of the mutation and crossover operators an efficient exploration-exploitation trade-off can be achieved. Exploration of the search space is achieved through random perturbation of the genes (mutation), while exploitation utilises the passing on of strong genes to the next generation (crossover). The fitness of a gene is evaluated on an objective function or in the case of multi-objective optimisation a set of objectives. For a more in-depth overview of GA's and multi-objective GA's in particular, the reader is directed towards [35].

In this work, the particular multi-objective GA employed is a variant of the fast non-dominated sorting genetic algorithm (NSGA-II) [37]. More specifically this is a controlled elitist GA and was chosen due to its speed and convergence properties. The encoding of the genes is usually as if not more important, as the stochastic search algorithm itself. Due to the nature of the parameter set it has been encoded as a mixed-integer problem, with $\{L, DF, t_m\} \in \mathbb{Z}$ and $\{\boldsymbol{\lambda}, \boldsymbol{\beta}\} \in \mathbb{R}$. To prevent unconstrained growth of the parameter set, bounds need to be placed on them. To ensure the solutions are not biased towards a single objective, additional operators such as the diversity operator are utilised. The advantage of using such an evolutionary algorithm is that any viable parameter set can potentially be reached and evaluated, thus aiding in the convergence of the algorithm to the global optimum. An additional benefit is that as new generations are evaluated, the Pareto front can be calculated directly, therefore lowering the number of redundant calculations being performed, in areas of the design space which are considered suboptimal or dominated. The drawback being that there are now additional GA parameters which need to be chosen. These algorithmic parameters are described in alongside their chosen values.

Table 2: The key parameters and functions used by the elitist NSGA-II algorithm. The parameter name and short description are given alongside the chosen implementation value used in this study.

Multi-objective GA parameter	Description	Chosen value
Population size	Number of chromosomes required in each generation. The more individuals used the more thorough the search will be at the expense of computational resources.	300
Mutation function	Specifies how the GA makes random changes to the chromosomes between generations. A uniform function selects a fraction of the genes based on the mutation rate and mutates these genes randomly within the specified range.	Uniform
Mutation rate	This is the probability rate associated with a gene being mutated. Due to the discontinuous nature of the search landscape, this has been set relatively high.	0.2
Crossover function	Specifies how the GA combines two individuals between generations. The scattered function creates a vector of random binary values and selects which parent's gene to inherit based on whether the corresponding value is high or low.	Scattered

Crossover fraction	The fraction of the next generation that are produced using the crossover operator.	0.6
Selection function	Specifies how the GA selects parents for the next generation. Stochastic universal sampling uses a single random variable to sample all of the solutions by selecting them at uniformly spaced intervals. The samples themselves are spaced proportionately to their fitness value.	Stochastic universal sampling
Distance function	Specifies how the GA maintains diversity by favouring individuals that are relatively far away from the Pareto front. The crowding distance associated with an individual provides an estimate of the density of solutions surrounding it. This can be done in the design space (genotype) or objective space (phenotype)	Genotype crowding distance
Elite count	The number of the elite individuals that are guaranteed to survive to the next generation. This helps the algorithm to eventually converge to a optimum.	2
Pareto fraction	The maximum fraction of individuals that can be on the Pareto front. Acts in collaboration with the mutation rate to maintain and control diversity.	0.5

3. Results and discussion

The experimental paradigm was devised to instigate large scale movement behaviours from the subjects. Over the course of a trial, two discrete phases were generated, namely a movement and rest phase. To highlight the ability of the detection algorithm to extract large-scale behaviours from LFP signals, it is used to extract these gross movement onsets. A movement phase is defined as a period where the subject is searching and locating a light source followed by orientation of their body towards it. A rest phase consisted of the subject waiting for a different light source to be illuminated, with the two phases alternating over the course of the experimental trial. Sometimes the subject would not orientate towards a new light location within the given 15 seconds. On these rare occasions, the subject was instructed to continue searching and would now find the reference light on thus orientating to this instead. In such cases, the subject would exhibit movement over this entire period and therefore it could be considered a single movement phase.

It was found that the EoG signal was erratic, saccadic and uncorrelated with the LFP signals. Thus the EoG signal was ignored in any analysis, as its associated LFP information within the BG-BS substructure was assumed to be overshadowed by the much larger movements of the body. Additionally it was found that during a movement phase, the head and trunk signals were generally the first and last to show motion and thus the feet and EMG channels could also be ignored. The head and trunk signals were segmented manually to ensure that movement and rest partitions were captured accurately. The first 100 seconds of the movement and rest phases, movement onsets and normalised stimuli, head and trunk signals for the three subjects are shown in .

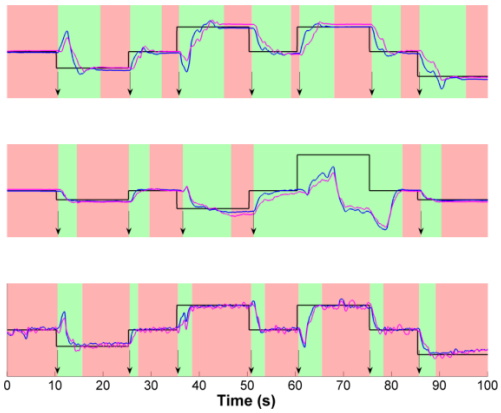


Fig. 3. Plots showing the manually partitioned movement and rest phases for the 3 subjects (Top – subject 1, middle – subject 2 and bottom – subject 3). Movement phases are shown with green patches while rest phases are red. The stimulus signal (black line) and anteroposterior rotation for the head (blue line) and trunk (purple line) signals are shown on top of the patches.

A movement phase includes periods where the head, trunk or both parts of the subject were shown to be in motion. A rest phase was recorded if the anteroposterior angle was unchanging and therefore the signal exhibited a flat region. These flat regions were required to be at least 2 seconds long, as previous studies have shown that the early MRP component within LFP signals, recorded in the basal ganglia generally occur up to two seconds preceding a movement [8][13]. Also for this reason, the time window for neural detection of a true positive has been set to ±2 seconds either side of the movement onset. If for whatever reason a shorter rest phase did occur, for example when the subject changed direction, it would be absorbed into the movement phase. This wouldn't affect the analysis, as only the movement onsets are of interest. Another reason for partitioning the phases manually is due to one subject exhibiting a low frequency postural tremor. As this is unintentional (relatively small-scale) motion, it is not considered part of the large-scale volitional movement behaviours of the individual and is generally visible only during the rest phase. Any high-pass filtering would cause ringing due to the sharp transitions within the signals and thus was avoided. By partitioning manually, the volitional movement and tremor phases could be effectively separated.

The LFP signals were initially band-limited to 128 Hz using a low-pass FIR filter and then re-sampled at 256 Hz. A 10^{th} order Butterworth notch filter with cut-off at 50±0.5 Hz was used to remove the mains noise. By using wavelet packet decomposition (WPD) at scale 7 alongside a discrete Meyer wavelet, the LFP signals could be further bandpass filtered at a resolution of 1Hz [38]. The WPD decomposes a signal, level by level, from time domain to a time-scale sub-band representation, presenting both approximation and detail subspaces in a binary tree [39]. Wavelet transforms give improved time-frequency resolution over traditional methods, such as the short-term Fourier transform (STFT). The spectrum of the discrete Meyer wavelet matches the spectrum of any band-limited signal in a least squares sense and has been shown to broadly match the oscillatory characteristics of LFP activity [40]. The WPD can be used to extract the standard neural rhythms shown in . It should be noted that all the filters are implemented causally and therefore don't distort the signal preceding an event onset [12]. Each rhythm could then be used as input to the detection algorithm, with its associated detection parameters optimised using either the grid search or multi-objective GA optimisation.

Table 3: The neural rhythms and their associated frequency range.

Neural rhythm	f_L(Hz)	f_H(Hz)
Full-band	1	90
δ	1	4
ϑ	4	8
α	8	12
β_L	12	20
β_H	20	30
γ_L	30	60
γ_H	60	90

gives an example of the detection and optimisation process for Subject 2, left PPN site using the full neural rhythm and the multi-objective GA optimisation approach. The top left of the figure displays the level diagram representation of the Pareto front comprising 2 scatter plots showing the Euclidean norm against N_{FN} and N_{FP} respectively. The Euclidean norm is calculated by normalising the entire Pareto front values to the range [0, 1] and then taking the 2-norm of each normalised objective vector $[N_{FN}, N_{FP}]$. The bottom row of the figure shows the associated Pareto set (or detection parameters) also plotted against the Euclidean norm of the objectives while the top right of the figure shows the full rhythm and its associated detection contour below it. Green and red patches are overlaid onto the neural rhythm and represent the movement phases. The left side of each patch corresponds to a movement onset. A green patch indicates a true positive detection was made while a red patch is indicative of a false negative. Additionally, vertical lines have been superimposed on to the graphs, with a black solid line indicating the position of a true positive neural event. Any neural event detected outside of the movement onset time window but within the movement phase is represented by a blue dashed line. These detections are ignored as the rate of neural activity will be constantly fluctuating throughout volitional movement due to changing cognitive load and motor control processes occurring. Finally, a red dashed line indicates a false positive detection.

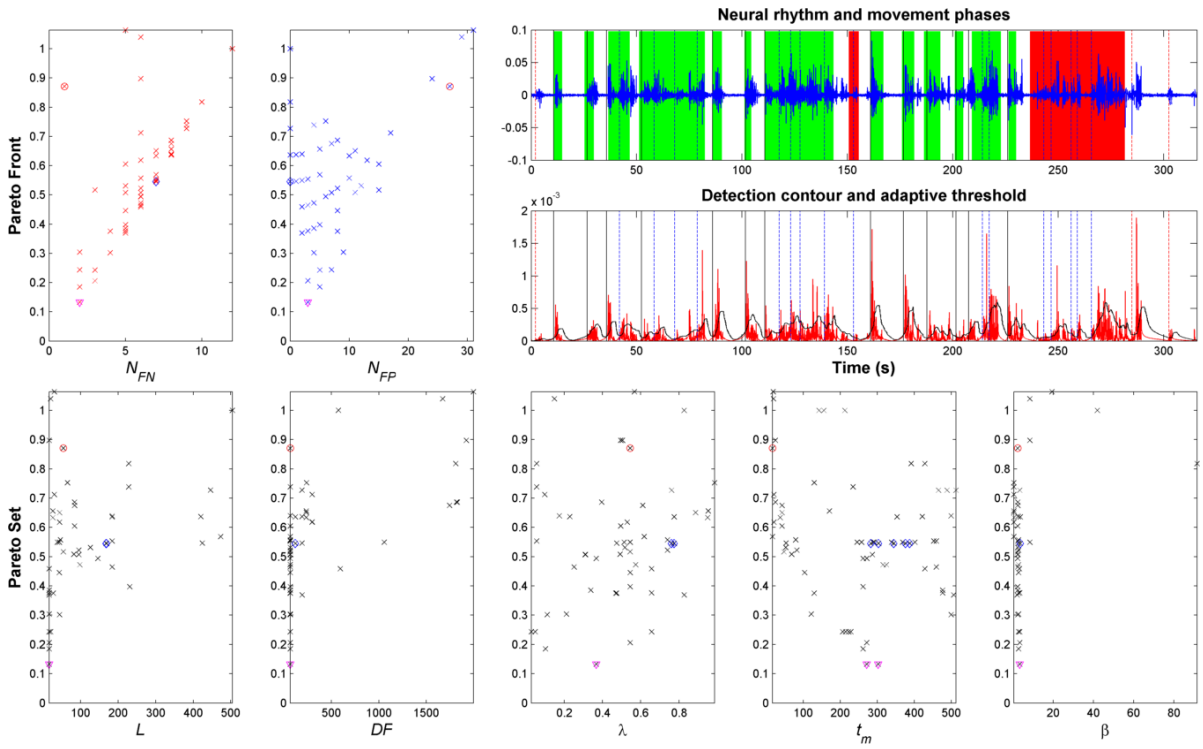


Fig. 4. Example plots associated with the multi-objective GA optimisation of the detection parameters for subject 2 using the full-band left PPN neural rhythm. The bottom row of the figure shows the Pareto optimal set for the five detection parameters using the level diagram representation. On the top row, the leftmost two plots show the corresponding Pareto optimal front for the N_{FN} and N_{FP} detection objectives. The y-axis of these seven plots is the Euclidean norm of the normalised detection objectives. The top right plots show the parameter set selected using the minimum Euclidean norm, with the upper plot showing the bandpass filtered LFP signal and the lower plot showing the associated detection contour and adapted threshold. For a full description of associated colourings and markings, the reader is directed towards the text

The adaptive threshold leading to these detection outcomes is shown on the detection contour as a black solid line on . The threshold corresponds to a set of detection parameters which were chosen based on the minimum Euclidean norm given as a pink inverted triangle on the Pareto plots. Also shown on the Pareto plots are a red circle and blue diamond highlighting the optimal points (based on the search performed) in terms of just the $\min\{N_{FN}\}$ and $\min\{N_{FP}\}$ respectively. It should be noted that multiple optimal points can be selected as different Pareto sets can lead to the same point on the Pareto front (especially as the objectives are discrete).

