Eye-Speech Affect Detection for Automatic Speech Recognition

By

ASHWAQ H. ALHARGAN

A Thesis Submitted for the Degree of

Doctor of Philosophy



Department of Electronic, Electrical and Systems Engineering

College of Engineering and Physical Sciences

University of Birmingham

November 2018

UNIVERSITY OF
BIRMINGHAM

University of Birmingham Research Archive

e-theses repository

This unpublished thesis/dissertation is copyright of the author and/or third parties. The intellectual property rights of the author or third parties in respect of this work are as defined by The Copyright Designs and Patents Act 1988 or as modified by any successor legislation.

Any use made of information contained in this thesis/dissertation must be in accordance with that legislation and must be properly acknowledged. Further distribution or reproduction in any format is prohibited without the permission of the copyright holder.

I would like to dedicate this thesis to my lovely kids Ziad and Ahmed – the reason for keeping me going . . .

Declaration

I hereby declare that the content of this thesis is my own work undertaken for the degree of Doctor of Philosophy and has not been submitted in whole or in part for consideration for any other degree or qualification at this or any other university. I confirm that:

- This thesis includes nothing which is the outcome of work done in collaboration, except where specifically indicated in the text.
- Some of the work in this thesis has been published in peer-reviewed conference proceedings.

Ashwaq Alhargan

Abstract

Human-computer interaction (HCI) is becoming increasingly natural. Machines are now able to recognise faces, to understand individual speech and to converse like a human would. However, they are still far from exhibiting humanlike intelligence. Affects play an important role in interaction, so understanding and responding to them are necessary steps towards more natural HCI.

This thesis reports the development and evaluation of affect detection systems suitable for use in real-life HCI applications (e.g. speech-enabled interfaces such as Alexa) using speech and eye movement modalities. A corpus of spontaneous affective responses in these modalities within an interactive virtual gaming environment, designed to elicit different affective states corresponding to the arousal and valence dimensions, was collected. A support vector machine was employed as a classifier to detect the affects elicited from both modalities. Several features of eye movement, namely pupillary response, fixation, saccade and blinking, are assessed for use in affect detection and new pupil response features based on the Hilbert transform are proposed. Acoustic and lexical characteristics of speech are investigated. The detection results suggest that eye movement is superior to speech, with pupillary response features based on Hilbert transform yielding superior performance on the arousal dimension, whereas saccade and fixation features perform better on the valence dimension. The improvement made by combining information from eye movement and speech modalities suggests that the two modalities carry complementary information for affect detection and that both warrant incorporation where feasible. An ASR application integrating affective information from both modalities for affect robustness was investigated. The best performing system uses affective information from eye movements, significantly reducing word error rates compared to the speech modality alone. This work highlights the potential of eye movements as an additional modality

to speech to enhance the accuracy of affect detection and facilitate the development of robust affect-aware speech-enabled interfaces.

Acknowledgements

First and foremost, I would like to express my deep sense of gratitude and sincere thanks to my supervisors: Dr Neil Cooke for his guidance, support and constructive criticisms throughout my PhD research and Professor Chris Baber for his advice and valuable comments. Thank you to my assessor, Dr Peter Jancovic, for his constructive feedback on my yearly progress reports. I am extremely thankful for their efforts, which helped to shape my ideas and taught me how to become an academic researcher. Thanks to Professor Robert stone and Mohammadhossein of the HIT Lab for providing me with the simulation environment used in this research, especially Mohammadhossein for his time in answering all my questions and emails. Many thanks to Jack, the welfare tutor, for his massive help and support. I would also like to take this opportunity to express my sincere thanks to the University of Birmingham and the EESE department for offering a studious environment and great facilities. Thanks to my colleagues at HCI lab for their informative discussions and help. I also would like to thank Nigel Hubbell for copy editing this thesis for conventions of language, spelling and grammar. Last but not least, I would like to express my deepest gratitude and appreciation to my parents for their emotional support, encouragement and care. Nothing would have been done without them. Additionally, I would like to thank my boys Ziad and Ahmed for their patience and tolerance throughout this journey. Finally, I express my sense of appreciation to all who have helped supported me throughout my PhD journey.

Table of Contents

1 Introduction 1
1.1 Overview 1
1.1.1 Eye tracking technology 4
1.1.2 Automatic Speech Recognition (ASR) 5
1.2 Research questions and methodology 6
1.3 List of publications 7
1.4 Thesis outline 8

2 Affect, Speech and Eye Movements 12
2.1 Affects 12
2.1.1 Models of affect 13
2.2 On affect detection 15
2.2.1 Affective data acquisition methods 16
2.2.2 Affect detection using the speech modality 19
2.2.3 Affect detection using the eye movement modality 29
2.3 Affects in automatic speech recognition 33
2.4 Summary 37

3 Affective Eye Movements and Speech Corpus 38
3.1 Motivation for developing the aGame Corpus 38
3.2 Description of the affective virtual gaming environment 40
3.3 Method of developing the aGame corpus 43
3.3.1 Participants 44
3.3.2 Apparatus 44
3.3.3 Recordings and software 45
3.3.4 Experimental task and procedure for corpus collection 46
3.4 Post-processing of recordings 49
3.4.1 Data synchronization 49
3.4.2 Transcription of speech recordings 51
3.4.3 Quality monitoring and assessment 51
3.5 Corpus size 54
3.6 Summary 55

4 Affect Elicitation Analysis and Detection from Eye Movements 56
4.1 Motivation 56
4.2 Affective Experiences Analysis 57
4.3 Detection method 59
4.3.1 Support vector machine (SVM) 60

4.3.2 Feature selection 64
4.3.3 Feature normalisation 65
4.3.4 Validation methods 66
4.3.5 Performance evaluation metrics 67
4.4 Development of an affect detection system based on eye movements 69
4.4.1 Preprocessing of pupillary response signal 69
4.4.2 Eye movements features extraction 73
4.4.3 Analysis of affect elicitation 76
4.4.4 Experimental settings 79
4.4.5 Results 79
4.5 Comparison with the state of the art 83
4.6 Summary 85
5 Affect Detection from Speech and Integration of Modalities 87
5.1 Motivation 87
5.2 Development of an affect detection system based on acoustic speech signals 88
5.2.1 Acoustic speech signal processing 89
5.2.2 Extraction of acoustic features 91
5.2.3 Affect elicitation analysis 97
5.2.4 Experiment settings 99
5.2.5 Results 100
5.3 Development of an affect detection system based on the lexical content of speech 103
5.3.1 Affective lexical dictionary 103
5.3.2 Lexical feature representation 106
5.3.3 Experimental settings 107
5.3.4 Results 108
5.4 Multimodal-based Affect detection system 110
5.4.1 Modality integration methodology 110
5.4.2 Results 112
5.5 Summary 115
6 Multimodal Affect-Robust ASR System 118
6.1 ASR system overview 119
6.1.1 Acoustic modelling 120
6.1.2 Language modelling 120
6.1.3 Pronunciation dictionary 121
6.1.4 Feature extraction 122
6.1.5 Decoding 122
6.2 ASR model adaptation techniques 123
6.2.1 Acoustic model adaptation techniques 123
6.2.2 Context-based language modelling adaptation 126
6.3 Integration of affects into ASR system architecture 126
6.3.1 Affect-adaptive language modelling 128
6.4 Evaluation method 130
6.4.1 Performance evaluation metrics 130
6.4.2 Training corpora 131
6.4.3 LibriSpeech corpus 132

6.4.4 Baseline ASR system 132
6.4.5 Baseline language model 133
6.4.6 Lattice rescoring 134
6.5 Experiments 134
6.6 Results and discussion 136
6.6.1 Acoustic model adaptation - Experiment 1 137
6.6.2 ALM performance – Experiment 2 139
6.6.3 Evaluation of ALM in ASR system – Experiment 3 141
6.6.4 Performance of affect-robust ASR-based ADS system – Experiment 4 142
6.7 Summary 146
7 Conclusion 148
7.1 Contributions 149
7.1.1 The aGame corpus 149
7.1.2 Eye movements-based affect detection 150
7.1.3 Speech-based affect detection and integration with eye movements 150
7.1.4 A multimodal affect-robust ASR system 151
7.2 Recommendations for Future Work 152
7.3 Summary 153
Appendix A 154
Appendix B 157
Appendix C 158
Appendix D 159
Appendix E 161
Appendix F 162
References 165

List of Figures

Figure 1.1: Block diagram of thesis structure. 9
Figure 2.1: Russell's Circumplex model of affect (simplified version), with eight affective keywords distributed in the two-dimensional space [5]. 15
Figure 3.1: A snapshot of the affective virtual gaming environment, a neutral scenario, adapted from [10]. 41
Figure 3.2: Experimental setup for eye movement and speech corpus collection. 43
Figure 3.3: A screenshot of the playback video of a neutral game session (Player 2) recorded using the EyeLink Screen Recorder software. The red and blue circles represent the position of the player's eyes on the screen (fixation event). The numbers show the timestamp (system clock) for host and display PCs. 46
Figure 3.4: Experimental procedure for collecting the aGame corpus. 47
Figure 3.5: Main experimental setup. 49
Figure 3.6: Scheme for synchronising recordings between PCs by sending keypress massages from the display PC to the host PC. 50
Figure 3.7: Example of time-aligned transcription (buttom) of a speech signal (top) using Praat software. The corresponding spectrum and acoustic featuers are shown in the middel. 51
Figure 3.8: Plot of a player's X and Y fixation location during a game session over all captured video frames. The player's focus was mostly on the center of the screen and the surrounding area. 53
Figure 3.9: Example of a player's speech in the aGame corpus. 54
Figure 4.1: An overview of the framework for the affect detection process used throughout this study. 59
Figure 4.2: Two-dimensional binary classes, linearly separable; (a) shows possible separating hyperplanes and (b) shows the optimal separating hyperplane. Red circles indicate support vectors for the two classes. 60
Figure 4.3: An illustration of a confusion matrix. 68
Figure 4.4: Overview of the front-end processing of eye movement signals. 69
Figure 4.5: Overview of the procedure for preprocessing pupillary response signals. Screen illumination is measured continuously during the gameplay session using a luxmeter to compensate for PLR. 70
Figure 4.6: Pupillary response unit conversion using an artificial pupil. 71
Figure 4.7: An illustration of pupillary response blink recovery. Missing samples due to blinking (blue) were recovered successfully using cubic interpolation. 72