The Pareto front has been calculated using not only the $\mathrm{J}_1 = N_{FN}$ and $\mathrm{J}_2 = N_{FP}$ but also the following objective function

$$\mathrm{J}_3 = \mathbb{E}\{\boldsymbol{t_{TP}}\} + 2^{2-N_{TP}}$$

where $\mathbb{E}$ denotes the expectation operator and $\boldsymbol{t_{TP}}$ is a set of latencies associated with the true positive detections (with a -ve value indicating that the neural event leads the movement onset while a +ve value indicates that it lags behind). The reason behind using this additional objective is that it allows the Pareto front to be populated with a larger set of solutions, which not only give reasonable behaviour in terms of N_{FN} but incorporates potential solutions which can also provide earlier detection of movement onsets (at the expense of increased N_{FN} and/or N_{FP}). If the only objectives used were J_1 and J_2, then the number of Pareto points is reduced to just the pink, red and blue highlighted points. By including this third objective, it gives the option for other points to be

included, with the potential of these points to only perform slightly worse (i.e. they are weakly dominated) in terms of J_1 and J_2 only, whilst providing earlier detection of the movement onsets. The $2^{2-N_{TP}}$ term prevents low N_{TP} solutions from being included due to them having Pareto optimal lead times but only making a very small number of true positive detections. It should be reiterated that the Euclidean norms have been calculated using just the 2 primary objectives, N_{FN} and N_{FP} as these are primarily what are to be minimised, with the detection latency considered an emergent quantity.

Results for both search methods are presented in , with the minimum Euclidean norm solutions extracted for the four neural sites using the full-band rhythm (1 – 90 Hz). The results are presented in terms of sensitivity (S_{nv}), specificity (S_{fc}), accuracy (Acc) and $\mathbb{E}\{\boldsymbol{t_{TP}}\} \pm 1sd$. The three statistical measures are defined as $S_{nv} = N_{TP}/(N_{TP} + N_{FN})$, $S_{fc} = N_{TN}/(N_{TN} + N_{FP})$ and $Acc = (N_{TP} + N_{TN})/(N_{TP} + N_{TN} + N_{FN} + N_{FP})$. These measures give an indication of the overall performance in a binary test. The detection algorithm along with tuned parameters is able to detect the onset of large scale movement behaviours with a high degree of sensitivity. Subject 1 and 3 achieve over 80% sensitivity across all neural sites, while subject 2 achieves over 80% for the left neural sites and over 60% for the right neural sites when using the GA search. This therefore shows that there is strong correlation between the LFP activity associated with the basal ganglia-brainstem subcortical circuit and the onset of large-scale motion. It should also be noted, that there is a general trend of a relatively decreasing sensitivity in response to a relative increase in specificity and is to be expected.

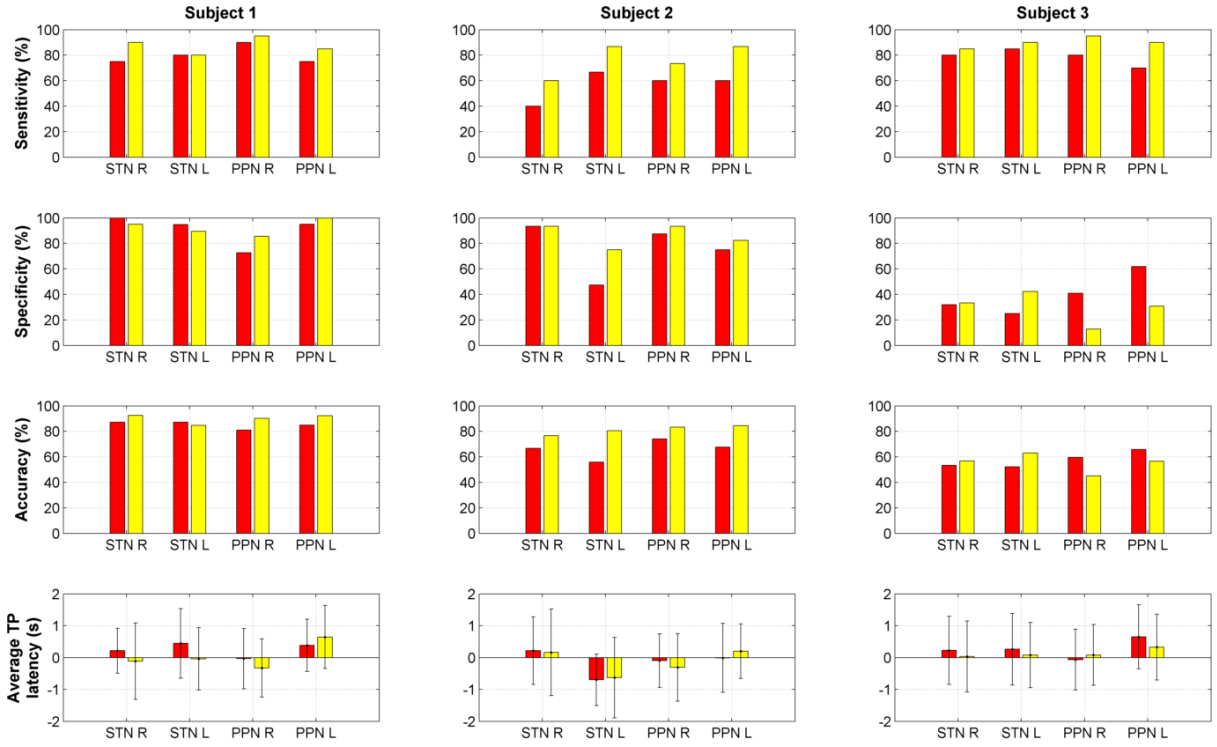


Fig. 5. The Pareto optimal solutions for each subject for the four LFP channels using the full-band LFP signal. The solutions are selected based on the minimum Euclidean norm as given by the minimum point on the level diagrams. Top row - Sensitivity, middle top row - Specificity, middle bottom row - Accuracy, bottom row – Average latencies associated with the TP detections ±1 std. The left red bars show the results using the grid search method while the right yellow bars are for the multi-objective GA optimisation approach.

A possible reason for the reduced sensitivity generally exhibited by subject 2 could be due to their capricious behaviour during the task. At three separate instances, stimulus changes were missed due to subject 2 being unsure and hesitant (only the first one is shown in). This could easily have a detrimental effect in terms of the neural signal detection rates, due for instance, to increased decision processes occurring or worry from the subject, leading to an increased cognitive load. This is especially noticeable in the STN R 12 channel where there is increased neural activity at roughly the halfway point during the experiment. The adaption process had to overcompensate for this and thus missed out on many of the successive movement onsets. Of the three subjects, the specificity associated with subject 3 was significantly worse. This most likely can be attributed to the postural tremor the subject exhibited throughout the rest phases and its associated neural representation being wrongly detected as volitional large-scale activity.

It can be seen that the multi-objective GA outperformed the grid search for the majority of subjects and neural sites across the three performance measures. This is further validated by performing a one-sided paired t-test which yields the following significance levels for the three measures: {Sensitivity, $p = 8.53 \times 10^{-5}$}, {Specificity, $p = 0.447$} and {Accuracy, $p = 0.0423$}. Although the GA specificity isn't significantly better than the grid search (at the 5% level), it should be noted that the grid optimal points were all found on the GA Pareto front and therefore could of ultimately been selected by the designer. The p-value for the expected TP latencies is 0.0614 and therefore although indicates no significant difference between the two search methods, implies that the GA method general found solutions which lead those found using the grid search. The reason for the specificity not exhibiting significance is mainly due to the PPN areas of subject 3. The increase in sensitivity when utilising the GA search method causes a much larger drop in the associated specificity. As mentioned previously this can most likely be attributed to the tremor exhibited by this subject and can be considered a significant part of the intrinsic noise associated with their LFP signals especially within the PPN locations.

The TP latencies show no consistent leading (or lagging) of the neural onsets to the physical onsets and as such, have wide standard deviation bars, with their distribution characterised by a uniform spread across the ±2 time window. Having said this, subjects 1 and 2 show slight average neural leading in the STN L 12, STN R 12 and PPN R 12 channels. In all subjects, the PPN L 12 has the highest TP latency and therefore generally lags the most. Only subject 2 has a brain area that consistently predicts the movement onsets, namely, STN L 12. It should also be noted that the detection contour is calculated causally in a moving window (that varies site-to-site) and therefore the size of this window will also increase the lag time.

By splitting the full-band rhythm down further into the standard narrow-band neural rhythms (as stipulated in), the effect on the detection process can be seen. shows results for the 3 subjects and 4 brain sites for the 7 neural rhythms. The detection parameters have been tuned using the multi-objective GA method. In all cases, there isn't a single band that performs better, as opposed to using the entire band, in terms of both sensitivity and specificity rates. There are generally individual rhythms which perform better than the full rhythm, but this varies subject-to-subject and site-to-site. One trend that should be mentioned is that for subjects 1 and 2, the delta band gives consistently (and comparatively) higher specificity across the neural sites. The converse is true for subject 3 and is potentially due to the LFP oscillations being frequency-coupled to the physical tremor of the individual which occurs at low frequencies of roughly 0.5 - 2 Hz. It is also noted that in

many cases of LFP investigation, it is considered that the very low frequency (δ) band does not carry much event related information and hence is usually filtered out during frequency band pre-processing. However, a recent study reported that the δ band carries significant information about movement-related events [37, 40]. This evidence is also strongly observed in this study. The lag times associated with the TP detections for subject 3 in these frequency bands has also increased. However in general, there are certain (narrow) bands which exhibit a faster lead time when compared to the full band case. However the specific bands are inconsistent subject-to-subject and site-to-site.

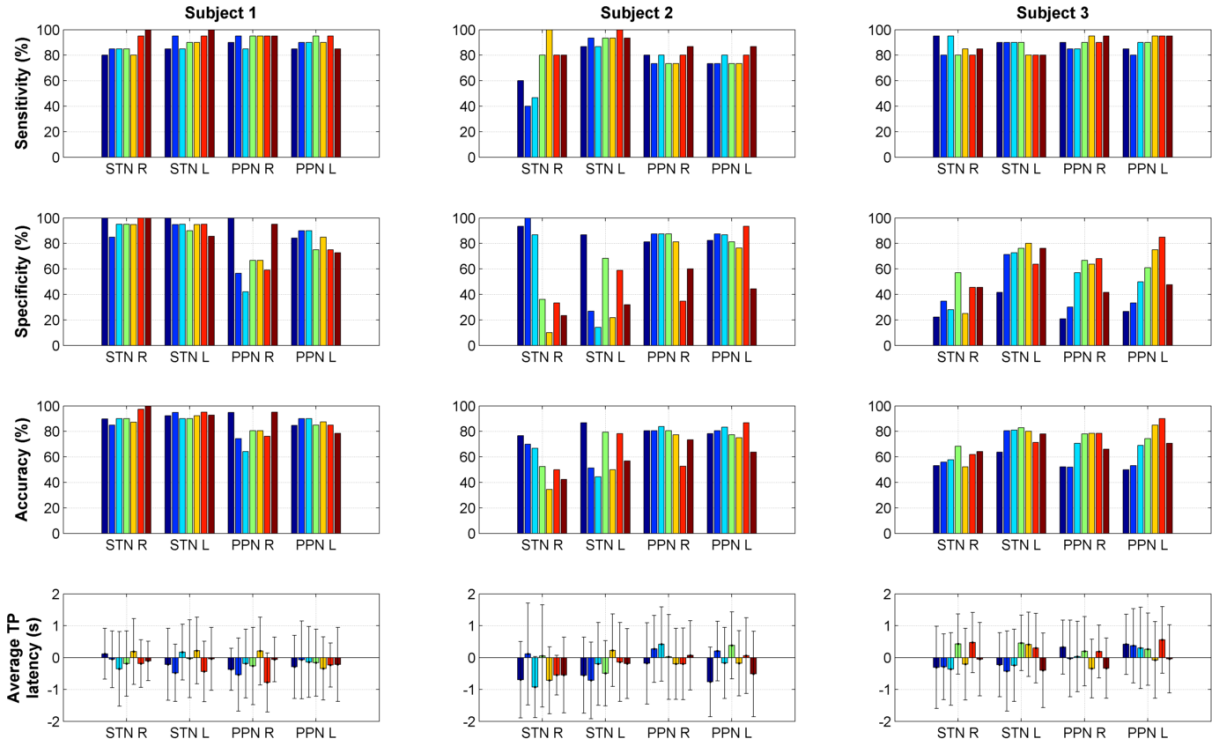


Fig. 6. The Pareto optimal solutions for each subject for the four LFP channels using the individual neural rhythms and the multi-objective GA approach. The solutions are selected based on the minimum Euclidean norm as given by the level diagrams. Top row - Sensitivity, middle top row - Specificity, middle bottom row - Accuracy, bottom row – Average latencies associated with the TP detections ±1 std. The bar groupings from left-right represent the individual neural rhythms from the lowest-highest frequency band.

One of the most significant aspects of this methodology lies in the paradigm shift it introduces in the current practice of clinical neurophysiology, especially in the use of peri-operative neurophysiology in stereotactic and functional neurosurgery. Most, if not all, such applications focus on micro-electrode recordings or on LFPs related to simple motor tasks, i.e., moving the wrist / fingers / toes actively or passively. Heretofore, there has been a hesitation about using LFPs generated by complex, large-scale tasks due to the considerable noise necessarily generated by such multi-system neural tasks. Also, these complex tasks in this severely affected subset of patients are much less amenable to averaging from multiple runs, simply due the limitations of the clinical situation. Refined methods of non-stationary signal analysis, including wavelet analysis, have moved this field forward in recent times (references). This model is an important step, advancing the use of such invasive neurophysiological data by permitting possible automation of detection and recording. This may have significant translational potential in the treatment of complex motor dysfunction like gait and locomotion problems as well as loss of postural control (which leads to frequent falls).

4. Conclusion

This paper has presented a new method for highlighting large-scale behaviours in neural signals. The proposed methodology allows neuronal events to be detected in an automated and systematic fashion. In particular, the algorithm's proposed use is in situations where standard ERP methods fail. This includes situations where there is high inter-trial variability of the ERP, intrinsic noise within the system and/or limited data available. Intrinsic noise is defined as non-ERP signals generated from alternative neural processes which are generally non-stationary, skewed, correlated and/or non-Gaussian. As such, it has a detrimental effect on the averaging process associated with extracting the buried potential when using standard ERP techniques. It should be noted however, that the method does rely on the neural process providing adequate SNR, as it is ultimately based on the relative energy contained within the potential. The algorithm itself is based on a detection contour, computed using a robust estimate of the STE and its instantaneous amplitude. This is utilised in conjunction with an adaptive threshold, with the final binary output based on logical combinations of these two signals. Finally, the associated detection parameters are optimised using an elitist multi-objective genetic algorithm.

Signals recorded from deep brain regions are highly likely to fall into the category of non-standard ERP analysis. This is mainly due to experimental subjects generally having neurological dysfunctions and thus extensive data collection cannot be performed, in particular given obvious constraints of surgical implantation. Deep brain signals such as LFPs are also more likely to have an increased intrinsic-to-extrinsic noise ratio as well as higher SNR then corresponding EEG signals. Additionally, if large scale behaviours are being explored then there will be significant increases in the inter-trial variability and SNR ratios.

To validate the proposed algorithm, neural onsets associated with complex movement behaviours were extracted from LFP's recorded bilaterally in the STN and PPN of Parkinsonian subjects. The exhibited physical behaviour included synergies between gross motor function, gait, vestibular control and muscle coordination in response to a searching and orientation task. As such, the task presented a challenging detection problem. Tested on three PD subjects, the algorithm was able to achieve over 80% sensitivity across the four brain sites, for two subjects (1 and 3) and over 80% sensitivity from the left brain locations and over 60% sensitivity for subject 2 for the full-band rhythms. The reason for the relative reduction in sensitivity of subject 2 can be attributed to them on occasion getting confused while searching for the light source, with the additional cognitive decision making acting as intrinsic noise within the LFP. The specificity associated with subjects 1 and 2 were all greater than 75%, while subject 3 achieved much lower specificity for all four neural sites. This is attributed to subject 3 exhibiting a low frequency postural tremor and thus during the rest phases there would be increased intrinsic noise within the LFP, leading to a significant decrease in specificity for this subject. Further to this analysis, detection was also performed on narrow-band neural rhythms. Although certain individual bands performed better than the full band in general, the analysis did not indicate any particular band or specific neural site being dominant for movement detection.

Using this new methodology, many of the aforementioned limitations associated with standard ERP techniques have been surmounted. This approach has been demonstrated on LFP activity recorded from deep brain structures, within a paradigm that promoted unconstrained and complex

movement behaviours. Despite these challenging conditions, strong correlations between the movement onsets and associated neurological activity were found. Furthermore, a significant benefit of the algorithm is that due to its causal nature, that once optimised, it can be implemented within an online system.

5. Acknowledgements

Nada Yousif is funded by a medical research grant from the Bupa Foundation. This research was supported in part bythe UK Engineering and Physical Sciences (EPSRC) Grant PS6571_MEME

6. References

[1] S. L. Bressler, "Event-Related Potentials," in *The Handbook of Brain Theory and Neural Networks*, M. A. Arbib, Ed. Cambridge MA: MIT Press, 2002, pp. 412–415.

[2] S. J. Luck, *An Introduction to the Event-Related Potential Technique*. MIT Press, 2005, pp. 1–388.

[3] D. Williams, A. Kühn, A. Kupsch, M. Tijssen, G. van Bruggen, H. Speelman, G. Hotton, C. Loukas, and P. Brown, "The relationship between oscillatory activity and motor reaction time in the parkinsonian subthalamic nucleus," *The European journal of neuroscience*, vol. 21, no. 1, pp. 249–58, Jan. 2005.

[4] S. Dandekar, J. Ales, T. Carney, and S. A. Klein, "Methods for quantifying intra- and inter-subject variability of evoked potential data applied to the multifocal visual evoked potential," *Journal of Neuroscience Methods*, vol. 165, no. 2, pp. 270–86, Sep. 2007.

[5] Z. H. Zhou, G. Z. Yuan, J. J. Yao, C. Li, and Z. H. Cheng, "An event-related potential investigation of deficient inhibitory control in individuals with pathological Internet use," *Acta Neuropsychiatrica*, vol. 22, no. 5, pp. 228–236, Sep. 2010.

[6] A. G. Androulidakis, P. Mazzone, V. Litvak, W. Penny, M. Dileone, L. M. F. D. Gaynor, S. Tisch, V. Di Lazzaro, and P. Brown, "Oscillatory activity in the pedunculopontine area of patients with Parkinson's disease," *Experimental Neurology*, vol. 211, no. 1, pp. 59–66, May 2008.

[7] G. Paradiso, D. Cunic, J. A. Saint-Cyr, T. Hoque, A. M. Lozano, A. E. Lang, and R. Chen, "Involvement of human thalamus in the preparation of self-paced movement," *Brain*, vol. 127, no. 12, pp. 2717–31, Dec. 2004.

[8] G. Paradiso, J. A. Saint-Cyr, A. M. Lozano, A. E. Lang, and R. Chen, "Involvement of the human subthalamic nucleus in movement preparation," *Neurology*, vol. 61, no. 11, pp. 1538–1545, Dec. 2003.

[9] I. K. Niazi, N. Jiang, O. Tiberghien, J. F. Nielsen, K. Dremstrup, and D. Farina, "Detection of movement intention from single-trial movement-related cortical potentials," *Journal of neural engineering*, vol. 8, no. 6, p. 066009, Dec. 2011.

[10] P. Rajdev, M. P. Ward, J. Rickus, R. Worth, and P. P. Irazoqui, "Real-time seizure prediction from local field potentials using an adaptive Wiener algorithm," *Computers in Biology and Medicine*, vol. 40, no. 1, pp. 97–108, 2010.

[11] I. Osorio, M. G. Frei, and S. B. Wilkinson, “Real-Time Automated Detection and Quantitative Analysis of Seizures and Short-Term Prediction of Clinical Onset,” *Epilepsia*, vol. 39, no. 6, pp. 615–627, 1998.

[12] D. J. Acunzo, G. Mackenzie, and M. C. W. van Rossum, “Systematic biases in early ERP and ERF components as a result of high-pass filtering,” *Journal of neuroscience methods*, vol. 209, no. 1, pp. 212–218, Jun. 2012.

[13] E. W. Tsang, C. Hamani, E. Moro, F. Mazzella, Y. Y. Poon, A. M. Lozano, and R. Chen, “Involvement of the human pedunculopontine nucleus region in voluntary movements,” *Neurology*, vol. 75, no. 11, pp. 950–9, Sep. 2010.

[14] M. Alegre, F. Alonso-Frech, M. C. Rodríguez-Oroz, J. Guridi, I. Zamarbide, M. Valencia, M. Manrique, J. A. Obeso, and J. Artieda, “Movement-related changes in oscillatory activity in the human subthalamic nucleus: ipsilateral vs. contralateral movements,” *European Journal of Neuroscience*, vol. 22, no. 9, pp. 2315–24, Nov. 2005.

[15] A. Priori, G. Foffani, A. Pesenti, A. Bianchi, V. Chiesa, G. Baselli, E. Caputo, F. Tamma, P. Rampini, M. Egidi, M. Locatelli, S. Barbieri, and G. Scarlato, “Movement-related modulation of neural activity in human basal ganglia and its L-DOPA dependency: recordings from deep brain stimulation electrodes in patients with Parkinson’s disease,” *Neurological Sciences*, vol. 23, pp. 101–102, 2002.

[16] N. Jenkinson, D. Nandi, T. Z. Aziz, and J. F. Stein, “Pedunculopontine nucleus: a new target for deep brain stimulation for akinesia,” *NeuroReport*, vol. 16, no. 17, pp. 1875–6, Nov. 2005.

[17] A. Stefani, A. M. Lozano, A. Peppe, P. Stanzione, S. Galati, D. Tropepi, M. Pierantozzi, L. Brusa, E. Scarnati, and P. Mazzone, “Bilateral deep brain stimulation of the pedunculopontine and subthalamic nuclei in severe Parkinson’s disease,” *Brain*, vol. 130, pp. 1596–607, Jun. 2007.

[18] E. A. C. Pereira, D. Nandi, N. Jenkinson, J. F. Stein, A. L. Green, and T. Z. Aziz, “Pedunculopontine stimulation from primate to patient,” *Journal of neural transmission*, vol. 118, no. 10, pp. 1453–60, Oct. 2011.