Figure 4.8: An illustration of the HT unwrapped instantaneous phase of pupillary responses during different affective games for player 15. N: neutral, NV: negative valence, LA:low arousal, PV: positive valence and HA: high arousal. 76
Figure 4.9: Box plots of different eye movement features of the three affective states on the arousal and valence dimensions. One-way repeated measures ANOVA results revealed significant differences between pupillary responses and fixation duration means on both dimensions ($p < 0.05$). Saccade amplitude and blinking rate means were significantly different on valence and arousal dimensions, respectively. 78
Figure 4.10: Bar graph of detection accuracies on the arousal (top) and valence (bottom) dimensions using different eye movement feature sets, with 95% confidence interval (n=14). Straight line (red) shows random accuracy. 81
Figure 4.11: Detection accuracies per affective state on the arousal (left) and valence (right) dimensions using different eye movement feature sets. 82
Figure 5.1: Overview of the front-end processing of acoustic speech signals. 89
Figure 5.2: An example of detected speech segments/voiced units (red). 90
Figure 5.3: A block diagram of the MFCC feature extraction process. 94
Figure 5.4: An illustration of the mel scale filter bank, with 26 triangular filters, for 8 kHz bandwidth. At the centre frequencies the filters' response is 1, then it decreases linearly towards 0. The filters have a response of 0 at the centre frequencies of adjacent filters, where they overlap. 96
Figure 5.5: Box plots showing different acoustic features on different levels of the arousal and valence dimensions. One-way repeated measures ANOVA results showed signfiicant differences between featuers' means across the affective states of both dimensions ($p < 0.05$). 99
Figure 5.6: Bar graph of detection accuracies on the arousal (top) and valence (bottom) dimensions using different acoustic feature sets with 95% confidence intervals (n=14). Better accuracy indicates greater importance of the feature set for the affect detection task. Straight line (red) shows random accuracy. 101
Figure 5.7: Detection accuracies per affective state on the arousal (left) and valence (right) dimensions using different acoustic feature sets. 103
Figure 5.8: An example of Game-oriented words found in the aGame corpus. 104
Figure 5.9: Bar graph of detection accuracies on the arousal and valence dimensions using different lexical features, with 95% confidence intervals (n=14). All features provided better accuracy than the random level of 33.33% (straight line in red). 109
Figure 5.10: Overview of the modality fusion framework using both feature-level (FLF) and decision-level fusion (DLF). 111
Figure 5.11: A bar plot comparing the best unimodal systems performances with multimodal-based ADS uaing DLF approach, on the arousal and valence dimensions, with 95% confidence intervals (n=14). 113
Figure 5.12: Confusion matrices for the best unimodal ADS and different modality integration approaches on the arousal and valence dimensions. 115
Figure 6.1: Basic framework of an automatic speech recognition system. 120

Figure 6.2: Integration of affective information into the speech recognition process. The affect detection system ADS serve as a front-end process for the ASR system. 127
Figure 6.3: Lattice rescoring framework. 134
Figure 6.4: Percentage of correct affective words recognised before and after lattice rescoring with ALMs. Rescoring with ALMs increase the prediction of affective words. 142
Figure 6.5: WER performance of affect-robust ASR system using affective information from different modalities on arousal and valence dimensions. Incorporating information from eye movement as additional modality to speech reduced the WER on arousal dimension significantly $(p < 0.05)$..... 144
Figure 6.6: Performance of affect-robust ASR system on the arousal (left) and valence (right) dimensions for each player in the aGame corpus. 145
Figure C.1: A screenshot of the questionnaire, from the interactive virtual gaming environment developed by [10]. 158
Figure D.1: Eyelink II eye movement tracking sensor, adapted from [27]. 159
Figure D.2: Shure WH20 microphone. 159
Figure D.3: Edirol UA-25 speech recorder. 160
Figure D.4: Logitech Wingman 3D force feedback joystick. 160
Figure D.5: Extech HD45 Luxmeter. 160
Figure F.1: A left-to-right five-state HMM, adapted from [202]. 163

List of Tables

Table 2.1: Review of some speech-based affect detection studies. 24
Table 2.2: A summary of affect detection literature using eye movement modality. Dep: subject dependent evaluation. Indep: subject independent evaluation. 34
Table 3.1: Five affective games with their parameter settings and ground truth lables, according to Russell's circumplex model of affect, selected from [10]. 42
Table 3.2: Details of sessions recorded per player in the aGame corpus. 54
Table 4.1: Signficant correlation between players' self-reporting and the expected affective experiences (the ground truth labels as defined in [10]). 58
Table 4.2: Features extracted from eye movement signals. 76
Table 4.3: Affect detection results for different feature sets of eye movement signals. PRHT performed best on arousal dimension and SAC+FIX performed best on valence dimension. 80
Table 4.4: A list of the most important features selected more frequently across the players. 83
Table 4.5: State-of-the-art comparison of different affect detection studies in gameplay scenarios using physiological signals. Eye movement signals seem promising for affect detection. 85
Table 5.1: Features extracted from speech signals. 97
Table 5.2: List of the most relevent MFCCs ($p < 0.05$) on both the arousal and valence dimensions. 98
Table 5.3: Affect detection results for different acoustic feature sets on the arousal and valence dimensions. Combined acoustic features performed best on arousal dimension, whereas ZCR performed best on valence dimension. 101
Table 5.4: List of words in the aGame corpus that maxiamully associated with affective states on both the arousal and valence dimention. 106
Table 5.5: Affect detection results using different lexical features. PPMI feature performed best on both dimensions. 108
Table 5.6: Affect detection performance using both eye movement and speech as input modalities, based on feature and decision level fusion approaches, on the arousal and valence dimensions. *Significant improvement (p <0.05) over the best unimodal ADSs. 112
Table 6.1: Word distribution and vocabulary size of the aGame corpus on arousal and valence dimensions. 131
Table 6.2: Baseline ASR performance on the LibriSpeech test sets. 133
Table 6.3: Acoustic model adaptation performance on aGame corpus speech data. 137
Table 6.4: Acoustic model adaptation performance, showing the influence of affective speech on the arousal dimension. 138

Table 6.5: Acoustic model adaptation performance, showing the influence of affective speech on the valence dimension. 138
Table 6.6: Perplexity scores for the adapted ALMs, averaged over players. * Significant improvement ($p < 0.05$) from baseline language model. 140
Table 6.7: Cross-affect testing of affect-adaptive language models on arousal and valence dimensions. 140
Table 6.8: ASR system performance before and after lattice rescoring with ALMs. *Significant improvement ($p < 0.05$) from the baseline. 141
Table 6.9: Affect-robust ASR system WER performance on aGame corpus using different affect detection systems, averaged over players. 143
Table 6.10: Affect-robust ASR system WER before and after ALM integration based on ADS output. 146
Table A.1: A list of available affective corpora. 156

List of Abbreviations

ACF	Autocorrelation function
ADS	Affect Detection System
ALM	Affective Language model
ANN	Artificial Neural Network
ANOVA	Analysis of Variance
ANS	Autonomic Nervous System
ASR	Automatic Speech Recognition
BOW	Bag-of-words
CMS	Cepstral Mean Subtraction
CMU	Carnegie Mellon University
CR	Corneal Reflection
DFT	Discrete Fourier Transform
DLF	Decision Level Fusion
DNN	Deep Neural Network
ECG	Electrocardiography
EEG	Electroencephalography
FDR	Fisher Discriminant Ratio
FIX	Fixation
FLF	Feature Level Fusion
GMM	Gaussian Mixture Model
GSR	Galvanic Skin Response
HA	High Arousal
HCI	Human-Computer Interaction
HMM	Hidden Markov Model
HNR	Harmonics to Noise Ratio
HT	Hilbert Transform
KNN	K-nearest neighbours
LA	Low Arousal
LLD	Low Level Descriptors
LMSF	Language Model Scaling Factor
MAP	Maximum A Posteriori
MFCC	Mel-Frequency Cepstral Coefficients
MLLR	Maximum Likelihood Linear Regression
N	Neutral
NV	Negative Valence
PLR	Pupillary Light Reflex
PMI	Pairwise Mutual Information
POS	Part-of-speech
PR	Pupillary Response

PRHT	Pupillary Response with the Hilbert Transform Features
PSD	Power Spectral Density
PV	Positive Valence
RBF	Radial Basis Function
SAC	Saccade
SAM	Self-Assessment Manikin
SD	Standard Deviation
SVM	Support Vector Machine
TF	Term Frequency
TF.IDF	Term Frequency-Inverse Document Frequency
VAM	"Vera am Mittag" corpus
VSM	Vector Space Modelling
WER	Word Error Rate
WIP	Word Insertion Penalty
WOz	Wizard-of-Oz
ZCR	Zero-crossing rate
f_0	Fundamental Frequency
2D	Two Dimensional
3D	Three Dimensional

Chapter 1

Introduction

'*When dealing with people, remember you are not dealing with creatures of logic, but with creatures of emotion.*' - *Dale Carnegie*

1.1 Overview

Affects are complex experiences (e.g. happiness, sadness and anger) that individuals exhibit in their everyday interactions. They play an important role in communication; in particular, they modulate individuals' interactions and can resolve ambiguities in messages. Affects are typically conveyed in different ways such as by changes in physiological responses (e.g. heartrate, sweating and eye movements) or in behaviour (e.g. facial expressions, body posture and speech). Individuals frequently express affects while interacting with machines, as they would with their peers [1]. Therefore, it is essential for machines to understand their users' affects and to respond accordingly.

Although computing machines have become increasingly intelligent and now have the ability to identify faces (e.g. Apple Face ID technology), to understand individual speech and to start a conversion like a human would (e.g. Amazon Alexa and Google Home, speech-enabled interfaces) they are still far from exhibiting humanlike intelligence. It is argued that affects are an important aspect of human-computer interaction (HCI) and must be considered in order to improve the quality of the interactions between individuals and machines [1]. The concept of affective computing, which emerged in 1995, aims to enable machines to detect and

respond to human affects and to make these interactions more intelligent, efficient and natural [2]. A great deal of effort has been put into developing affect detection systems (ADSs) in recent decades. Researchers have utilised various sensory information modalities such as speech, facial expression, body posture and physiological responses [3, 4] as input to machine learning algorithms in order to detect the affective states of individuals. An affect representation model that has been widely adopted in affect detection is Russell's circumplex model of affect [5]. This model represents affects in two dimensions, called arousal and valence. The arousal dimension ranges from excitement to sleepiness, while valence ranges from pleasure to displeasure.

Despite remarkable progress in the field, affect-intelligent machines are not a consumer reality yet [6]. Obtaining appropriate and realistic affective data relevant to the affects that individuals experience in real-life HCI is considered a fundamental requirement for the applicability of affect-aware interfaces in the real world [6]. Most research depends heavily on three types of data to develop affect detection systems. These are acted data (e.g. [7]), spontaneous/real-life scenarios (e.g. call centres [8]) and elicited data (e.g. [9]). Acted data is meaningless for real-life HCI applications, as it does not realistically reflect individuals' unprompted affective responses. As to spontaneous data, this requires more effort to collect, and affects are typically defined as perceived by others rather than as felt by individuals, due to the absence of any ground-truth reference. In addition, such data is applicable only to behavioural signals as they are visually/auditory observable. On the other hand, it is believed that elicited data can represent the spontaneous affective responses felt by individuals when exposed to stimuli designed to elicit specific affects, which they report subjectively. Hence, it is feasible to establish a ground-truth reference to the individuals' felt affects. Nevertheless, the majority of affect elicitation methods so far employed (e.g. movies [9]) are considered passive (noninteractive) and limited in their ability to simulate real-life HCI experiences. Recently, however, the use of virtual reality environments (e.g. gaming [10]) has emerged as an active affect elicitation method. Thanks to their capacity for strengthening the sense of presence in the real world, it is believed that virtual reality environments can facilitate the capture of more realistic affective experiences, which can be generalised to real-life HCI application (e.g. speech-enabled interfaces, interactive games, etc.).

Since the speech modality is considered an important aspect of a natural HCI, being the primary mode of interaction, as well as the most natural and favourable, it follows that detecting

affect from speech is desirable [11]. The issue arises, however, of the reliability of using this modality to detect affects, especially in spontaneous cases [12]. As affects are manifested multimodally, considering information from other modalities can improve the robustness of the affect detection process [4]. The majority of multimodal contributions have focused on the so-called audio-visual modality, combining information from facial expressions and body posture with speech, whereas incorporating information from physiological responses has been largely ignored especially from eye movement [4, 6].