[19] M. Cassidy, P. Mazzone, A. Oliviero, A. Insola, P. Tonali, V. Di Lazzaro, and P. Brown, “Movement-related changes in synchronization in the human basal ganglia,” *Brain*, vol. 125, no. 6, pp. 1235–46, Jun. 2002.

[20] M. R. DeLong and T. Wichmann, “Circuits and circuit disorders of the basal ganglia.,” *Archives of neurology*, vol. 64, no. 1, pp. 20–4, Jan. 2007.

[21] J. Mena-Segovia, J. P. Bolam, and P. J. Magill, “Pedunculopontine nucleus and basal ganglia: distant relatives or part of the same family?,” *Trends in Neurosciences*, vol. 27, no. 10, pp. 585–8, Oct. 2004.

[22] K. Takakusaki, K. Saitoh, H. Harada, and M. Kashiwayanagi, “Role of basal ganglia-brainstem pathways in the control of motor behaviors,” *Neuroscience research*, vol. 50, no. 2, pp. 137–51, Oct. 2004.

[23] S. Roy and X. Wang, “Wireless multi-channel single unit recording in freely moving and vocalizing primates,” *Journal of neuroscience methods*, vol. 203, no. 1, pp. 28–40, Jan. 2012.

[24] W. Thevathasan, A. Pogosyan, J. A. Hyam, N. Jenkinson, T. Foltynie, P. Limousin, M. Bogdanovic, L. Zrinzo, A. L. Green, T. Z. Aziz, and P. Brown, “Alpha oscillations in the pedunculopontine nucleus correlate with gait performance in parkinsonism,” *Brain*, vol. 135, no. 1, pp. 148–60, Jan. 2012.

[25] O. F. do Nascimento, K. D. Nielsen, and M. Voigt, “Influence of directional orientations during gait initiation and stepping on movement-related cortical potentials,” *Behavioural Brain Research*, vol. 161, no. 1, pp. 141–54, Jun. 2005.

[26] M. Shoushtarian, A. Murphy, and R. Iansek, “Examination of central gait control mechanisms in Parkinson’s disease using movement-related potentials,” *Movement disorders*, vol. 26, no. 13, pp. 2347–53, Nov. 2011.

[27] I. Osorio and M. G. Frei, “Real-time detection, quantification, warning and control of epileptic seizures: The foundations for a scientific epileptology,” *Epilepsy and Behavior*, vol. 16, no. 3, pp. 391–396, 2009.

[28] N. Jenkinson, D. Nandi, K. Muthusamy, N. J. Ray, R. Gregory, J. F. Stein, and T. Z. Aziz, “Anatomy, physiology, and pathophysiology of the pedunculopontine nucleus,” *Movement disorders*, vol. 24, no. 3, pp. 319–28, Feb. 2009.

[29] D. Nandi, T. Z. Aziz, X. Liu, and J. F. Stein, “Brainstem Motor Loops in the Control of Movement,” *Movement disorders*, vol. 17, no. 3, pp. 22–27, 2002.

[30] J. V. Jacobs and F. B. Horak, “Cortical control of postural responses,” *Journal of neural transmission*, vol. 114, no. 10, pp. 1339–48, Jan. 2007.

[31] A. Singh, S. Kammermeier, A. Plate, J. H. Mehrkens, J. Ilmberger, and K. Bötzel, “Pattern of local field potential activity in the globus pallidus internum of dystonic patients during walking on a treadmill,” *Experimental neurology*, vol. 232, no. 2, pp. 162–7, Dec. 2011.

[32] J. G. Burgess, K. Warwick, V. Ruiz, M. N. Gasson, T. Z. Aziz, J. S. Brittain, and J. Stein, “Identifying tremor-related characteristics of basal ganglia nuclei during movement in the Parkinsonian patient,” *Parkinsonism & Related Disorders*, vol. 16, no. 10, pp. 671–5, Dec. 2010.

[33] C.-S. Wei, C. H. Tsai, S. M. Chiou, Y. T. Hseu, T. W. Liu, H. Y. Lai, T. F. Chien, Y. H. Kuo, W. T. Zhao, Y. S. Tang, S. Y. Su, and Y. Y. Chen, “Nonlinear analysis of movement-related changes in human subthalamic local field potentials,” in *Proc. of the 32nd Annual Int. Conf. of the IEEE Engineering in Medicine and Biology Society*, 2010, vol. 2010, pp. 4132–5.

[34] X. Blasco, J. M. Herrero, J. Sanchis, and M. Martínez, “A new graphical visualization of n-dimensional Pareto front for decision-making in multiobjective optimization,” *Information Sciences*, vol. 178, no. 20, pp. 3908–3924, Oct. 2008.

[35] A. Konak, D. W. Coit, and A. E. Smith, “Multi-objective optimization using genetic algorithms: A tutorial,” *Reliability Engineering & System Safety*, vol. 91, no. 9, pp. 992–1007, Sep. 2006.

[36] K. Deb, “Multi-objective genetic algorithms: Problem difficulties and construction of test problems.,” *Evolutionary computation*, vol. 7, no. 3, pp. 205–30, Jan. 1999.

[37] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, “A Fast and Elitist Multiobjective Genetic Algorithm: NSGA-II,” *IEEE Transactions on Evolutionary Computation*, vol. 6, no. 2, pp. 182–197, 2002.

[38] K. A. Mamun, M. Mace, M. E. Lutman, J. Stein, X. Liu, T. Aziz, R. Vaidyanathan, and S. Wang, “A Robust Strategy for Decoding Movements from Deep Brain Local Field Potentials to Facilitate Brain Machine Interfaces,” in *In Proc. of the 4th Int Conf. on Biomedical Robotics and Biomechatronics*, 2012.

[39] S. Mallet, *A wavelet tour of signal processing*, 2nd ed. Academic Press, 1998, pp. 1–668.

[40] V. J. Samar, A. Bopardikar, R. Rao, and K. Swartz, “Wavelet analysis of neuroelectric waveforms: A conceptual tutorial,” *Brain and language*, vol. 66, no. 1, pp. 7–60, Jan. 1999.

Appendix D

Publication list

D.1 Journal Papers (*Pending*)

- M. Mace, K. A. Mamun, A. A. Naeem, L. Gupta, S. Wang and R. Vaidyanathan,"A heterogeneous framework for real-time decoding of bioacoustic tongue signals: Applications to assistive interfaces and prosthesis control," (*Submitted to Expert Systems with Applications*)
- M. Mace, N. Yousif, M. Naushahi, K. A. Mamun, S. Wang, D. Nandi, R. Vaidyanathan, "An automated approach towards detecting complex behaviours from deep brain structures," (*Submitted* to the Journal of Neuroscience Methods, Elsevier)
- K. A. Mamun, M. Mace, M. E. Lutman, R. Vaidyanathan and S. Wang, "Weighted Sequential Feature Selection (WSFS): an efficient strategy to facilitate high dimensional small sample size biosignal classification," (*In preparation*, to be submitted in IEEE Transactions on Neural Systems and Rehabilitation Engineering or IEEE Transactions on Pattern Analysis and Machine Intelligence)
- M. Mace, L. Gupta and R. Vaidyanathan, "Characterisation of a novel robust template creation process for exemplifying morphological patterns in time series data sets," (*In preparation*)

D.2 Journal Papers (*Accepted*)

- K. A. Mamun, M. Mace, L. Gupta, C. A. Verschuur, M. E. Lutman, M. Stokes, R. Vaidyanathan, S. Wang, "Robust real-time identification of tongue movement commands from interferences," Neurocomputing, Elsevier, vol. 80, pp 83-92, 2011
- M. Mace, R. Vaidyanathan, S. Wang and L. Gupta, "Tongue in cheek: A novel concept in assistive human-machine interface," Journal of Assistive Technologies, vol. 3, pp

14-26, 2009

D.3 Conference Papers (*Accepted*)

- K. A. Mamun, M. Mace, R. Vaidyanathan, M. E. Lutman, J. Stein, X. Liu , T. Aziz, and S. Wang, "A robust strategy for decoding movements from deep brain local field potentials to facilitate brain machine interfaces," 4th IEEE RAS/EMBS International Conference on Biomedical Robotics and Biomechatronics, pp.320-325, Roma, Italy, Jun. 24-28, 2012

- M. Mace, K. A. Mamun, S. Wang, L. Gupta and R. Vaidyanathan, "Ensemble classification for robust discrimination of multi-channel, multi-class tongue-movement ear pressure signals," Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp.1733-1736, Boston, USA, Aug. 30-Sept. 3 2011 (*Attended and presented*)

- K. A. Mamun, M. Banik, M. Mace, M. E. Lutman, R. Vaidyanathan and S. Wang, "Multi-layer neural network classification of tongue movement ear pressure signal for human machine interface," Proceedings of the 13th International Conference on Computer And Information Technology (ICCIT 2010), Dhaka, Bangladesh, Dec. 23-25, 2010

- M. Mace, K. A. Mamun, Ravi Vaidyanathan, Shouyan Wang and Lalit Gupta, "Real-time Implementation of a Non-invasive Tongue-based Human-Robot Interface," Proceedings of the 2010 IEEE/RSJ International Conference on Intelligent Robots and Systems, pp.5486 - 5491, Taipei, Taiwan, Oct. 18-22, 2010

- M. Mace, K. A. Mamun, S. Wang, L. Gupta and R. Vaidyanathan, "Human-machine interface for tele-robotic operation using tongue movement ear pressure (TMEP) signals," Proceedings of the 11th Annual Towards Autonomous Robotic Systems Conference (TAROS), pp.148-153, Plymouth, UK, Aug. 31-Sep. 2, 2010 (*Attended and presented*)

- K. A. Mamun, M. Mace, M. E. Lutman, R. Vaidyanathan, L. Gupta and S. Wang, "Multivariate Bayesian classification of tongue movement ear pressure signals based on the wavelet packet transform," Proceedings of the 2010 IEEE International Workshop on Machine Learning for Signal Processing, pp.208-213 , Kittilä, Finland, Aug. 29-Sep. 1, 2010

- S. Kota, M. Mace, L. Gupta and R. Vaidyanathan, "A DCT-Gaussian Classification Scheme for Human-Robot Interface," Proceedings of the 2009 International Confer-

ence on Intelligent Robots and Systems, pp.5503-5508, St Louie, USA, Oct. 10-15, 2009

Bibliography

[1] O. Tonet, M. Marinelli, L. Citi, P. M. Rossini, L. Rossini, G. Megali, and P. Dario, "Defining brain-machine interface applications by matching interface performance with device requirements.," *Journal of neuroscience methods*, vol. 167, pp. 91–104, Jan. 2008. 1, 10, 14, 29, 41, 178

[2] B. A. Myers, "A Brief History of Interaction Technology," *ACM interactions*, vol. 5, no. 2, pp. 44–54, 1998. 1, 12, 17

[3] T. M. Vaughan, J. R. Wolpaw, and E. Donchin, "EEG-based communication: prospects and problems," *IEEE Transactions on Rehabilitation Engineering*, vol. 4, no. 4, pp. 425–430, 1996. 2, 29

[4] R. Raya, D. Torricelli, E. Rocon, R. Ceres, A. Frizera, and J. L. Pons, "Feasibility in human machine interfaces for elderly people," *Gerontechnology*, vol. 7, Apr. 2008. 2, 11, 12

[5] A. V. Heest, "Congenital Upper Extremity Differences: Undestanding the Etiology and Treatment," *A Pediatric Perspective*, vol. 10, no. 6, pp. 1–4, 2001. 2, 12

[6] C. M. Gibson, M. D. Donnelly, and T. Turner, "One Degree of Separation: Paralysis and Spinal Cord Injury in the United States," tech. rep., Christopher and Dana Reeve Foundation, 2009. 12, 13

[7] J. Bourn, "Reducing Brain Damage : Faster access to better stroke care," tech. rep., National Audit Office, 2006. 12, 13

[8] A. Compston and A. Coles, "Multiple sclerosis," *Lancet*, vol. 372, pp. 1502–17, Oct. 2008. 13

[9] G. Bauer, F. Gerstenbrand, and E. Rumpl, "Varieties of the locked-in syndrome," *Journal of neurology*, vol. 221, pp. 77–91, Aug. 1979. 13

[10] E. Smith and M. Delargy, "Clinical review: Locked-in syndrome," *Bristish Medical Journal*, vol. 330, pp. 406–409, 2005. 2, 13

[11] W. H. Organisation, "The world report on disability," tech. rep., World Health Organisation, July 2011. 3, 13

[12] C. G. Pinheiro, E. L. M. Naves, P. Pino, E. Losson, A. O. Andrade, and G. Bourhis, "Alternative communication systems for people with severe motor disabilities: a survey," *Biomedical Engineering Online*, vol. 10, pp. 1–31, Jan. 2011. 3, 10, 20

[13] S. Guo, R. A. Cooper, M. L. Boninger, A. Kwarciak, and B. Ammer, "Development of power wheelchair chin-operated force-sensing joystick," in *Proc. of the Second Joint EMBS/BMES Conf.*, vol. 3, (Houstan), pp. 2373–2374, IEEE, 2002. 3, 18

[14] M. Betke, J. Gips, and P. Fleming, "The Camera Mouse: Visual Tracking of Body Features to Provide Computer Access for People With Severe Disabilities," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 10, no. 1, pp. 1–10, 2002. 3, 19

[15] L. J. Hadjileontiadis, S. M. Panas, and I. T. Rekanos, "Bioaoustic signals," in *Wiley Encyclopedia of Biomedical Engineering*, pp. 1–17, 2006. 3, 4, 49, 107

[16] L. K. Shark, H. Chen, and J. Goodacre, "Knee acoustic emission: A potential biomarker for quantitative assessment of joint ageing and degeneration.," *Medical Engineering and Physics*, vol. 33, pp. 534–545, July 2011. 4

[17] C. Harrison, D. Tan, and D. Morris, "Skinput : Appropriating the Body as an Input Surface," in *In Proc. of the 2010 ACM Conf. on Human Factors in Computing Systems*, (Atlanta), pp. 453–462, ACM, 2010. 4, 52, 81

[18] S. Harada, J. A. Landay, J. Malkin, X. Li, and J. A. Bilmes, "The Vocal Joystick: Evaluation of Voice-based Cursor Control Techniques," *Disability and Rehabilitation: Assistive Technology*, vol. 3, no. 1, pp. 22–34, 2007. 36, 37

[19] T. Simpson, M. Gauthier, and A. Prochazka, "Evaluation of tooth-click triggering and speech recognition in assistive technology for computer access.," *Neurorehabilitation and neural repair*, vol. 24, pp. 188–94, Feb. 2010. 4, 37, 38, 41

[20] M. Ghovanloo, "Tongue operated assistive technologies.," in *Proc. of the 29th Annual Int. Conf. of the IEEE Engineering in Medicine and Biology Society*, vol. 2007, (Lyon), pp. 4376–9, IEEE, Jan. 2007. 5, 33

[21] G. Fesl, B. Moriggl, U. D. Schmid, T. P. Naidich, K. Herholz, and T. A. Yousry, "Inferior central sulcus: variations of anatomy and function on the example of the motor tongue area," *NeuroImage*, vol. 20, pp. 601–610, Sept. 2003. 5, 38

[22] J. Watanabe, M. Sugiura, N. Miura, Y. Watanabe, Y. Maeda, Y. Matsue, and R. Kawashima, "The human parietal cortex is involved in spatial processing of tongue movement-an fMRI study," *NeuroImage*, vol. 21, pp. 1289–99, Apr. 2004.

[23] A. G. He, L. H. Tan, Y. Tang, G. A. James, P. Wright, M. A. Eckert, P. T. Fox, and Y. Liu, "Modulation of neural connectivity during tongue movement and reading," *Human brain mapping*, vol. 18, pp. 222–32, Mar. 2003. 5, 38

[24] R. Vaidyanathan, B. Chung, L. Gupta, H. Kook, S. Kota, and J. D. West, "Tongue-Movement Communication and Control Concept for Hands-Free Human-Machine Interfaces," *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, vol. 37, no. 4, pp. 533–546, 2007. 5, 55, 56, 58, 73, 77, 84, 87, 88, 130, 132, 152, 197, 198, 199

[25] R. Vaidyanathan, M. P. Fargues, R. Serdar Kurcan, L. Gupta, S. Kota, R. D. Quinn, and D. Lin, "A Dual Mode Human-Robot Teleoperation Interface Based on Airflow in the Aural Cavity," *The International Journal of Robotics Research*, vol. 26, pp. 1205–1223, Nov. 2007. 5, 55, 58, 59

[26] L. R. Hochberg and J. P. Donoghue, "Sensors for Brain-Computer Interfaces: Options for Turning Thought into Action," *IEEE Engineering in Medicine and Biology Magazine*, no. October, pp. 32 – 38, 2006. 6, 14, 15, 30, 32

[27] K. A. Mamun, M. Mace, L. Gupta, C. A. Verschuur, M. E. Lutman, M. Stokes, R. Vaidyanathan, and S. Wang, "Robust real-time identification of tongue movement commands from interferences," *Neurocomputing*, vol. 80, pp. 83–92, Nov. 2011. 9, 62, 157, 173, 174, 175, 183, 195, 197, 200, 205, 207

[28] A. M. Cook and J. M. Polgar, *Assistive Technologies: Principles and Practice*. Mosby, 3rd ed., 2007. 10

[29] U. S. Congress, "Assistive Technology Act of 2004," tech. rep., 2004. 11

[30] M. R. DeLong and T. Wichmann, "Circuits and circuit disorders of the basal ganglia.," *Archives of neurology*, vol. 64, pp. 20–4, Jan. 2007. 12

[31] S. P. Agnew, J. Ko, M. De La Garza, T. Kuiken, and G. Dumanian, "Limb transplantation and targeted reinnervation: a practical comparison.," *Journal of reconstructive microsurgery*, vol. 28, pp. 63–8, Jan. 2011. 12

[32] M. Tibble, "User's guide to disability estimates and definitions," Tech. Rep. March, Department for Work and Pensions, 2004. 13

[33] G. Brown, *The living end: the future of death, aging and immortality*. Palgrave Macmillan, 2007. 13

[34] K. Carroll, S. Murad, and A. Majeed, "Stroke incidence and risk factors in a population - based prospective cohort study," *Health Statistics Quarterly*, vol. 12, pp. 18–26, 2001. 13

[35] K. Ziegler-Graham, E. J. Mackenzie, P. L. Ephraim, T. G. Travison, and R. Brookmeyer, "Estimating the prevalence of limb loss in the United States: 2005 to 2050," *Archives of physical medicine and rehabilitation*, vol. 89, pp. 422–9, Mar. 2008. 13

[36] J. Qiu, "China Spinal Cord Injury Network: changes from within," *The Lancet Neurology*, vol. 8, pp. 606–607, July 2009. 13

[37] W. W. Abbott and A. A. Faisal, "Ultra-low-cost 3D gaze estimation: an intuitive high information throughput compliment to direct brain-machine interfaces," *Journal of Neural Engineering*, vol. 9, pp. 1–11, Aug. 2012. 14, 24, 25, 26, 27, 32, 41

[38] R. E. Cowan, B. J. Fregly, M. L. Boninger, L. Chan, M. M. Rodgers, and D. J. Reinkensmeyer, "Recent trends in assistive technology for mobility," *Journal of neuroengineering and rehabilitation*, vol. 9, p. 20, Jan. 2012. 14

[39] J. Musić, M. Cecić, and M. Bonković, "Experimental Measurement of Information Transfer Rate of the Inertial Sensor Based Human-Computer Interface for the Disabled," vol. 8, no. 9, pp. 1072–1082, 2009. 17, 18, 41

[40] J. Silva, T. Chau, and A. Goldenberg, "MMG-based multisensor data fusion for prosthesis control," in *Proc. of the 25th Annual Int. Conf. of the IEEE Engineering in Medicine and Biology Society*, pp. 2909–2912, Ieee, 2003. 17, 19

[41] I. S. MacKenzie, T. Kauppinen, and M. Silfverberg, "Accuracy measures for evaluating computer pointing devices," *Proceedings of the SIGCHI conference on Human factors in computing systems - CHI '01*, pp. 9–16, 2001. 17, 41

[42] R. Raya, J. Roa, E. Rocon, R. Ceres, and J. Pons, "Wearable inertial mouse for children with physical and cognitive impairments," *Sensors and Actuators A: Physical*, vol. 162, pp. 248–259, Aug. 2010. 18

[43] R. G. Radwin, G. C. Vanderheiden, and M. L. Lin, "A method for evaluating head-controlled computer input devices using Fitts' law.," Aug. 1990. 18

[44] D. G. Evans, R. Drew, and P. Blenkhorn, "Controlling mouse pointer position using an infrared head-operated joystick," *IEEE Transactions on rehabilitation engineering*, vol. 8, no. 1, pp. 107–117, 2000.

[45] Y. L. Chen, F. T. Tang, W. H. Chang, M. K. Wong, Y. Y. Shih, and T. S. Kuo, "The new design of an infrared-controlled human-computer interface for the disabled," *IEEE Transactions on Rehabilitation Engineering*, vol. 7, no. 4, pp. 474–481, 1999. 18, 41

[46] O. Takami, N. Irie, C. Kang, T. Ishimatsu, and T. Ochiai, "Computer interface to use head movement for handicapped people," in *Proc. of the IEEE TENCON on Digital Signal Processing Applications*, pp. 468–472, IEEE, 1996. 18

[47] A. Frisoli, F. Salsedo, M. Bergamasco, B. Rossi, and M. C. Carboncini, "A force-feedback exoskeleton for upper-limb rehabilitation in virtual reality," *Applied Bionics and Biomechanics*, vol. 6, pp. 115–126, July 2009. 18

[48] M. Baklouti, J. AbouSaleh, E. Monacelli, and S. Couvet, "Human Machine Interface in Assistive Robotics: Application to a Force Controlled Upper-Limb Powered Exoskeleton," in *Robotics 2010 Current and Future Challenges*, ch. 12, pp. 211–222, InTech, 2010. 18

[49] T. Harada, T. Sato, and T. Mori, "Pressure distribution image based human motion tracking system using skeleton and surface," in *Proc. of the IEEE Int. Conf. on Robotics and Automation*, (Seoul), pp. 3201–3207, IEEE, 2001. 18

[50] T. Tezuka, A. Goto, K. Kashiwa, H. Yoshikawa, and R. Kawano, "A study on space interface for teleoperation system," in *Proc. of the 3rd IEEE International Workshop on Robot and Human Communication*, pp. 62–67, IEEE, 1994. 18, 19

[51] K. Taniguchi, A. Nishikawa, S. Kawanishi, and F. Miyazaki, "KOMEKAMI Switch: A Novel Wearable Input Device Using Movement of Temple," *Journal of Robotics and Mechatronics*, vol. 20, no. 2, 2008. 19, 41

[52] R. B. Reilly and M. J. O'Malley, "Adaptive noncontact gesture-based system for augmentative communication.," *IEEE transactions on Rehabilitation Engineering*, vol. 7, pp. 174–82, June 1999. 20

[53] A. Kushki, A. J. Andrews, S. D. Power, G. King, and T. Chau, "Classification of activity engagement in individuals with severe physical disabilities using signals of the peripheral nervous system.," *PloS one*, vol. 7, p. e30373, Jan. 2012. 20

[54] B. U. Köhler, C. Hennig, and R. Orglmeister, "The principles of software QRS detection," *IEEE Engineering in Medicine and Biology magazine*, vol. 21, no. 1, pp. 42–57, 2002. 20, 79, 80

[55] R. Ahsan and M. I. Ibrahimy, "EMG Signal Classification for Human Computer Interaction: A Review," *European Journal of Scientific Research*, vol. 33, no. 3, pp. 480–501, 2009. 20

[56] M. A. Oskoei and H. Hu, "Myoelectric control systems - A survey," *Biomedical Signal Processing And Control*, vol. 2, pp. 275–294, 2007. 20, 21

[57] X. Navarro, T. B. Krueger, N. Lago, S. Micera, T. Stieglitz, and P. Dario, "A critical review of interfaces with the peripheral nervous system for the control of neuroprostheses and hybrid bionic systems," *Journal of the peripheral nervous system*, vol. 10, pp. 229–58, Sept. 2005. 21, 22

[58] T. Felzer and B. Freisleben, "HaWCoS: The "Hands-free" Wheelchair Control System," in *Proc. of the ACM SIGCAPH Conf. on Assistive Technologies*, (Edinburgh), pp. 127–134, ACM Press, 2002. 21

[59] M. Gasson, B. Hutt, I. Goodhew, P. Kyberd, and K. Warwick, "Bi-directional human machine interface via direct neural connection," in *Proc. of the 11th IEEE Int Workshop on Robot and Human Interactive Communication*, (Berlin), pp. 265–270, IEEE, 2002.