Facial expressions and body posture are behavioural signals, which, in common with speech, individuals can deliberately control and manipulate, making the felt affects difficult to measure [13]. By contrast, physiological signals originate in the central/autonomic nervous systems, making them very difficult or impossible to control deliberately [3]. It is thus believed that incorporating information from physiological modalities with speech is beneficial and can provide a richer and more reliable representation of affects. A major difficulty is that most methods of measuring physiological responses, e.g. electrocardiogram (ECG)[1], galvanic skin response (GSR)[2], or electroencephalography (EEG),[3] are intrusive compared to speech. Lately, eye movement tracking has emerged as a means of measuring affect (e.g. [9, 14]), because the signals are natural and their observation is more efficient and less intrusive than for other physiological signals [15]. Hence, eye movement is a potential candidate signals for representing affective responses via physiological behaviour to enhance the affect detection process and facilitate natural interaction.

As a contribution to making affect-aware interfaces feasible in a natural HCI, this thesis focuses on the detection of affects based on spontaneous affective eye movements and speech data, elicited in an interactive virtual environment, namely a gameplay scenario. The research has many potential applications in HCI field. For example, researchers from Amazon Inc. are working on their speech-enabled interface (Alexa, a virtual assistants) to be able to understand and interact accordingly to the individual's affective states. Another application example is gaming, to enable adaptive gameplay experiences [16] e.g. Brainball [17] is an interactive game in which a physical ball moves according to the player relaxation state, and Microsoft Xbox games. E-learning environment can also benefit from this research, to support adaptive teaching (e.g. change the content style based on the individual's affective state) [18]. The following

[1] An ECG measures heart rhythm, rate and electrical activity using electrodes attached to the individual's chest.
[2] GSR or skin conductivity is a measure of the electrical conductivity of the skin.
[3] An EEG records the electrical activity of the brain through the scalp.

subsections provide an overview on eye movement tracking technology and discuss automatic speech recognition (ASR) systems highlighting the impact of affective speech on ASR system performance, an application of an affect-aware interface for natural HCI.

1.1.1 Eye movement tracking technology

Eye movement tracking is a sensing technology that facilitates machine awareness of an individual's eye movement behaviour, such as focus of attention and scan path, in a given environment. Eye movement tracking sensors commonly measure fixation, saccade, pupillary responses and blinking events [19]. Fixation occurs when a subject is focusing on a static target on the screen. Saccade is defined as fast eye movement and occurs between fixations, when the individual switches his attention from one target to another in the environment. Pupillary responses and blinking are physiological phenomena, the former manifested as dilation and contraction of the pupil and the latter as brief closure of the eyelid.

Eye movement tracking was initially used in physiological and psychological studies in order to understand human cognition and scene perception processes [20]. Recently, it has been applied in other fields, such as marketing research [21]. In the context of HCI, there has been an increasing amount of research into the use of eye movement tracking technology for evaluating the usability of interfaces [22], understanding user attention and actions in context-sensitive situations [23], remote collaborative interfaces [24] and replacing traditional input devices to facilitate HCI tasks for disabled people [25].

There are in general two types of eye movement tracking sensors: wearable and remote. Wearable sensors require physical contact with the subject and can be categorised into video-based and non-video-based trackers [26]. Video-based trackers can be worn as glasses (e.g. Tobii glasses), or head mounted (e.g. EyeLink II). They typically use two high speed cameras with infrared illumination to capture video of the eyes [27]. Eye movements are tracked by directing infrared illumination towards the pupils, detecting the reflections of this light from the corneas and decoding them using image processing software [19]. The non-video-based eye movement tracking measures skin potentials thought electrodes attached to the skin around the eyes [26], so called Electrooculogram (EOG) sensor, the measured changes in the skin can be analysed to detect blinks. EOG is less reliable in term of estimating other eye movements

measures (e.g. gaze location/fixation) compared video-based eye movement tracking and it is mainly used for clinical applications [26].

Remote eye movement tracking sensors, on the other hand, are based on multiple cameras located in the environment which fixate on the subject's face to capture movements of the eyes and head and are also video-based. Some remote eye trackers are built into the display screen (e.g. Acer Predator[4]). Remote trackers are less invasive than wearable ones and easier to set up. Wearing an eye tracker may also feel unnatural and can cause discomfort, unlike remote trackers. However, wearable head-mounted eye movement tracking sensors offer superior tracking quality [27] and were therefore used in this research to ensure accurate data collection.

Eye movement tracking is now an established technology; in recent years, it has been integrated as interaction modality in smartphones (e.g. Samsung Galaxy), virtual reality headsets (e.g. Fove[5]) and gaming (e.g. Deus Ex[6]). Indeed, eye movement tracking is very likely to become a standard interaction modality in many end-user products and applications. For example, Nacke and Mandryk [28] found that interaction through eye movement is more preferable than other physiological input (e.g. GSR) in a simple-side scrolling game. Therefore, research into effect detection via eye movements is valuable for the HCI community.

1.1.2 Automatic Speech Recognition (ASR)

Speech-enabled interface is a human-computer interface that uses speech as input modality to understand and communicate naturally with individuals. It employs an ASR system, also known as speech-to-text, to interpret human speech sounds. Advances in statistical modelling and machine learning techniques since the 1980s have led to ASR systems becoming usable in various practical applications [29]. These systems typically involve modelling the time-varying acoustic features extracted sequentially from the speech waveform using hidden Markov models (HMMs) to estimate the most likely word sequence corresponding to a given speech signal [30]. A language model is used to estimate the probability of the word sequence. ASR represents a major breakthrough in technology and has changed the way that individuals interact

[4] https://www.acer.com/
[5] https://www.getfove.com/
[6] Using Tobii's game-enhancing eye movement tracking features https://tobiigaming.com/

with machines, from automated call centre systems (e.g. "*How may I help you?*" [31]) to virtual assistants.

Despite the remarkable success of ASR systems in various applications, their performance remains an open issue in term of recognising spontaneous speech in natural environment settings due to a mismatch between system training and recognition conditions. Among the variety of challenging factors that can affect ASR system performance accuracy are environmental noise, speaker differences (e.g. age, and gender) and affects [32]. Typically, the performance of ASR systems tends to deteriorate during affective speech, due to changings in individuals vocal behaviour (e.g. fundamental frequency) [33]. Since speech is essential for a natural HCI, ASR systems should consider the affective state of the speaker in order recognise the verbal content more accurately. This can help to improve the interaction with the user and make it more natural. Little research has addressed the problem of affective speech in ASR systems, however. Proposed approaches have mainly focused on adapting the system based on the speaker's affective state as detected from the acoustic aspects of speech signals [33-35]. The issue is that affects may not be detected from the spontaneous speech with high accuracy, as it can be deliberately controlled and manipulated by individuals.

Recently, researchers have demonstrated that combining speech sounds with other modalities (e.g. lip movements) can improve the performance of ASR systems [36]. In addition, the use of the eye movement modality in ASR has been explored in the literature with promising results, suggesting that integrating information from eye movements (e.g. the visual focus of attention) can aid the speech recognition process in speech-enabled interfaces and make it more efficient in term of detecting what is being said related to the visual field (e.g. [37, 38]). Given the difficulties in detecting affects from spontaneous speech data, this thesis investigates the use of eye movements as an additional modality to improve the recognition of affective speech in ASR systems. Hence, facilitating affect-aware speech-enabled interfaces for natural HCI.

1.2 Research questions and methodology

The ultimate aim of this study is to investigate the use of eye movement and speech modalities for automatic detection of individuals' affective experiences, elicited in a natural setting of a

virtual interactive environment, and its application to ASR system for affect robustness. The following research questions are investigated:

1. Can information from eye movements be used to detect affective states elicited in a virtual interactive environment?
2. How can spontaneous speech be elicited in a virtual interactive environment and to what degree can the experienced affective state be detected?
3. How can affective information from eye movements and speech be combined effectively?
4. How can an ASR system benefit from the use of affective information to improve its performance?

To answer the research questions, the following methodology is adopted:

- Collecting a corpus of eye movements and speech during interaction with an affective virtual environment.
- Investigating eye movement features associated with affect elicited in the virtual interactive environment.
- Investigating speech features associated with affect elicited in the virtual interactive environment.
- Developing an affect detection system to infer the elicited affective state from eye movement and speech signals, unimodally and multimodally.
- Developing a framework to account for affects in ASR systems and recognising affective speech.
- Evaluation of the affect-robust ASR systems and their benefit in terms of improving performance during affective speech.

1.3 List of publications

Some of the work reported in this thesis has been published in peer-reviewed conference reports as follows:

- A. Alhargan, N. Cooke, and T. Binjammaz. "Eye tracking-based affect detection using Hilbert transform features". 19th European Conference on Eye Movements. Accepted. (Chapter 4)
- A. Alhargan, N. Cooke, and T. Binjammaz. "Affect recognition in an interactive gaming environment using eye tracking". Seventh International Conference on Affective Computing and Intelligent Interaction (ACII 2017).[7] (Chapter 3 and part of Chapter 4)
- A. Alhargan, N. Cooke, and T. Binjammaz. "Multimodal affect recognition in an interactive gaming environment using eye tracking and speech signals". In Proceedings of the 19th ACM International Conference on Multimodal Interaction (ICMI 2017).[8] (Chapter 5 and part of Chapter 4)
- A. Alhargan and N. Cooke. "EyeS: Towards an Affect-robust Automatic Speech Recognition (ASR) system". (In preparation – Chapter 6)

1.4 Thesis outline

The structure of the thesis is illustrated in Figure 1.1. Following this introductory chapter, the remainder of the thesis consists of six chapters. The main research contributions are presented in Chapters 3, 4, 5 and 6, then Chapter 7 concludes the thesis. This chapter has offered an overview of the topic, declared the aims of the study, research questions and listed publications resulting from the work. An outline of the remaining chapters follows.

- **Chapter 2** describes affects and models of affects, then reviews the literature on affect detection from speech and eye movements. Finally, affects issue in ASR systems is discussed.
- **Chapter 3** describes the motivation and the experimental method used for the collection of eye movement and speech data in the virtual interactive environment, a gameplay scenario. In addition, it discusses the post-processing of the recordings, including data synchronisation and quality assessment.

[7] [39] A. Alhargan, N. Cooke, and T. Binjammaz, "Affect recognition in an interactive gaming environment using eye tracking," in *17th International Conference on Affective Computing and Intelligent Interaction (ACII)*, 2017.

[8] [40] A. Alhargan, N. Cooke, and T. Binjammaz, "Multimodal affect recognition in an interactive gaming environment using eye tracking and speech signals," in *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, 2017: ACM Press, pp. 479-486.

- **Chapter 4** analyses the players' reported affective experience elicited while playing in the virtual interactive gaming environment and describes the detection methods used throughout the research. It discusses the development of the affect recognition system using eye movement signals, namely pupillary responses, fixation, saccade and blinking. To improve performance, a new pupillary response feature based on the Hilbert transform is proposed. This chapter adds to the work reported in the literature by considering affect detection using eye movement signals elicited in the more natural setting of the interactive virtual gaming environment, making it suitable for real-life natural HCI applications.

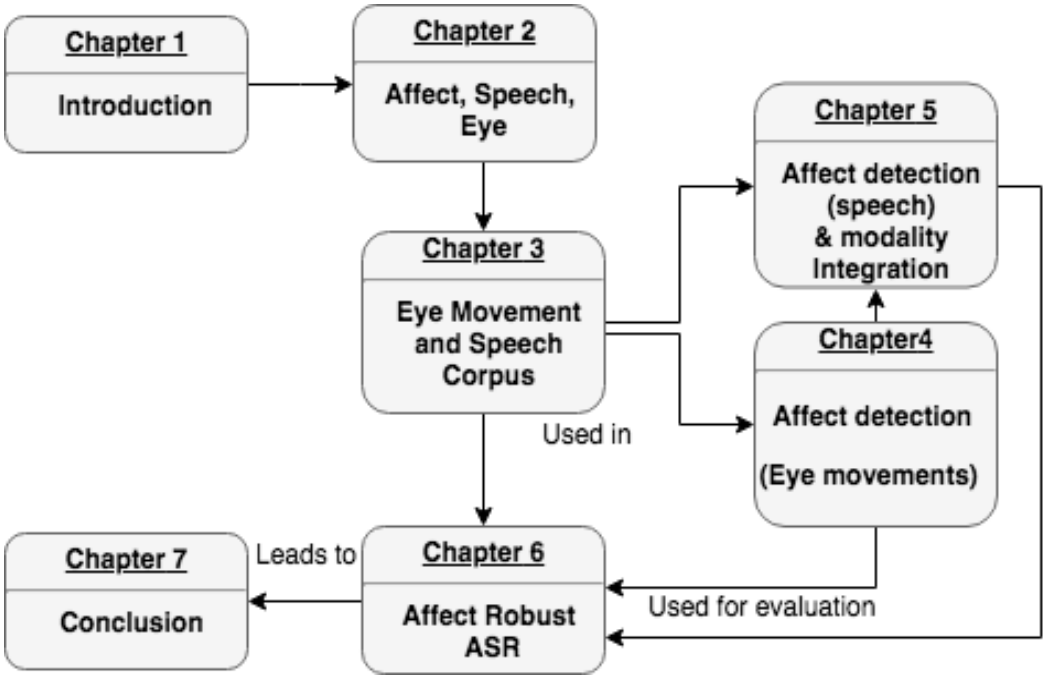


Figure 1.1: Block diagram of thesis structure.

- **Chapter 5** details the development of an affect recognition system using the acoustic and lexical content of speech signals, collected in the same interactive virtual gaming environment. In order to improve performance, it investigates a multimodal approach to the use of eye movements as additional modality to speech in affect detection.
- **Chapter 6** tackles a problem in ASR systems, which is enhancing their capabilities to recognise affective speech. It describes the development of a research-level baseline ASR system, detailing the proposed integration framework for an affect-robust ASR

system that incorporates affective information from eye movements and speech signals, unimodally and multimodally. The results reported in chapters 4 and 5 are used to evaluate the system. The results highlight the benefit of using eye movement as an additional modality to speech in order to improve the automatic recognition of affective speech, in terms of word error rate reduction.

- **Chapter 7** concludes the thesis by summarising the research, discussing the key contributions and making recommendations for future work.

Chapter 2

Affect, Speech and Eye Movements

This chapter explains affects and describes how they are modelled. After a review of affect detection from speech and eye movements, it discusses the impact of affects in the context of automatic speech recognition (ASR) systems.

2.1 Affects

Several answers have been proposed to the question "What are affects/emotions?" [41, 42]. Darwin defines affects as an expressive communication of the underlying internal state [43]. He suggests that the existence of affects is related to their survival value over the evolutionary lifecycle [43]. In pioneering work, James [44] presented the idea that affects are experienced in response to psychological changes, suggesting that affects are characterised by different peripheral physiological responses. The most widely accepted definition is the one proposed by Scherer [41]. He distinguishes between affects and other human characteristics (e.g. mood and attitude), defining affects as episodes of interrelated responses to internal or external events which involve continuous changes in cognition, physiological reactions and motor expression. It follows from this definition that affects are complex phenomena which:

- Cause behavioural changes (e.g. in facial expression, vocal pitch, etc.) and physiological changes (e.g. in muscle tension, heart rate, etc.).

- Occur in response to internal stimuli (e.g. memories and thoughts) or external stimuli (e.g. high-pressure situations, behaviour of others, etc.).
- Can be observed and measured.
- Are episodic in nature and can occur repeatedly, depending upon the occurrence of internal or external stimuli.

2.1.1 Models of affect

An important issue in developing affect detection systems is to determine which set of affects that are experienced by individuals are to be identified and detected by the machine. Two main models of affect have been widely adopted by the affective computing community to represent affective states. These are the categorical and dimensional models.

Categorical model

In the categorical model, an individual's affective states are represented in terms of discrete sets of affects (e.g. happy or sad). Psychologists have introduced several sets of keywords to model affect. For example, Ekman et al. [45] carried out cross-cultural experiments to study the universality of affects and defined the six basic affects (keywords) that can be detected from facial expressions, which are: happiness, sadness, surprise, fear, anger and disgust. Scherer [41] suggests that there may be from six to fourteen basic categories of affect, each characterised by different facial expressions, speech and physiological behaviour. This model has been widely used in affect detection systems, especially for detecting the basic affects from facial expression [12]. However, discrete affects may not describe the complex affective experiences which people exhibit in everyday interactions, conveyed through various sensory information modalities (e.g. face, body postures, speech and eyes) [12]. Moreover, the interpretation of the keywords used for categorising affective states can differ across languages and cultures [46].

Dimensional model

In the dimensional model, alternatively, affective states are represented within a set of continuous dimensions instead of discrete categories. These dimensions, which include

arousal and valence [5], are independent and can be combined to create multidimensional models. The two best-known multidimensional models are Russell's circumplex model [5] and Mehrabian's three-dimensional model [47]. The circumplex model [5], describes affects in two dimensions: arousal and valence. The arousal dimension refers to the current state of excitement, which ranges from arousal to sleepiness, while valence refers to the negativity or positivity of the current affective state, in a range from pleasure to displeasure (see Figure 2.1). The model incorporates keywords (discrete affects) which individuals often use to describe their affective states and which are spread round a circle representing the two-dimensional arousal-valence space. As shown in Figure 2.1, for example, 'happy' corresponds to relatively high arousal and positive valence (pleasure) while 'bored' represents low arousal (sleepiness) and relatively negative valence (displeasure). An alternative multidimensional model, proposed by Mehrabian [47], describes affects in a three-dimensional space of arousal, valence and dominance. The model is thus similar to Russell's, with the addition of the dominance dimension, which refers to the level of control the subject has over a situation or environment. Mehrabian claims that these three dimensions are sufficient to describe all affects experienced by humans [47]. However, Russell [5] argues that adding this dimension accounts for only a small amount of variance and that the circumplex model more efficiently models affective states. Furthermore, Moghimi et al. [10] found that valence and dominance dimensions are correlated and that Russell's circumplex model is sufficient to represent the affective states experienced by individuals, while they interacted with a virtual gaming environment, than Mehrabian's model.

The present study adopts the Russell's circumplex model to represent the affective states, because of its wide use in physiology, behavioural science [48] and affective computing (e.g. [9, 13, 14, 49, 50]). The advantage of this model is that it sufficiently describes both the intensity of affects and the affective categories (keywords) which are distributed round a circle in a single two-dimensional model.

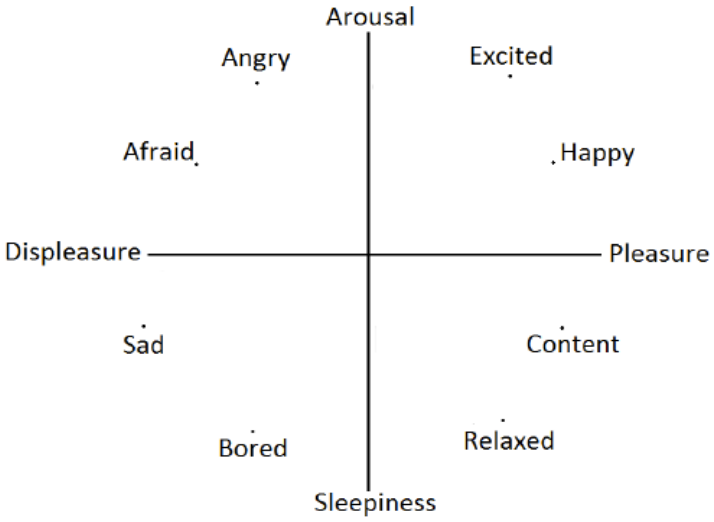


Figure 2.1: Russell's Circumplex model of affect (simplified version), with eight affective keywords distributed in the two-dimensional space [5].

2.2 On affect detection

Since affects play an important role in interactions among people, psychological research has given prominence to understanding and modelling affective phenomena [51]. Nass and Beeves [52] argue that individuals interact naturally with machines as they do with their peers, while Tao and Tan add that they naturally express affective responses to machines [1]. Thus, affect has attracted attention within the human-computing interaction (HCI) community with the desire to make these interactions more natural and intelligent. This emerging field of HCI research, referred to as *affective computing* after Picard [2], aims to facilitate the development of affect-aware interfaces that can automatically detect and respond to individuals' affects in order to enhance the effectiveness and naturalness of their interactions with machines. Underpinning affect-awareness in machines is the affect detection process, which uses computational techniques for extracting affective information and machine learning algorithms to develop a system that can associate human behaviours, observed in different modalities, with affects [6].

Researchers have utilised both behavioural and physiological modalities for detecting affects. Early studies focused on behavioural modalities such as speech, facial expression and body posture (e.g. [12, 53]), in unimodal approaches or by combining them multimodally for performance improvement [4]. However, individuals can relatively easily

manipulate and mask their behaviour, over which they have more control than over their physiological responses (e.g. pupil dilation in response to arousing stimuli); this makes felt affective states difficult to measure accurately from behavioural modalities in natural scenarios. More recently, the focus has shifted towards using physiological signals, because they can provide continuous measurements and seem to offer a more efficient and reliable means of affect detection [3]. Physiological signals provide measures of responses originating both from the autonomic nervous system (ANS)—such as electrocardiogram (ECG) (e.g. [54]), galvanic skin response (GSR) (e.g. [55]) and eye movement signals (e.g. [9])—and from the central nervous system, such as electroencephalogram (EEG) (e.g. [9, 55]) readings. However, the majority of these signals are inappropriate for use real-life HCI application due to their intrusiveness in terms of restrictions they pose on users (e.g. sensors have to be attached to the body) [56]. Eye movement sensing seems to be promising approach for affect detection in real-life HCI application, because eye movements can be observed naturally and the sensing is less intrusive than other physiological signals [15], with the availability of remote sensing devices.

With the expanding role of speech modality in many in HCI applications (e.g. Alexa, Google glasses, etc.) speech-based affect detection is considered vital in affect-aware interfaces [1]. Therefore, this thesis focuses on affect detection via speech as a behavioural source of affective information and eye movements as a source of physiological information. The following subsections provide a review of the literature on affective data acquisition methods and on affect detection via speech and eye movements.

2.2.1 Affective data acquisition methods

The first step in any affect detection system is the availability of suitable affective data. The literature refers to three types of data that have been used in developing and evaluating affect detection systems, namely acted, spontaneous and elicited affective data.