[60] M. A. Oskoei and H. Hu, "Myoelectric based virtual joystick applied to electric powered wheelchair," in *Proc. of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, (Nice), pp. 2374–2379, IEEE, 2008. 21

[61] C. Choi, H. Han, C. Kim, and J. Kim, "Development and Performance Evaluation of a Neural Signal-based Assistive Computer Interface," in *Proc. of the 16th IEEE Int. Conf. on Robot and Human Interactive Communication*, (Jeju), pp. 944–949, IEEE, 2007. 21

[62] T. R. Farrell and R. F. Weir, "Pilot Comparison of Surface vs. Implanted EMG for Multifunctional Prosthesis Control," in *Proc. of the IEEE 9th Int. Conf. on Rehabilitation Robotics*, (Chicago), pp. 277–280, IEEE, 2005. 21

[63] S. Micera, P. M. Rossini, J. Rigosa, L. Citi, J. Carpaneto, S. Raspopovic, M. Tombini, C. Cipriani, G. Assenza, M. C. Carrozza, K. P. Hoffmann, K. Yoshida, X. Navarro, and P. Dario, "Decoding of grasping information from neural signals recorded using peripheral intrafascicular interfaces.," *Journal of neuroengineering and rehabilitation*, vol. 8, p. 53, Jan. 2011. 21, 22

[64] M. Gasson, B. Hutt, I. Goodhew, P. Kyberd, and K. Warwick, "Invasive neural prosthesis for neural signal detection and nerve stimulation," *International Journal of Adaptive Control and Signal Processing*, vol. 19, pp. 365–375, 2005. 22

[65] R. F. Weir, P. R. Troyk, G. A. DeMichele, D. A. Kerns, J. F. Schorsch, and H. Maas, "Implantable Myoelectric Sensors (IMESs) for Intramuscular Electromyogram Recording," *IEEE Transactions on Biomedical Engineering*, vol. 56, pp. 159–171, Jan. 2009. 22

[66] J. Badia, T. Boretius, D. Andreu, C. Azevedo-Coste, T. Stieglitz, and X. Navarro, "Comparative analysis of transverse intrafascicular multichannel, longitudinal intrafascicular and multipolar cuff electrodes for the selective stimulation of nerve fascicles," *Journal of neural engineering*, vol. 8, pp. 1–13, June 2011. 22, 23

[67] L. A. Miller, K. A. Stubblefield, R. D. Lipschutz, B. A. Lock, and T. A. Kuiken, "Improved Myoelectric Prosthesis Control Using Targeted Reinnervation Surgery: A Case Series," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 16, no. 1, pp. 46–50, 2008. 22

[68] T. a. Kuiken, L. a. Miller, R. D. Lipschutz, B. a. Lock, K. Stubblefield, P. D. Marasco, P. Zhou, and G. a. Dumanian, "Targeted reinnervation for enhanced prosthetic arm function in a woman with a proximal amputation: a case study.," *Lancet*, vol. 369, pp. 371–80, Feb. 2007. 22

[69] T. A. Kuiken, B. A. Lock, R. D. Lipschutz, L. A. Miller, K. A. Stubblefield, and K. B. Englehart, "Targeted Muscle Reinnervation for Real-time Myoelectric Control of Multifunction Artificial Arms," *Journal of the American Medical Association*, vol. 301, no. 6, pp. 619–628, 2009. 23, 41

[70] N. Memarian, A. N. Venetsanopoulos, and T. Chau, "Infrared thermography as an access pathway for individuals with severe motor impairments.," *Journal of neuroengineering and rehabilitation*, vol. 6, p. 11, 2009. 24, 41

[71] N. Memarian, A. N. Venetsanopoulos, and T. Chau, "Body Functions and Structures Pertinent to Infrared Thermography-Based Access for Clients With Severe Motor Disabilities," *Assistive Technology*, vol. 23, pp. 53–64, June 2011. 24

[72] H. Yunfei and P. Phukpattaranont, "Wheelchair Control Based on Bioimpedance," vol. 3, no. 1, pp. 13–17, 2010. 24

[73] N. Alves and T. Chau, "Mechanomyography as an access pathway: corporeal contraindications," *Disability and rehabilitation: Assistive technology*, vol. 6, pp. 552–63, Jan. 2011. 24

[74] J. Silva and T. Chau, "A mathematical model for source separation of MMG signals recorded with a coupled microphone-accelerometer sensor pair," *IEEE Transactions on Biomedical Engineering*, vol. 52, pp. 1493–501, Sept. 2005. 24, 41, 50

[75] M. J. Stokes, "Acoustic Myography: Applications and considerations in measuring muscle performance," *Isokinetics and Exercise Science*, vol. 3, no. 1, pp. 4–16, 1993. 24, 50

[76] C. Topal, O. N. Gerek, and A. Dogan, "A head-mounted sensor-based eye tracking device: Eye touch system," in *Proc. of the 2008 symposium on Eye tracking research and applications*, vol. 1, pp. 87–90, 2008. 25

[77] C. Hennessey and P. Lawrence, "Noncontact binocular eye-gaze tracking for point-of-gaze estimation in three dimensions," *IEEE transactions on biomedical engineering*, vol. 56, pp. 790–9, Mar. 2009. 25, 26

[78] D. Rozado, F. B. Rodriguez, and P. Varona, "Low cost remote gaze gesture recognition in real time," *Applied Soft Computing*, vol. 12, pp. 2072–2084, Aug. 2012. 25, 27

[79] A. B. Usakli, S. Gurkan, F. Aloise, G. Vecchiato, and F. Babiloni, "On the Use of Electrooculogram for Efficient Human Computer Interfaces," *Computational intelligence and neuroscience*, pp. 1 – 5, Jan. 2010. 25, 27

[80] T. Doyle, Z. Kucerovsky, and W. Greason, "Design of an Electroocular Computing Interface," in *Proc. of the IEEE Canadian Conference on Electrical and Computer Engineering*, no. May, (Ottawa), pp. 1458–1461, IEEE, 2006. 27

[81] A. Bulling, J. A. Ward, H. Gellersen, and G. Tröster, "Robust Recognition of Reading Activity in Transit Using Wearable Electrooculography," *Pervasive*, vol. LNCS 5013, pp. 19–37, 2008. 25

[82] D. D. Salvucci and J. R. Anderson, "Intelligent gaze-added interfaces," in *Proc. of the SIGCHI conf. on Human factors in computing systems*, vol. 2, (New York, New York, USA), pp. 273–280, ACM Press, 2000. 25

[83] R. Atienza and A. Zelinsky, "Intuitive interface through active 3D gaze tracking," in *Proc. of the Int. Conf. on Active Media Technology (AMT 2005)*, pp. 16–21, IEEE, 2005.

[84] J. P. Hansen, K. Tørning, A. S. Johansen, K. Itoh, and H. Aoki, "Gaze typing compared with input by head and hand," in *Proc. of the Eye tracking research and applications symposium on Eye tracking research and applications - ETRA 2004*, vol. 1, (New York), pp. 131–138, ACM Press, 2004. 25

[85] D. Decker and J. A. Piepmeier, "Gaze Tracking Interface for Robotic Control," in *Proc. of the 40th Southeastern Symposium on System Theory (SSST)*, (New Orleans), pp. 274–278, IEEE, 2008. 26

[86] E. D. Guestrin and M. Eizenman, "General theory of remote gaze estimation using the pupil center and corneal reflections," *IEEE transactions on bio-medical engineering*, vol. 53, pp. 1124–33, June 2006. 26

[87] R. Barea, L. Boquete, J. M. Rodriguez-Ascariz, S. Ortega, and E. López, "Sensory system for implementing a human-computer interface based on electrooculography," *Sensors*, vol. 11, pp. 310–28, Jan. 2011. 27

[88] A. Bulling, J. A. Ward, H. Gellersen, and G. Tröster, "Eye movement analysis for activity recognition using electrooculography.," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, pp. 741–53, Apr. 2011. 27, 28

[89] R. Shaw, E. Crisman, A. Loomis, and Z. Laszewski, "The eye wink control interface: using the computer to provide the severely disabled with increased flexibility and comfort," in *Proc. of the Third Annual IEEE Symposium on Computer-Based Medical Systems*, pp. 105–111, IEEE, 1990. 28

[90] J. P. Donoghue, "Connecting cortex to machines: recent advances in brain interfaces," *Nature Neuroscience*, vol. 5, pp. 1085–1088, 2002. 29

[91] E. C. Leuthardt, G. Schalk, D. Moran, and J. G. Ojemann, "The emerging world of motor neuroprosthetics: a neurosurgical perspective.," *Neurosurgery*, vol. 59, no. 1, pp. 1–14; discussion 1–14, 2006. 29

[92] Z. A. Keirn and J. I. Aunon, "Man-machine communications through brain-wave processing," *IEEE Engineering in Medicine and Biology Magazine*, vol. 9, pp. 55–57, Mar. 1990. 29

[93] J. R. Wolpaw, N. Birbaumer, D. J. McFarland, G. Pfurtscheller, and T. M. Vaughan, "Brain-computer interfaces for communication and control," *Clinical neurophysiology*, vol. 113, pp. 767–91, June 2002. 31, 192

[94] M. Fatourechi, R. K. Ward, and G. E. Birch, "A self-paced brain-computer interface system with a low false positive rate.," *Journal of neural engineering*, vol. 5, pp. 9–23, Mar. 2008.