Acted data acquisition

The traditional approach to collecting affective data from individuals is to use the acted method, whereby actors are typically employed to follow scripts especially designed to

portray specific affective responses. The advantage of this approach is that collecting acted affects is relatively easy and requires lower resources than spontaneous data. However, they are unnatural and ecologically invalid [57], in that they do not realistically reflect actual affective behaviour. For example, Shahid et al. [58] found that acted and natural/spontaneous affective speech are perceived differently. One of the most popular acted speech corpora is the Berlin Database of Emotional Speech (Emo-DB) [7]. The affective speech in this corpus was obtained by asking actors to verbalise sentences in seven predefined affects: neutral, anger, fear, joy, sadness, disgust and boredom.

Spontaneous data acquisition

In the spontaneous approach, affective data is collected in natural/real-life situations such as call centres [8] or TV talk shows [59]. The advantage is that natural affective responses can be captured as they occur in real life. However, this approach requires effort to annotate the data (typically by expert), which can be expensive and time-consuming, as there is no prior knowledge (ground truth) about the affective states experienced by the subjects. Moreover, affective responses may not be present, may be hard to distinguish and may not be relevant to real-life HCI experiences. Finally, this method is applicable only to behavioural signals, such as speech, which can be visually/auditory observed. A popular spontaneous affective speech corpus is the VAM corpus [59], comprising spontaneous affective speech and visual data obtained from a TV talk show called "Vera am Mittag", annotated on the arousal, valence and dominance dimensions of affect, using the Self-Assessment Manikin (SAM) questionnaire [60].[9]

Elicited data acquisition method

A semi-natural and less expensive approach to collecting spontaneous affective data is to expose participants to stimuli or situations designed to elicit particular affective responses which they can subjectively report, thus providing a ground truth reference to the individual felt affects. Elicited data is typically recorded in carefully controlled laboratory conditions. Among the stimuli that have been developed for evoking affective responses are images,

[9] One of the most used tools for affect assessment and annotation. It comprises of manikins repressing different level of the three dimensions from low to extreme, an ordinal scale.

audio and movies. The individual is presented with a set of images, each of which may evoke a different affective state (e.g. [61]); for example, a family image can be used to evoke happiness. The most popular database of affective images is the International Affective Picture System (IAPS) [62], which consists of over a thousand images that have been evaluated on three affective dimensions using the SAM questionnaire. Alternatively, individuals are asked to watch short movie clips (e.g. [9]), or to listen to audio clips (e.g. [63]) to elicit specific affective responses that can be captured via both their behavioural and physiological signals. The most common affective auditory database is the International Affective Digital Sounds (IADS) database [64], comprising acoustic stimuli that have been evaluated on the same three dimensions as IAPS. Images, movies and audio are considered to be passive stimuli [57], to which affective reaction can be weak and with limited ability to stimulate real-life HCI experiences. Although they have been validated as reliable affective stimuli, it is not clear whether their content can be manipulated to evoke specific affective responses [65].

Spontaneous affective data can be elicited in an active way by simulating a HCI where individuals are asked to interact with an artificial system agent or robot, known as a Wizard-of-Oz (WOz) scenario [66]. In the simulation, the agent/robot is controlled and manipulated by a human operator to evoke the desired affects. This approach is commonly used to collect spontaneous affective speech data. Examples of corpora collected using this method are the Belfast Sensitive Artificial Listener (SAL) [67], SmartKom [68] and FAU Aibo [69].[10] Although this method is likely to elicit natural affective responses, it will succeed only if the participants are unaware that their affective responses are of interest; otherwise their expressed affects may be unnatural and deliberately acted. Moreover, there is no guarantee that the WOz scenario will reliably evoke affective responses, as this kind of interaction is considered a low impact manipulation [57].

An emerging active method for eliciting affects is virtual reality, defined as a medium "able to induce the experience of 'presence' in a computer-generated world. Presence is usually defined as the 'sense of being there' or the 'feeling of being in a world that exists outside the self'" [65]. Riva et al. [65] investigated the reliability of virtual reality environments as a method for eliciting affects. They used a 'virtual park' designed to elicit relaxation, neutral affect and anxiety, which subjects self-reported. They found that this

[10] Refer to Appendix A for corpora details.

virtual environment could effectively manipulate the individual affective experience and that the feeling of being present increased in the affective virtual environment (e.g. one designed to elicit relaxation) compared to the neutral one. This suggests that an affective response may not exist if the environment itself is unable to produce a sense of presence. Participants in another study were asked to interact with a virtual park designed to evoke sadness, while their EEG signals were monitored [70]. EEG activity during the experience was found to be different from the baseline measured before the experiment. More recently, Moghimi et al. [10] designed the affective virtual reality environment, used in this study, based on a gameplay scenario and subjectively evaluated it within Russell's circumplex model [5]. Based on the evaluation, they reliably concluded that the virtual gaming environment is capable of evoking different affective experiences by controlling the environmental parameters (e.g. faulty timer and changing the input modality from mouse to joystick). By employing the environment in an affect detection experiment and monitoring the evoked affective experiences via EEG, GSR and heart rate while participants interacted with the virtual gaming environment, they were able to detect them with high accuracy up to 98% [55].

Virtual reality is a very promising method for eliciting affective responses, as it can incorporate a rich and relatively natural interaction with individuals. More importantly, it increases the sense of being present in the real world. Arguably, gaming itself can naturally trigger various affective experience (e.g. frustration) [71]. To this end, a gameplay scenario in a virtual reality environment offers more realistic means of eliciting affects where interactions can be manipulated to influence individual affective experiences, unlike other forms of HCI (e.g. WOz). Consequently, this setup has been chosen for use in this study to elicit affective responses from participants.

2.2.2 Affect detection using the speech modality

Speech is a good modality for estimating affects, as it is the most natural and is widely used in communication. Speech conveys a variety of information about the speaker's affective state that is embedded implicitly in the signal, in the form of 'paralinguistic cues' [72-74]. These are nonverbal aspects of vocal behaviour associated with how speech is performed, which can be found in acoustic (e.g. high pitch) and linguistic (e.g. hesitation) features of

the speech signal [72]. The source of this nonverbal behaviour is the vocal tract, which is responsible for modulating the speech signal acoustically to produce speech with different characteristics [75]. Affective meanings can also be conveyed explicitly by the linguistic aspects of speech, e.g. semantically by the use of words such as 'sad', 'angry' or 'happy'. This linguistic content can provide a rich source of information for detecting underlying affective states. Both acoustic and linguistic features have been employed in affect detection, as the following subsections explain.

Acoustic aspects of speech

Several studies reported in the psychology literature have examined the correlation between acoustic features and affects [76]. It is argued that individuals modulate their speech to express different affective states, which can be observed from prosodic features of the speech [77]. Williams and Stevens [78] found that fundamental frequency (also called pitch), energy-related features and duration of pauses can indicate the arousal level of the speaker. This finding was confirmed later by Scherer [79], who also argued that valence is more relevant to speech articulation characteristics. Recently, Belyk and Brown [80] found fundamental frequency to be weakly associated with valence when analysing vocalised exclamations in response to valenced images. Murray and Arnott [81] identified the most relevant acoustic features that are influenced by speakers' basic affective states (anger, happiness, sadness, fear and disgust), namely fundamental frequency, speaking rate, intensity and voice quality. Perception studies have found a strong association between affects and voice quality [82]. Since the evidence of acoustic affective responses is well established in the literature, the acoustic aspect of speech has been extensively utilised to detect speakers' affects automatically.

The acoustic features most commonly employed to detect affective states are prosodic and spectral features [53, 83]. Prosodic features include fundamental frequency, energy and duration, while spectral features include formants and mel-frequency cepstral coefficients (MFCCs) [53, 83]. Prosodic features are generally considered to be effective at conveying affective information [83, 84]. Fundamental frequency was found to be more accurate than energy and formants in detecting four basic affects (neutral, happiness, anger and sadness) in acted speech in five languages of Pakistan [85]. MFCCs are also proven to convey useful affective information and can be used as complementary information to the prosodic features

[53, 86]. The majority of the contributions to the first INTERSPEECH Affective Challenge employed MFCC features to detect various basic affective states based on 2-class and 5-class problems in [87] from spontaneous speech in the FAU Aibo corpus [69]. For example, Vlasenko et al. [88] employed 13 MFCCs with their delta and acceleration coefficients, which are standard features of ASR systems; good results were achieved for both problems using a simple Gaussian mixture model (GMM) classifier compared to using HMMs, see Table 2.1. Bedoya-Jaramillo et al. [89] used MFCC features to detect anger, boredom and neutral affective states from acted speech in the Emo-DB corpus [7], using k-nearest neighbours (KNN), Bayesian linear and quadratic classifiers. They found that MFCC features were very effective in discriminating between angry and neutral speech. Wu et al. [50] found that MFCCs were the most informative features for detecting the arousal, valence and dominance dimensions from spontaneous speech data in the VAM affective corpus [59], in comparison to fundamental frequency, duration related features (e.g. speaking rate) and energy. In [90], MFCCs were found to be the most closely correlated with the three dimensions and the best detection results were achieved on the arousal dimension. Another study by [91] found MFCC features to be most effective in detecting affects on the valence dimension when evaluating spontaneous speech in a cross-corpus setting. Espinosa et al. [92] suggest the use of a different feature set for each affective dimension when estimating affects from spontaneous speech in the VAM corpus [59]. Most of the early work on speech-based affect detection systems focused on prosodic and spectral features of the speech signal. Recently, voice quality features such as jitter, shimmer and harmonics-to-noise ratio (HNR) have been used in measuring affects [53]. For example, [93] investigated various voice quality features of spontaneous speech to detect positive and negative valence and found them to perform well compared to prosodic and spectral features.

Prosodic features, spectral features and voice quality features are characterised as low-level descriptors (LLDs) computed over short time windows. Statistical (also called functional) features calculated from LLDs over the whole utterance or turn are extensively used in affect detection (e.g. [91, 94]). On the other hand, few studies have employed LLDs directly (e.g. [87, 88]). It is agreed that functional features are superior to LLDs in terms of detection accuracy, due to the long-term nature of affects in speech signals [53, 83]. Schuller et al. [94] compared different types of LLDs and functional features in order to find the set that would best improve detection performance in four affective categories (neutral,

motherese, emphatic and angry) of speech data collected while interacting with an Aibo robot. They found that duration and energy features were the most relevant LLDs and that the most relevant functional features for the detection task were sequential and combinatorial. It is suggested that better performance can be achieved by reducing and selecting the most relevant features. Other studies have attempted to use novel features extracted from speech signals which can be effective for the task. For example, Wu et al. [86] introduced modulation spectral features to capture the acoustic and temporal modulation frequency characteristics of the speech signal. Compared with the conventional spectral features, MFCCs and perceptual linear predictive (PLP) coefficients, these had superior performance in detecting basic affective categories of acted speech data in the Emo-DB corpus [7]. The authors conclude that modulation spectral features can provide complementary information to prosodic features in terms of the accuracy of detection of affective categories. However, for continuous detection of dimensional affects from spontaneous speech in the VAM corpus [59], MFCCs had superior performance, with higher correlation on the arousal and dominance dimensions than on valence, both alone and when combined with prosodic features. Sun et al. [95] evaluated the use of weighted spectral features to detect affect in three different acted speech corpora,[11] reporting better results when compared with typical spectral features (such as MFCCs) and when combined with prosodic features in both speaker-independent and speaker-dependent experiments using a support vector machine (SVM) classifier. They suggest that this approach effectively improves the performance of affect detection systems and serves as a complement to prosodic features.

As discussed above, the measurement of affects via the acoustic characteristics of speech has been studied extensively using a variety of speech datasets (acted and spontaneous and in various languages), showing that extracting the most relevant and informative features for the task is a vital step in developing speech-based affect detection systems. However, the best feature set for that task has not yet been identified, perhaps because of the use of different databases with diverse languages and cultures. For example, Schuller et al. [87, 96-98] defined different features sets for different databases. LLD features such as fundamental frequency, energy and MFCC are better established than

[11] Emo-DB, Surrey Audio-Visual Expressed Affect Database (SAVEE): http://kahlan.eps.surrey.ac.uk/savee/ and Speech Affect Database of Institute of Automation Chinese Academy of Sciences (CASIA): http://english.ia.cas.cn/

functional features. Eyben et al. [99] propose an affect detection toolkit called openSMILE,[12] with numerous predefined features that can be extracted and used for speech-based affect detection. Recently, the most common way to detect affects from speech has been to extract as many features as possible using brute-force extraction [53, 72], then to apply feature selection techniques to reduce the size of the feature set and select those with the potential to maximise detection performance [72]. Different feature selection approaches have been employed to detect affects from acoustic features, either wrapper or filter based [100], such as sequential forward feature selection (SFS) [86, 101] and Fisher discriminant ratio (FDR) [86]. Finally, the selected features are used to train machine learning algorithms for affect detection. Several detection methods have been employed to detect affects from speech. Table 2.1 lists some of speech-based affect detection studies grouped according to the employed detection method. The techniques most often used are: SVMs [50, 86, 91-95, 101], KNNs [89, 101], HMMs [87, 88] and GMMs [87, 88]. Other techniques such as, fuzzy logic [101], random forest [94] and Bayesian classifiers [89] are also reported in the literature. Grimm and Kroschel [101] found that SVM perform better than KNN and fuzzy logic when detecting affects on the three-dimensional space from spontaneous speech in the VAM corpus, see Table 2.1. Compared with HMMs and GMMs, which are usually used to model time-varying LLDs directly over time, SVMs have advantages such as global optimisation and good data generalisation bounds [102]. Moreover, due to the long-term nature of affective responses in the acoustic aspect of speech signals, SVMs tend to be preferred. Affect detection using deep learning-based methods have also attracted much attention recently with promising results (e.g. [103]). However, a large amount of data (typically hundreds of hours) would be required to achieve satisfactory results [104]. As can be seen in Table 2.1, for example, SVMs provided better results compared to deep learning-based networks for detecting dimensional affects from spontaneous speech in the VAM corpus, which consist of 12 hours of audio recordings. It seems that SVMs can achieve promising performance in affect detection from acoustic features and have been widely used in the literature [53]. The following discuss affect detection from linguistic aspect of speech signal.

[12] Other toolkits and software such as Praat and MATLAB can also be used for extracting acoustic features.

Table 2.1: Review of some speech-based affect detection studies.

Detection method	**Reference**	**Best results and remarks**
SVM	Wu et al. [86]	91.6% (acted)/ Emo-DB corpus, 7 affective categories. 73% (spontaneous)/ VAM corpus, 3 dimensions.
	Eyben et al. [91]	57.3% (spontaneous)/ within-corpora setting, 3 classes of valence. 53.3% (spontaneous)/ cross-corpora setting, 3 classes of valence.
	Espinosa et al. [92]	68% (spontaneous)/ VAM corpus, 3 dimensions.
	Schuller et al. [94]	63.1% (spontaneous)/ FAU Aibo corpus, 4 affective categories.
	Sun et al. [95]	75%-98% (acted).
	Grimm and Kroschel [101]	69% (spontaneous)/ VAM corpus, 3 dimensions.
KNN	Bedoya-Jaramillo et al. [89]	86% (acted)/ Emo-DB corpus, 3 affective categories.
	Grimm and Kroschel [101]	68% (spontaneous)/ VAM corpus 3 dimensions.
HMM	Vlasenko et al. [88]	68% (spontaneous)/ 2 affective categories. 35.2% (spontaneous)/ 5 affective categories.
GMM	Vlasenko et al. [88]	69.2% (spontaneous)/ 2 affective categories. 41.4% (spontaneous)/ 5 affective categories.
Fuzzy logic	Grimm and Kroschel [101]	58.3% (spontaneous)/ VAM corpus, 3 dimensions.
Random forest	Schuller et al. [94]	61.7% (spontaneous)/ FAU Aibo corpus, 4 affective categories.
Bayesian classifiers	Bedoya-Jaramillo et al. [89]	88%(acted)/ Emo-DB corpus, 3 affective categories.
Deep networks model	Amer et al. [103]	66.3% (spontaneous)/ VAM corpus, 3 dimensions.

Linguistic aspects of speech

The relation between linguistic content and affects is well studied in psychology [105]. Utilising linguistic content to detect affects depends on whether the speech data is acted or

spontaneous [53]. It has no value in the case of acted data, as this is based on predefined scripts. In spontaneous situations, speakers tend to express their affects explicitly using words which describe their current affective state [105]. Thus, spoken words can provide significant information about a speaker's affects [53]. In an early study, Lee et al. [106] introduced the concept of 'affective salience', which is defined as a measure of mutual information between the spoken word and a specific affective category. This concept was used to detect words in a spoken utterance collected in a call centre communication scenario, related to negative affects. Target affective states were determined by maximum likelihood estimation from the probability of the affective state given the salience of words. Improvement in detection was observed when the linguistic information was combined with acoustic features at the decision level, compared with using acoustic information only. Recent approaches to affect detection from linguistics aspects of speech have relied mainly on bag-of-words (BOW) representation,[13] a vector space modelling of the linguistic content. Schuller [90] compared different representations of BOW with SVM string kernels to detect affects continuously on the arousal, valence and dominance dimensions in the VAM corpus [59], using support vector regression (SVR). String kernels performed poorly compared to BOW representations, while a combination of BOW features and acoustic ones produced a better performance on the arousal and dominance dimensions only, indicating that valence is better modelled by linguistic content. Similarly, Truong et al. [13] found that linguistic content based on BOW representation was better at predicting the valence dimension, when evaluating affective speech captured from dyadic conversations while playing a multiplayer game. Polzehl et al. [107] show that features based on affective salience outperform BOW representations in detecting angry speech. Other features based on the concept of affective salience have been investigated and show promising performance compared to BOW representations (e.g. [108, 109]).

It has been suggested that discourse features such as rejection, repetition and rephrasing convey information about a speaker's affective state [72]. These can be combined with acoustic features at the decision level to enhance affect recognition performance. For example, [8] considered combining discourse information with acoustic and lexical information to detect negative affect in call centre dialogues. They found relative improvement over the use of acoustic information only. Earlier, Batliner et al. [110]

[13] This well-known approach to document retrieval is explained in Chapter 5, Section 5.3.2.

investigated whether discourse information could be used to detect 'trouble' from acted and spontaneous speech, collected in a WOz scenario where individuals engaged in a conversation with a dialogue system. They found that acoustic information (namely prosodic features) tended to be less reliable in the spontaneous case and that combining discourse information such as repetition with acoustic features improved performance. Part-of-speech (POS) tagging (e.g. nouns, adjectives, verbs and pronouns) has also been investigated to improve affect recognition task [53]. For example, Bulut et al. [111] found that affective states (namely happy, angry, sad and neutral) can be shown to differ at the POS level. Schuller et al. [94] investigated the use of different linguistic information, including disfluencies, POS, higher semantics and BOW to detect neutral, motherese, emphatic and angry affects in the FAU Aibo corpus [69]. The best performance was achieved with BOW representations using an SVM classifier. Moreover, an improvement was observed when this was combined with acoustic information. The superiority of BOW was observed in [112], when compared with SVM string kernels, POS and higher semantics using both correct transcription and ASR output.

Most of the above studies investigated the use of various kinds of linguistic information to detect affects and showed the possibility of combining linguistic and acoustic information to improve the accuracy of speech-based affect detection systems. Although the complementary value of linguistic content has been shown, obtaining this information in real life is challenging. The reference to spoken words has to be first obtained by an ASR system, which cannot reliably detect affective speech [33, 113]. Several studies have found that relying on ASR transcription degrades the detection performance from linguistic content compared to using the true transcription (e.g. [90, 107, 109, 112, 114]). Metze et al. [114] investigated the impact of ASR transcription quality when used for affect detection from angry speech, from data in FAU Aibo corpus [69]. They found that detection accuracy was degraded and that it could be relatively constant when the word accuracy of the ASR system was within a reasonable range. In their experiment the ASR system was adapted to the FAU Aibo corpus.

Issues on affect detection from speech

Despite progress in literature on speech-based affect detection systems focusing on investigating different acoustic and linguistic features (e.g. [86, 90]), and using advanced

affect detection algorithms (e.g. [103]), an important issue that needs to be tackled before speech-based affect detection can be considered ready for commercial use in real-life HCI applications is obtaining realistic and reliable affective speech data suitable for HCI applications, as the following explains.

- **Realistic affective speech data for HCI applications**

The availability of affective data plays a key role and is the major requirement for more realistic systems with high generalisability to real-life HCI applications. Speech-based affect detection systems are generally evaluated on three types of data: acted speech, spontaneous speech captured in real life and speech elicited while individuals are interacting with a conversational agent or robot, i.e. a WOz scenario (see appendix A for details of the datasets used). Although apparently promising results can be obtained with acted data, they are meaningless and irrelevant, because the data does not reflect realistic spontaneous affective speech responses. For example, affect detection systems evaluated on acted speech reach accuracies of approximately above 80% (e.g. [89, 95]), but this decreases significantly on spontaneous speech (e.g. [86, 91, 94, 103]), see Table 2.1. Spontaneous speech elicited in WOz scenarios might be suitable for HCI applications; however, the affects expressed can be falsely portrayed if the speakers are aware that their affective speech is of interest. Moreover, the likelihood that the WOz scenario will evoke affective responses can be relatively low [57]. Recently, spontaneous speech has been captured during dyadic conversations between friends while playing in multiplayer game scenarios [13]. The researchers showed that the captured speech could be used to detect the experienced affective states on different levels of arousal and valence. However, it is not clear whether the affective speech responses were evoked by interaction with the game, since social interactions may have affected their affective responses vocally. Therefore, the captured speech responses may not be suitable for HCI applications. Exploiting speech in an interactive virtual environment while chatting with an agent or commenting on a game can be a valuable approach to collecting realistic affective speech responses suitable for HCI applications.

- **Reliability of speech modality**

As mentioned above, speech-based affect detection systems perform very well when evaluated on acted data. This is because acted speech is very expressive and exaggerated and can be easily distinguished across different affects. On the other hand, the performance decreases significantly on spontaneous speech (e.g. [86, 91, 94, 103]). Batliner et al. [110] argue that information from the speech signal, namely its prosodic features, becomes less reliable in the case of spontaneous speech. One of the main reasons is that individuals can deliberately control and manipulate their speech to mask or fake their affective responses, making it difficult to distinguish their experienced affects from their speech [13]. Moreover, utilising linguistic content depends on ASR systems, which cannot deal reliably with affective speech [33, 113]. Some work has been done to improve affect recognition by considering affective information from other modalities in addition to speech, in a multimodal approach.[14] For example, [115] performed multimodal affect recognition by capturing speech, facial expressions and body gestures while participants interacted with an agent in a WOz scenario, using a Bayesian classifier. Spectral and prosodic features (namely MFCC, fundamental frequency, energy, speaking rate related features) were extracted from speech, while visual features were extracted from video recordings. The modalities were combined at both feature and decision levels. The researchers found that combining speech with each of the other modalities improved detection performance and that the best results were achieved by combining speech with body gestures. Moreover, adding information from facial expressions improved performance further. The majority of research in this area has used visual (facial expression/body posture) modalities, with facial expression being the most well studied [4]. These are behavioural signals which, in common with speech and unlike physiological signals, can be deliberately controlled, masked and faked.