[95] M. K. Hazrati and A. Erfanian, "An online EEG-based brain-computer interface for controlling hand grasp using an adaptive probabilistic neural network," *Medical Engineering and Physics*, vol. 32, pp. 730–9, Sept. 2010. 29

[96] L. F. Nicolas-Alonso and J. Gomez-Gil, "Brain computer interfaces, a review," *Sensors*, vol. 12, pp. 1211–79, Jan. 2012. 30, 31, 108, 109, 110, 134, 145, 153

[97] K. A. Mamun, M. Mace, M. E. Lutman, J. Stein, X. Liu, T. Aziz, R. Vaidyanathan, and S. Wang, "A Robust Strategy for Decoding Movements from Deep Brain Local Field Potentials to Facilitate Brain Machine Interfaces," in *In Proc. of the 4th Int Conf. on Biomedical Robotics and Biomechatronics*, (Roma), IEEE, 2012. 31, 33, 120, 127, 174, 199, 206, 207

[98] A. Nijholt and D. Tan, "Brain-Computer Interfacing for Intelligent Systems," *IEEE Intelligent Systems*, vol. 23, pp. 72–79, May 2008. 32

[99] T. N. Lal, T. Hinterberger, G. Widman, M. Schr, J. Hill, W. Rosenstiel, C. E. Elger, and B. Sch, "Methods Towards Invasive Human Brain Computer Interfaces," in *IN ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS 17*, pp. 737–744, MIT Press, 2005. 32

[100] J. P. Donoghue, A. Nurmikko, M. Black, and L. R. Hochberg, "Assistive technology and robotic control using motor cortex ensemble-based neural interface systems in humans with tetraplegia," *The Journal of physiology*, vol. 579, no. 3, pp. 603–611, 2007. 32

[101] L. R. Hochberg, D. Bacher, B. Jarosiewicz, N. Y. Masse, J. D. Simeral, J. Vogel, S. Haddadin, J. Liu, S. S. Cash, P. V. D. Smagt, and J. P. Donoghue, "Reach and grasp by people with tetraplegia using a neurally controlled robotic arm," *Nature*, vol. 485, no. 7398, pp. 372–375, 2012. 32

[102] C. Mehring, J. Rickert, E. Vaadia, S. Cardosa de Oliveira, A. Aertsen, and S. Rotter, "Inference of hand movements from local field potentials in monkey motor cortex," *Nature neuroscience*, vol. 6, pp. 1253–4, Dec. 2003. 33

[103] N. F. Ince, R. Gupta, S. Arica, A. H. Tewfik, J. Ashe, and G. Pellizzer, "High accuracy decoding of movement target direction in non-human primates based on common spatial patterns of local field potentials," *PloS one*, vol. 5, p. e14384, Jan. 2010. 33

[104] A. A. Kühn, D. Williams, A. Kupsch, P. Limousin, M. Hariz, G. H. Schneider, K. Yarrow, and P. Brown, "Event-related beta desynchronization in human subthalamic nucleus correlates with motor performance," *Brain*, vol. 127, pp. 735–46, Apr. 2004. 33

[105] G. C. Vanderheiden, "A journey through early augmentative communication and computer access," *Journal of rehabilitation research and development*, vol. 39, no. 6, pp. 39–53, 2002. 33

[106] C. Gerdtman and M. Lindén, "Six-button Click Interface for a Disabled User by an Adjustable Multi-level Sip-and-Puff Switch," in *SIGRAD 2010*, pp. 59–63, 2010. 33, 34

[107] A. Plotkin, L. Sela, A. Weissbrod, R. Kahana, L. Haviv, Y. Yeshurun, N. Soroker, and N. Sobel, "Sniffing enables communication and environmental control for the severely disabled (Supporting Information)," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 107, no. 32, pp. 1–9, 2010. 33, 41

[108] B. House, J. A. Bilmes, and J. Malkin, "The VoiceBot: a voice controlled robot arm," in *Proc. of the 27th int. conf. on Human factors in computing systems*, (Boston), pp. 183–192, ACM, 2009. 34, 36, 37

[109] A. Karpov, A. Ronzhin, A. Nechaev, and S. Chernakova, "Multimodal system for hands-free PC control," in *Proc. of 13th European Signal Processing Conf. EUSIPCO 2005*, vol. 7, (Antalya), Citeseer, 2005.

[110] M. C. Su and M. T. Chung, "Voice-controlled human-computer interface for the disabled," *Computing and Control Engineering Journal*, vol. 12, no. 5, pp. 225–230, 2001. 34, 36

[111] M. Benzeghiba, R. De Mori, O. Deroo, S. Dupont, T. Erbes, D. Jouvet, L. Fissore, P. Laface, a. Mertins, C. Ris, R. Rose, V. Tyagi, and C. Wellekens, "Automatic speech recognition and speech variability: A review," *Speech Communication*, vol. 49, pp. 763–786, Oct. 2007. 34, 35, 62

[112] M. A. Little, P. Mcsharry, I. Moroz, and S. Roberts, "Nonlinear, biophysically-informed speech pathology detection," in *Proc. of the 2006 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 14–19, IEEE, 2006. 35

[113] S. Furui, "50 years of progress in speech and speaker recognition," *Acoustic Society of America Journal*, vol. 116, no. 4, pp. 2497–2498, 2004. 35

[114] J. M. Baker, L. Deng, J. Glass, S. Khudanpur, C. H. Lee, N. Morgan, and D. O. Shaughnessy, "Research Developments and Directions in Speech Recognition and Understanding, Part 1," *IEEE Signal Processing Magazine*, no. May, pp. 75–80, 2009. 35

[115] N. Miller, "The neurological bases of apraxia of speech," *Seminars in speech and language*, vol. 23, pp. 223–30, Dec. 2002. 37

[116] T. Hueber, E. L. Benaroya, G. Chollet, B. Denby, G. Dreyfus, and M. Stone, "Development of a silent speech interface driven by ultrasound and optical images of the tongue and lips," *Speech Communication*, vol. 52, pp. 288–300, Apr. 2009. 37

[117] T. Hirahara, M. Otani, S. Shimizu, T. Toda, K. Nakamura, Y. Nakajima, and K. Shikano, "Silent-speech enhancement using body-conducted vocal-tract resonance signals," *Speech Communication*, vol. 52, no. 4, pp. 301–313, 2010. 37

[118] T. Simpson, C. Broughton, M. J. A. Gauthier, and A. Prochazka, "Tooth-click control of a hands-free computer interface," *IEEE transactions on bio-medical engineering*, vol. 55, pp. 2050–6, Aug. 2008. 37, 38, 41

[119] D. R. Corfield, K. Murphy, O. Josephs, G. R. Fink, A. Frakowiak, A. Guz, L. Adams, and R. Turner, "Cortical and subcortical control of tongue movement in humans: a functional neuroimaging study using fMRI," *Journal of Applied Physiology*, vol. 86, no. 5, pp. 1468–1477, 1999. 38

[120] O. Takai, S. Brown, and M. Liotti, "Representation of the speech effectors in the human motor cortex: somatotopy or overlap?," *Brain and language*, vol. 113, pp. 39–44, Apr. 2010. 38

[121] B. Haslinger, P. Erhard, F. Weilke, A. O. Ceballos-Baumann, P. Bartenstein, H. Gräfin von Einsiedel, M. Schwaiger, B. Conrad, and H. Boecker, "The role of lateral premotor-cerebellar-parietal circuits in motor sequence control: a parametric fMRI study.," *Brain research. Cognitive brain research*, vol. 13, pp. 159–68, Apr. 2002. 38

[122] Y. J. Kingma and P. J. Sabourin, "Development of a mouth-mouse for quadriplegics," in *Proc. of the 11th Annual Int. Conf. of the IEEE Engineering in Medicine and Biology Society*, no. 403, pp. 1516–1517, 1989. 38, 39

[123] D. Kim, M. Tyler, and D. J. Beebe, "Development of a Tongue-Operated Switch Array as an Alternative Input Device," *International Journal of Human-Computer Interaction*, vol. 18, pp. 19–38, Feb. 2005. 39, 41

[124] L. N. Struijk, "Tongue-computer interface for disabled people," in *Proc. of the 6th International Conf. on Disability, Virtual Reality and Associated Technologies*, pp. 153–158, 2006. 39

[125] Q. Peng and T. F. Budinger, "ZigBee-based Wireless Intra-oral Control System for Quadriplegic Patients," in *Proc. of the 29th Annual Int. Conf. of the IEEE Engineering in Medicine and Biology Society*, vol. 2007, pp. 1647–1650, IEEE, Aug. 2007. 39

[126] T. S. Saponas, D. Kelly, B. A. Parviz, and D. S. Tan, "Optically sensing tongue gestures for computer input," in *Proc. of the 22nd annual ACM symposium on User interface software and technology - UIST '09*, (Victoria), pp. 177–180, ACM Press, 2009. 38, 39

[127] L. N. S. A. Struijk, "An inductive tongue computer interface for control of computers and assistive devices," *IEEE Transactions on Biomedical Engineering*, vol. 53, pp. 2594–2597, Dec. 2006. 39

[128] H. A. Caltenco, E. R. Lontis, B. Bentsen, and L. N. S. Andreasen Struijk, "Effects of sensory feedback in intra-oral target selection tasks with the tongue," *Disability and rehabilitation. Assistive technology*, pp. 1–10, July 2012. 39

[129] C. Salem and S. Zhai, "An isometric tongue pointing device," in *Proc. of the SIGCHI conf. on Human factors in computing systems*, (New York, New York, USA), pp. 538–539, ACM Press, 1997. 40

[130] W. Nutt, C. Arlanch, S. Nigg, and G. Staufert, "Tongue-mouse for quadriplegics," *Journal of Micromechanics and Microengineering*, vol. 8, pp. 155–157, 1998. 40

[131] G. Krishnamurthy and M. Ghovanloo, "Tongue drive: a tongue operated magnetic sensor based wireless assistive technology for people with severe disabilities," in *Proc. of the IEEE International Symposium on Circuits and Systems*, p. 4, IEEE, 2006. 40

[132] X. Huo, J. Wang, and M. Ghovanloo, "Wireless control of powered wheelchairs with tongue motion using tongue drive assistive technology.," in *Proc. of the 30th Annual Int. Conf. of the IEEE Engineering in Medicine and Biology Society*, vol. 2008, (Vancouver), pp. 4199–4202, IEEE, Jan. 2008. 40

[133] X. Huo and M. Ghovanloo, "Evaluation of a wireless wearable tongue-computer interface by individuals with high-level spinal cord injuries.," *Journal of neural engineering*, vol. 7, p. 26008, Apr. 2010. 40, 41

[134] J. Hori, K. Sakano, M. Miyakawa, and Y. Saitoh, "Eye movement communication control system based on EOG and voluntary eye blink," in *Proc. of the 11th Int. Conf. on computers helping people with special needs*, pp. 950–953, Springer, 2006.

[135] S. Harada, J. A. Landay, J. Malkin, X. Li, and J. A. Bilmes, "The Vocal Joystick: Evaluation of Voice-based Cursor Control Techniques," in *In Proc. of the 8th Int. ACM SIGACCESS Conf. on computers and accessibility*, pp. 197–204, 2006.

[136] B. Denby, T. Schultz, K. Honda, T. Hueber, J. Gilbert, and J. Brumberg, "Silent speech interfaces," *Speech Communication*, vol. 52, pp. 270–287, Apr. 2010.

[137] E. McIntosh, A. Gray, and T. Aziz, "Estimating the costs of surgical innovations: the case for subthalamic nucleus stimulation in the treatment of advanced Parkinson's disease.," *Movement disorders: official journal of the Movement Disorder Society*, vol. 18, pp. 993–9, Sept. 2003.

[138] X. Huo, J. Wang, and M. Ghovanloo, "Introduction and preliminary evaluation of the Tongue Drive System: Wireless tongue-operated assistive technology for people with little or no upper-limb function," *The Journal of Rehabilitation Research and Development*, vol. 45, no. 6, pp. 921–930, 2008. 41

[139] M. V. Scanlon, "Acoustic monitoring of first responder's physiology for health and performance surveillance," *Proceedings of SPIE*, vol. 4708, pp. 342–353, 2002. 49, 51

[140] T. W. Beck, T. J. Housh, J. T. Cramer, J. P. Weir, G. O. Johnson, J. W. Coburn, M. H. Malek, and M. Mielke, "Mechanomyographic amplitude and frequency responses during dynamic muscle actions: a comprehensive review," *Biomedical Engineering Online*, vol. 4, p. 67, Jan. 2005. 50

[141] J. Silva, W. Heim, and T. Chau, "A self-contained, mechanomyography-driven externally powered prosthesis.," *Archives of physical medicine and rehabilitation*, vol. 86, pp. 2066–70, Oct. 2005. 50

[142] N. Alves and T. Chau, "The design and testing of a novel mechanomyogram-driven switch controlled by small eyebrow movements," *Journal of neuroengineering and rehabilitation*, vol. 7, p. 22, Jan. 2010. 50, 51

[143] J. D. Bass, M. V. Scanlon, and T. K. Mills, "Getting two birds with one phone: An acoustic sensor for both speech recognition and medical monitoring.," *The Journal of the Acoustical Society of America*, vol. 106, no. 4, pp. 2180–2185, 1999. 51

[144] P. Henry and T. R. Letowski, "Bone Conduction: Anatomy, Physiology and Communication," Tech. Rep. May, 2007. 51