Relatively few researchers have investigated multimodal affect detection using speech combined with physiological signals. A notable example is the Audio/Visual Emotion Challenge (AVEC) [116], which undertook multimodal affect detection using speech, visual and physiological signals (comprised of ECG and GSR), focusing on the detection of arousal and valence using the RECOLA corpus [117].[15] The multimodal integration results showed that information from ECG provided roughly up to 30% contribution to the overall all

[14] Multimodal approach for affect detection refers to the fusion of information from different modalities (e.g. speech and facial expression) during the detection process in order to enhance performance.
[15] The RECOLA corpus is briefly reviewed in Appendix A.

detection of both arousal and valence. However, these physiological signals tend to be contaminated with noise, and are difficult to measure as sensors have to be attached to humans' body [3], with no potential for remote sensing. This limits their scalability to real-life natural HCI applications, in contrast to eye movement sensing, a modality that has received increasing attention recently for use in affect detection (Section 2.2.3). However, it seems that combining the speech and eye movement modalities in affect recognition remains largely uninvestigated.

This thesis reports such an investigation of the combination of eye movement and speech modalities in affect detection tasks. Readers interested in audio-visual affect detection and multimodal detection in general should refer to the following review [4].

2.2.3 Affect detection using the eye movement modality

The eye is the most active modality in the human sensory system, providing a rich source of ANS activity, including that which indicates affective processes [118, 119]. There are multiple measures of eye behaviour, including pupillary response, fixation, saccade and blinking. Fixation occurs when a subject is focusing on a static target, while saccade comprises rapid eye movement which occurs between fixations, when the subject switches his attention from one target to another [120]. Pupillary responses are the physiological phenomena of slow dilations and contractions, mainly in response to changing light levels [119]. Finally, blinking is the momentary closure of the eyelid; it can occur voluntarily, reflexively or spontaneously [121]. Measures of these eye movements have been shown to provide an insight into individuals' affective states.

Going beyond the effects of light, pupillary responses due to affective processes have been explored. For example, Portal and Surakka [63] investigated pupillary responses while subjects were listening to different affective audio stimuli, selected from the IADS dataset. They found that pupillary response was significantly larger while listening to positive and negative audio clips compared to listening to neutral ones with medium arousal. This suggests that pupillary response is sensitive to highly arousing stimuli, regardless of positive or negative valence. Another study by [122] examined variations in pupillary responses to affective auditory storytelling content. They found that the responses were significantly larger when listening to a story with highly arousing negative content compared with neutral

content. These two studies indicate the possibility of using pupillary responses as a measure of an individual's affective experiences. Recently, this conclusion has been confirmed using visual stimuli. For example, Bradley et al. [61] showed that pupillary responses were significantly larger when subjects were presented with positive or negative arousing pictures, selected from the IAPS [62]. Kawai et al. [123] report that responses to negative images were larger than to positive images, suggesting that pupillary responses can be used to estimate negative and positive affects. Herbon et al. [124] found that pupillary responses to everyday computer tasks (e.g. searching) were strongly correlated with the valence dimension of affect.

The above studies have focused on pupillary response as a parameter to evaluate affective responses to different stimuli, suggesting that pupillary response reflects the affective process manifested through the ANS, mainly arousal level. Other eye movement parameters have also been investigated. For example, spontaneous blinking has been reported to be associated with negative affects, anxiety and arousal [125-127]. Tanaka [127] examined blinking rate and duration in response to high and low arousing tasks and found a tendency for both measures to increase during low arousal tasks. Gaze features have also been related to affects. For example, Subramanian et al. [128] showed that eye movement features (namely fixation frequency, fixation duration and saccade amplitude) were influenced by the affective content of movies. They found that these features were more strongly influenced by the affective content of positive and negative valence movie clips compared to neutral ones. Porta et al. [18] report that fixation duration, blink rate, in addition to pupillary responses, were good indicators of users' affective states in an e-learning environment.

Eye movement signals and their correlation with affective states have been studied extensively. It is believed that the eye movement modality has promise for measuring affective states, since it provides continuous measures of individuals' affective experiences and can be observed naturally and less intrusively. This is in contrast to other physiological signals such as EEG or GSR, which are often inappropriate for real-life HCI applications where the goal is to make the interaction with the machine as natural as possible, and which have no potential for the remote sensing that can be used to detect eye movements. To date, little work has been reported on the use of the eye movement modality in affect detection systems. A summary of the scant existing literature on affect detection via the eye movement

modality is given in Table 2.2, which shows that the studies can be categorised as either unimodal or multimodal, i.e. utilising eye movement data alone or combining this with EEG readings.

In 2012, Soleymani et al. [9] employed the eye movement modality in addition to EEG data to automatically detect individuals' affects in response to videos, on the arousal and valence dimensions. The videos were selected based on the affective content in movies and their ability to evoke affects was evaluated via an online questionnaire to form the ground truth. The eye movement and EEG data was collected as part of the MAHNOB-HCI audio-visual database [129]. The affective states were defined on the two dimensions respectively as low, neutral and high arousal, and negative, neutral and positive valence. The features extracted from eye movements were pupillary responses, blinking and gaze distance from the screen. The analysis of eye movement features showed that eye blink rate was higher while watching low arousal videos, while blink duration was shown to be higher for negative valence videos and pupillary responses were shown to be higher when watching neutral arousal videos. An SVM classifier with radial basis function, evaluated subject independently, was employed to classify the affective responses from both modalities. The researchers found that eye movements yielded a better detection performance than the EEG modality on both dimensions, achieving a detection accuracy of 71.1% and 66.6%, compared to 62.1% and 50.5%, on arousal and valence respectively[16]. By performing modality fusion at the decision level, the EEG modality contributed to improving the robustness of affect detection from eye movements on the arousal and valence dimension, by 5% and 2% respectively. Similarly, in [129], eye movement data including fixation features outperformed the EEG and peripheral modalities.[17] However, the detection accuracy from eye movement improved on the valence dimension only compared to the results obtained in [9], giving an accuracy of 68.8%. The results for combined eye movement and EEG were 67.7% and 76.1% on the arousal and valence dimensions respectively.

In [14], pupillary responses were investigated to detect affects on the valence dimension in response to movie clips, which were classified as either positive, neutral, or negative movies. It was found that the pupillary responses were larger when watching negative and positive videos compared with neutral ones. Two pupillary response features were extracted:

[16] The random accuracy was 33.3%.
[17] Peripheral signals include: GSR, ECG, breathing, and skin temperature.

power spectral density (PSD) and entropy over four frequency bands up to 1 Hz. The researchers employed SVM with linear kernel, subject dependently, to detect affects across the valence dimension and achieved better accuracy with entropy-based features compared to PSD. However, pupillary response did not outperform the EEG modality. The best result was achieved by multimodal fusion at feature level, giving 73.6% accuracy. Detection of affective states on the valence dimension was improved by considering other eye movement signals including fixation, saccade and blinking in addition to pupillary responses [130]. A performance similar to the EEG modality was achieved, giving 77.8% and 78.5% accuracy respectively, on a subject-dependent setting. Moreover, the study found that the two modalities carried complementary information and that combining them yielded better performance, up to 87.6%, at the decision level.

In a unimodal study, Raudonis et al. [131] developed a pupil analysis algorithm based on an artificial neural network (ANN) in order to detect neutral, disgust, interest and funny affects in response to images. Three eye movement features were employed: pupillary responses, gaze position and speed of the eye. The eye movement features analysis showed that the pupillary responses, user's focus of attention (fixation) and speed varied among the different affective states. The study achieved an average detection accuracy of 87% (random sampling of training and testing), with the funny and interest states being the easiest to detect. Similarly, Aracena el al. [49] performed affect detection using pupil responses and gaze position (fixation) in response to positive, neutral and negative images from the IAPS dataset. A decision tree neural network, a hierarchical classification, was used to resolve the confusion between neutral and positive states, achieving an average accuracy of 72.1%, on a subject-dependent basis. Alghowinem et al. [132] employed GMM and SVM classifiers to detect positive and negative affective experiences from eye movements, in response to video clips of affective movies followed by interview. The eye movement features used were pupillary responses, head rotation, blinking and saccade speed. They found that pupillary responses and blinking rate were higher in response to negative videos and the interview, while head rotation to the left was longer during an interview after watching a negative movie clip. Average accuracy of 65.3% and 66.7% was achieved for affect detection in response to movies and interview respectively.

The literature on affect detection using the eye movement modality has shown promising results, with better performance than the EEG and peripheral modalities (e.g. [9,

129]). This could be due to the nature of these signals as they tend to show high variability among individuals especially for EEG [133]. The research has focused on passive/noninteractive stimuli such as videos or images, from which it is difficult to make generalisations applicable to real-life HCI applications. It appears that using eye movement for affect detection in a virtual reality environment remains unexplored. This thesis reports on affect detection from eye movements by utilising a virtual reality environment for affect elicitation.

2.3 Affects in automatic speech recognition

Over the last few years, automatic speech recognition systems, ASR, has facilitated more natural human-computer interactions. So-called speech-enabled interfaces have been widely used in many applications. Popular examples are voice search systems (e.g. Google Voice Search, which enables users to search the internet by voice command), virtual assistants and games (e.g. Microsoft Xbox 360 games where players can interact with game characters using speech). Recognising and understanding users' speech is important for the interaction process. However, it has been shown that individuals' underlying affective states influence their speech behaviour [77], with changes in fundamental frequency, energy and MFCC features being commonly reported (Section 2.2.2). This makes the verbal content difficult to recognise and degrades the performance of ASR systems, thus compromising the naturalness of the interaction [113].

Few researchers have investigated the impact of affective speech on ASR system performance and suggested ways of improving it. For example, [134] used data from the FAU Aibo corpus [69] to compare the performance of an ASR system on children's spontaneous speech produced in a neutral affective state with its performance on motherese, emphatic and angry speech, in two scenarios. They first compared the performance of speech recognisers trained exclusively for each affective state in a cross-affect setting. They found that using a speech recogniser trained on the corresponding affective state always gave a better performance than one trained on neutral speech. Secondly, affect adaptation was performed by adding affective speech data to the neutral speech to train an affective recogniser for each of the three states. The adaptation was based on adding several copies of the affective speech data. Performance was compared against the speech recogniser trained

on neutral speech. The results showed that incorporating affective speech data with the neutral data improved recognition performance for angry and emphatic speech only. The authors conclude that ASR system performance can be enhanced during affective speech by the adaptation of acoustic and linguistic models to the corresponding affective state.

Table 2.2: A summary of affect detection literature using eye movement modality. Dep: subject dependent evaluation. Indep: subject independent evaluation.