[145] W. F. B. Brinkman, E. H. A. M. Marres, and J. Tolk, "The mechanism of bone conduction: An experimental study," *Acta Oto-laryngologica*, vol. 59, no. 2-6, pp. 109–115, 1965. 51

[146] P. Henry, P. Tran, and T. Letowski, "Comparison of Bone-Conduction Technologies," Tech. Rep. January, 2009. 51

[147] M. McBride, P. Tran, T. Letowski, and R. Patrick, "The effect of bone conduction microphone locations on speech intelligibility and sound quality," *Applied ergonomics*, vol. 42, pp. 495–502, Mar. 2011. 52

[148] M. Graciarena, H. Franco, K. Sonmez, and H. Bratt, "Combining standard and throat microphones for robust speech recognition," *IEEE Signal Processing Letters*, vol. 10, no. 3, pp. 72–74, 2003. 52

[149] S. N. Ghadiali, J. Banks, and J. D. Swarts, "Finite element analysis of active Eustachian tube function," *Journal of applied physiology*, vol. 97, pp. 648–54, Aug. 2004. 57

[150] L. R. Rabiner, "Algorithm for determining the endpoints of isolated utterances," *The Journal of the Acoustical Society of America*, vol. 56, no. 1, p. 31, 1974. 79, 81, 198

[151] P. Rajdev, M. P. Ward, J. Rickus, R. Worth, and P. P. Irazoqui, "Real-time seizure prediction from local field potentials using an adaptive Wiener algorithm," *Computers in Biology and Medicine*, vol. 40, no. 1, pp. 97–108, 2010. 79, 80, 110

[152] S. J. Orfanidis, *Optimum Signal Processing*. McGraw-Hill Publishing Company, 2007. 79, 80, 109, 110

[153] B. Wu, X. Ren, C. Liu, and Y. Zhang, "A Robust, Real-Time Voice Activity Detection Algorithm for Embedded Mobile Devices," *International Journal of Speech Technology*, vol. 8, pp. 133–146, 2005. 79

[154] X. Huang, A. Acero, and H. W. Hon, *Spoken Language Processing: A Guide to Theory, Algorithm and System Development*. Prentice Hall, 1st ed., 2001. 79

[155] K. Sakhnov, E. Verteletskaya, and B. Simak, "Approach for Energy-Based Voice Detector with Adaptive Scaling Factor," *International Journal of Computer Science*, vol. 36, no. 4, 2009. 79

[156] J. Ramírez, J. M. Górriz, and J. C. Segura, "Voice Activity Detection. Fundamentals and Speech Recognition System Robustness," in *Robust Speech Reconition and Understanding*, no. June, pp. 1–22, 2007. 79

[157] I. Osorio, M. G. Frei, and S. B. Wilkinson, "Real-Time Automated Detection and Quantitative Analysis of Seizures and Short-Term Prediction of Clinical Onset," *Epilepsia*, vol. 39, no. 6, pp. 615–627, 1998. 81

[158] R. Duda and P. Hart, *Pattern Classification*. Wiley, 2 ed., 2001. 97, 99, 125, 129, 130, 131, 135

[159] A. K. Jain, R. P. W. Duin, and J. Mao, "Statistical pattern recognition: a review," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 1, pp. 4–37, 2000. 98, 105, 107, 108, 114

[160] C. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, 1 ed., 2007. 98, 100, 103, 115, 129, 131, 135, 156, 202

[161] S. Datta, V. Pihur, and S. Datta, "An adaptive optimal ensemble classifier via bagging and rank aggregation with applications to high dimensional data.," *BMC bioinformatics*, vol. 11, p. 427, Jan. 2010. 98

[162] M. A. Little, *Biomechanically Informed Nonlinear Speech Signal Processing*. PhD thesis, University of Oxford, 2006. 99

[163] A. Jain and D. Zongker, "Feature Selection: Evaluation, Application, and Small Sample Performance," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 2, pp. 153–158, 1997. 99, 107, 119, 121, 126

[164] P. Somol, J. Novoviǒová, and P. Pudil, "Efficient Feature Subset Selection and Subset Size Optimization," *Pattern Recognition, Recent Advances*, pp. 75–98, 2010. 99, 121, 122, 126, 137

[165] J. P. Hoffbeck and D. A. Landgrebe, "Covariance Matrix Estimation and Classification with Limited Training Data," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, no. 7, pp. 763–767, 1996. 99, 131, 137

[166] B. Diri and S. Albayrak, "Visualization and analysis of classifiers performance in multi-class medical data," *Expert Systems with Applications*, vol. 34, pp. 628–634, Jan. 2008. 100

[167] L. I. Kuncheva, *Combining pattern classifiers, methods and algorithms*. Wiley-Interscience, 2004. 101, 155, 159, 160, 165, 171

[168] T. G. Dietterich, "Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms," *Neural Computation*, vol. 10, pp. 1895–1923, Oct. 1998. 103

[169] R. Kohavi, "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection," in *Proc. of the Int. Joint Conf. on Artificial Intelligence (IJCAI)*, (Montreal), pp. 1137–1143, Morgan Kaufmann, 1995. 103

[170] S. Mallet, *A wavelet tour of signal processing*. Academic Press, 2nd ed., 1998. 107, 110, 205

[171] K. Sayood, *Introduction to data compression*. San Diego: Morgan Kaufmann, 2nd ed., 2000. 107, 113

[172] A. Dufaux, *Detection and Recognition of Impulsive Sound Signals*. PhD thesis, University of Neuchatel, Switzerland, 2001. 108

[173] M. E. Davies and C. J. James, "Source separation using single channel ICA," *Signal Processing*, vol. 87, pp. 1819–1832, Aug. 2007. 108

[174] R. Vaidyanathan and C. J. James, "Independent component analysis for extraction of critical features from tongue movement ear pressure signals," in *Procs of the 29th Annual Int. Conf. of the IEEE Engineering in Medicine and Biology Society*, vol. 2007, pp. 5481–4, Jan. 2007. 108

[175] J. B. Tenenbaum, V. de Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction.," *Science (New York, N.Y.)*, vol. 290, pp. 2319–23, Dec. 2000. 108

[176] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding.," *Science (New York, N.Y.)*, vol. 290, pp. 2323–6, Dec. 2000.

[177] T. Balli and R. Palaniappan, "Classification of biological signals using linear and nonlinear features," *Physiological Measurement*, vol. 31, pp. 1–18, 2010. 108

[178] A. Schlögl, "A comparison of multivariate autoregressive estimators," *Signal Processing*, vol. 86, pp. 2426–2429, Sept. 2006. 110, 209

[179] J. H. McClellan, R. W. Schafer, and M. A. Yoder, *Signal Processing First*. Pearson Prentice-Hall, 2003. 110

[180] N. Ahmed, T. Natarajan, and K. R. Rao, "Discrete Cosine Transform," *IEEE Transactions on Computers*, no. January, pp. 90–93, 1974. 113, 153

[181] M. Turk and A. Pentland, "Eigenfaces for recognition," *Journal of cognitive science*, vol. 3, no. 1, pp. 71–86, 1991. 116

[182] L. Yu and H. Liu, "Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution," in *In Proc. of the 20th Int. Conf. on Machine Learning*, (Washington DC), pp. 856–863, 2003. 119, 120, 121

[183] Y. Saeys, I. Inza, and P. Larranaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics*, vol. 23, pp. 2507–17, Oct. 2007. 120

[184] Z. Zhao, S. Shashvata, A. Anand, F. Morstatter, S. Alelyani, and H. Liu, "Advancing Feature Selection Research - ASU feature Selection Repository," tech. rep., School of Computing, Informatics, and Decision Systems Engineering, Arizona State University, Tempe, 2010. 120, 124, 125

[185] H. Peng, F. Long, and C. Ding, "Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance and Min-Redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226–1238, 2005. 121

[186] S. Güneş, K. Polat, and Ş. Yosunkaya, "Multi-class f-score feature selection approach to classification of obstructive sleep apnea syndrome," *Expert Systems with Applications*, vol. 37, pp. 998–1004, Mar. 2010. 124, 125

[187] A. Golugula, G. Lee, and A. Madabhushi, "Evaluating feature selection strategies for high dimensional , small sample size datasets," in *In Proc. of the 33rd IEEE Int. Conf. of Engineering in Medicine and Biology Society*, pp. 949–952, 2011. 126

[188] F. Lotte, M. Congedo, A. Lécuyer, F. Lamarche, and B. Arnaldi, "A review of classification algorithms for EEG-based brain-computer interfaces.," *Journal of neural engineering*, vol. 4, pp. R1–R13, June 2007. 129

[189] H. Zhang, "The Optimality of Naive Bayes," in *Proc. of the 17th Int. FLAIRS conf. (FLAIRS 2004)*, (Florida), AAAI Press, 2004. 129, 130

[190] B. Zhang, Y. Zhang, and R. Begg, "Gait classification in children with cerebral palsy by Bayesian approach," *Pattern Recognition*, vol. 42, pp. 581–586, Apr. 2009. 130

[191] X. Wu, V. Kumar, R. Q. J., J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, P. S. Yu, Z. Zhou, M. Steinbach, D. J. Hand, and D. Steinberg, "Top 10 algorithms in data mining," *Knowledge and Information Systems*, vol. 14, pp. 1–37, Dec. 2007. 130

[192] P. M. Clarkson, *Optimal and adaptive signal processing*. Florida: CRC Press, 1993. 134

[193] J. Demsar, "Statistical Comparisons of Classifiers over Multiple Data Sets," *Journal of Machine Learning Research*, vol. 7, pp. 1–30, 2006. 138, 139, 140, 199

[194] T. Schmah, G. Yourganov, R. S. Zemel, G. E. Hinton, S. L. Small, and S. C. Strother, "Comparing classification methods for longitudinal fMRI studies.," *Neural computation*, vol. 22, pp. 2729–62, Nov. 2010. 139, 140

[195] R. Polikar, "Ensemble based systems in decision making," *IEEE Circuits And Systems*, vol. 6, no. 3, pp. 21–45, 2006. 154, 155, 156, 157, 159, 165, 171, 199

[196] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas, "On Combining Classifiers," *IEEE trans. on pattern analysis and machine intelligence*, vol. 20, pp. 226–239, Jan. 1998. 154, 155, 156, 157

[197] K. T. Leung and D. S. Parker, "Empirical comparisons of various voting methods in bagging," in *Proc. of the 9th ACM Int. Conf. on Knowledge discovery and data mining - SIGKDD '03*, (Washington DC), pp. 595–600, ACM Press, 2003. 155, 156, 157

[198] R. Caruana, A. Niculescu-Mizil, G. Crew, and A. Ksikes, "Ensemble selection from libraries of models," in *In Proc. of the 21st Int. Conf. on Machine learning - ICML '04*, no. 1996, (Banff), p. 18, ACM Press, 2004. 155

[199] N. Jankowski and K. Grabczewski, "Heterogenous committees with competence analysis," *Fifth International Conference on Hybrid Intelligent Systems (HIS'05)*, p. 6 pp., 2005. 155

[200] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of Computer and System Sciences*, vol. 55, no. 1, pp. 119–139, 1997. 160

[201] J. F. Borisoff, S. G. Mason, and G. E. Birch, "Brain Interface Research for Asynchronous Control Applications," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 14, pp. 160–164, June 2006. 178

[202] B. Waryck, "Comparison of two myoelectri multi-articulating prosthetic hands," in *Proc. of the 2011 Myoelectric Controls/Powered Prosthetics Symposium*, (New Brunswick), pp. 1–4, 2011. 202

[203] C. M. Light, P. H. Chappell, and P. J. Kyberd, "Establishing a standardized clinical assessment tool of pathologic and prosthetic hand function: Normative data, reliability, and validity," *Archives of Physical Medicine and Rehabilitation*, vol. 83, pp. 776–783, June 2002. 203

[204] A. Walden, "Wavelet analysis of discrete time series," in *Proc. of the 3rd European congress of mathematics*, (Barcelona), pp. 627–641, 2001. 205, 206