Study	**Type**	**Stimuli**	**Affects**	**Subject**	**Method**	**Best Result**
Soleymani et al. [9]	Multimodal	Videos	Arousal Valence	24	SVM, Indep	76.4% (arousal – multimodal) 68.5% (valence– multimodal)
Soleymani et al. [129]	Multimodal	Videos	Arousal Valence	27	SVM, Indep	67.7% (arousal – multimodal) 76.1% (valence– multimodal)
Zheng et al. [14]	Multimodal	Videos	Valence	12	SVM, Dep	73.6% (multimodal)
Lu et al. [130]	Multimodal	Videos	Valence	9	SVM, Dep	87.6% (multimodal)
Raudonis et al. [131]	Unimodal	Images	neutral, disgust, interest and funny	47	ANN, Random	87%
Aracena et al. [49]	Unimodal	Images	Valence	4	Decision tree neural networks, Dep	72.1%
Alghowinem et al. [132]	Unimodal	Videos interview	Positive and negative	60	SVM, GMM, Indep	65.3% (videos) 66.7% (interview)

Vlasenko et al. [135] evaluated the performance of an ASR system on affective speech in a cross-corpus setting. The acoustic model was adapted to acted affective speech using the maximum likelihood linear regression (MLLR) and maximum a posteriori (MAP)[18] adaptation techniques. When evaluated on affective spontaneous speech, a relative improvement of 7.1% in word accuracy over the ASR system trained on neutral speech was observed. These results suggest that training a speech recognition system on neutral speech, then adapting it to acted affective speech, can improve the performance of a system when used to recognise affective spontaneous speech.

Schuller et al. [33] investigated whether a speaker's underlying affective state can be used to enhance the performance of an ASR system during angry speech, by adapting the ASR acoustic model to the speaker's affective state as inferred from the acoustic features (namely MFCC and energy-related features) of his/her speech using an SVM classifier. They achieved a 32.9% relative reduction in word error rate on single-digit data and 16.59% for angry sentences.[19] Although this shows the feasibility of integrating speakers' affective states into an ASR via acoustic modal adaptation, the reported performance is expected to be achieved only on acted affective speech data, because it can be distinguished and modelled more easily than spontaneous affective speech. Similarly, Vegesna et al. [34] showed some improvements in the ASR system performance during affective speech when a speaker's affective states (namely happy, angry, compassion and neutral), as inferred from acoustic features, were used to select the corresponding affective ASR acoustic model that was trained on synthetically generated affective speech from prosody modification of acted speech data captured in the natural state.

Some research effort has focused on incorporating affective information to enhance language models in ASR systems. One of the first attempts was made by [113], who argued that recognition of spontaneous affective speech can be improved by increasing the representation of affective words in the neutral corpus, in particular by training the language model on affective content. To do so, they derived affective sentences from a general-purpose corpus, the British National Corpus (BNC),[20] using the Whissel affective dictionary [136]. They then used these sentences to augment the BNC corpus, thus creating an affect-oriented corpus. The language model trained on this corpus achieved an average 20%

[18] An overview of ASR system and adaptation techniques are provided in Chapter 6.
[19] speakers verbalising single digit (0-9) and selected sentences form the Wall-Street-Journal (WSJ) database in natural and anger states.
[20] http://www.natcorp.ox.ac.uk/

improvement in recognition rate when used in an ASR system to recognise spontaneous affective speech, captured while speakers engaged in conversation with an artificial agent. While this technique showed some improvements, these were not necessarily due to the underlying affective states of the speakers, because the researchers added multiple copies of the sentences that contained affective words, thus increasing the occurrence of easily recognisable neutral words (e.g. 'a'). The method also completely ignored the underlying relation between the words and the affective state of the speakers. Karkhedkar [35] proposed an affective language model based on the acoustic features of spoken affective words, which were first identified according to the detected affective state of the speakers on the arousal, valence and dominance dimensions. The baseline language model, trained on the Switchboard[21] corpus [137], was then adapted by redistributing the word probabilities according the detected affective states. The results showed a 1.38% reduction in perplexity, a measure of a language model predictably of a word sequence[22], over the baseline language model and no improvement in ASR performance was reported. The small improvement in perplexity probably occurred because the affect detection system performed poorly on spontaneous speech from the Switchboard corpus, showing low correlation with the observers' annotations. This indicates that relying on the speech modality as a source of affective information to enhance the language modelling of spoken affective words may not be reliable.

Studies that have evaluated the impact of affective speech on the performance of ASR systems, either by training the models on affective speech data [134] or by using adaptation techniques [135], suggests that to enhance the capabilities of an ASR system during affective speech, it is important to select the right models for the recognition task, corresponding to speakers' affective states. Research involving the adaptation of ASR models to speakers' affective states have tended to use information from the speech modality, mainly acoustic features [33-35]. However, this method may not be suitable for dealing with spontaneous speech, because speech-based affect detection systems are less reliable when dealing with spontaneous speech (Section 2.2.2). The present study investigates the use of affective information from the eye movement modality to enhance the capabilities of an ASR system during affective speech.

[21] Telephonic conversations corpus typically used for training speech recognition systems.
[22] Refer to Section 6.4.1 for future details.

2.4 Summary

This chapter has outlined the definition of affect and the models used to represent it (Section 2.1), then reviewed work on affect detection using speech and eye movement modalities (Section 2.2). This included discussion on affective data acquisition approaches highlighting the use of virtual environments for eliciting more realistic affective responses suitable for real-life HCI applications. Based on the discussion of the existing affect detection studies, using speech modality can be less reliable in the case of spontaneous speech. Some proposed approaches to improve the performance is done by considering affective information from other modalities (e.g. physiological signals) in addition to speech. Eye movement seems to be underutilised modality for affect detection, especially in a natural setting (e.g. virtual environments), and the existing studies are presented with the promise that eye movements are suitable for measuring affects compared with other physiological signals, such as EEG. In addition, they can be observed naturally using remote sensors. For natural HCI, therefore, it is more reasonable to consider using eye movement to aid affect detection from speech modality without compromising the naturalness of the interaction.

The chapter ended by discussing the issue of affects in ASR systems and some proposed approaches of performance enhancement during affecive speech (Section 2.3). Among the mentioned approaches for affect-robust ASR systems, some have showed the benefit of adapting ASR models trained on affective speech, while others considered using affective information as inferred from the speech modality for ASR adaptation. In the case of spontaneous speech, using affective information from the speech modality may not be reliable. Thus, considering other source of affective information, from eye movement modality, is likely to be more beneficial for affect robustness in ASR systems.

This thesis aims to investigate the use of eye movement and speech data captured in an interactive virtual environment for the affect detection task and its application for affect-robust ASR system, in order to facilitate affect-aware interfaces for natural HCI. The following chapter details the experimental work to collect eye movement and speech data in an interactive virtual environment.

Chapter 3

Affective Eye Movements and Speech Corpus

In order to develop and evaluate affective detection systems suitable for real-life HCI application, a corpus of realistic affective behaviour occurring during human-computer interaction that can lead to a meaningful result in the real world is required. This chapter describes the experimental method for developing an affective corpus of eye movements and speech utilising an interactive virtual gaming environment. The motivation behind collecting the corpus and the interactive virtual gaming environment in which affects were elicited are described. The chapter characterises the participants and describes the apparatus, experimental task and procedure for data collection. In addition, discusses the data synchronisation scheme and quality assessment procedure. Hereinafter, the affective game corpus developed is referred to as the 'aGame' corpus.

3.1 Motivation for developing the aGame Corpus

Various corpora are widely available for affective computing research. Early corpora contained affective data from a single modality mainly speech (e.g. EmoDB corpus [7]) . In recent years, affective computing researchers have shown increasing interest in more complex multimodal data to improve recognition performance (e.g. [138]). The majority of the existing affective corpora consist either of acted data, based on a predefined script to portray a range of affects, or of spontaneous data gathered in a real-life situation such as a

call center [8]. Acted data inaccurately represent realistic affective behaviour, while spontaneous data requires more effort to collect and may be compromised by biased annotation. In response to the need for affective detection systems based on natural unacted behaviour, some corpora have induced affective responses using passive stimuli, such as videos [129], allowing developers to capture more natural data compared with the acted method. However, it is not clear how the content of passive stimuli can be manipulated to evoke specific affective responses and it is difficult to make generalisations applicable to real-life HCI applications. Others have used WOz scenario to simulate HCI while inducting affects (e.g. [67]), however, the likelihood for WOz method to induce affective responses can be relatively low. Details about existing affective corpora can be found in appendix A.

Riva et al. [65] have shown that interactive stimuli, namely virtual reality, can be used efficiently to elicit affects. This method is more natural than laboratory experiments which passively induce affects, because it can give the feeling of being in a real-life situation. Virtual reality environments are very promising for eliciting affective responses in more natural, less constrained settings. In addition, playing games naturally triggers affective responses [71]. It is believed that utilizing games in virtual reality context can elicit natural affective responses [10]. Hence, the aGame corpus was collected in an interactive virtual gaming environment, described in Section 3.2, to allow the quasi-natural elicitation of affects close to real-life responses.

Speech modality has been widely used in HCI applications, such as virtual assistances, smart glasses (e.g. Google glasses) and gaming (e.g. Microsoft Xbox games). Exploiting speech in a virtual interactive gaming environment while chatting or commenting on the game can allow collecting naturalistic affective responses. However, speech is a behavioural signal, which, in common with facial expressions and body posture, can be controlled and manipulated by individuals, making the felt affects difficult to measure. Physiological signals, on the other hand, can provide continuous measurements and seem to offer a more efficient and reliable method of affect recognition [3]. Combining data from both behavioural and physiological modalities can be beneficial [116]. A major difficulty is that most methods of measuring physiological signals (e.g. EEG and GSR) are intrusive with no potential for remote sensing and are often inappropriate for real-life HCI applications such as interactive gaming, where the objective is to make the interaction between the player and the game as nearly natural as possible. Moreover, setting up such sensors is not easy and can

discourage users. Lately, eye movement tracking has emerged as a means of measuring affect, because the signals are natural and their observation is more efficient and less intrusive than for other physiological signals [15]. The present study therefore uses eye movements to represent affective responses via physiological behaviour.

A few studies have captured both speech and eye movement data for HCI applications. The MAHNOBA-HCI corpus [129] captured both audio-visual and physiological data, including eye movements in response to short movie clips. However, the amount of audio data is insufficient for affect analysis and consists only of laughter. Cooke [139] collected an eye/speech corpus of synchronised speech and eye movement data during collaborative visual tasks, for the development and evaluation of gaze-contingent ASR systems. Shen [140] developed a similar corpus to investigate the gaze and speech relationship and used it to evaluate ASR in noisy environments. However, these datasets specifically contain visually oriented information collected in a laboratory setting, unavailable publicly, and include no affective information. To the best of the author's knowledge, no corpus of affective eye movements and speech has been collected in a more natural setting such as the interactive virtual gaming environment used in this study.

The motivation for collecting the aGame corpus is to obtain a set of appropriate and realistic affective eye movements and speech data elicited in a more natural environment in order to develop an affect detection system suitable for real-life HCI applications. This corpus facilitated the evaluation of an affect-robust ASR system in the present research (Chapter 6).

3.2 Description of the affective virtual gaming environment

This study uses an affective virtual gaming environment proposed by Moghimi et al. [10]. The environment is capable of eliciting various affective responses, as described by Russell's circumplex model of affect [5]. It was originally developed for virtual reality healthcare research [141]. It is based on a single player speedboat simulation scenario and consists of different sets of sub-games that target specific affective responses. In the neutral case of the game, the player drives a small boat freely. S/he can collect floating life rings to gain points and can either finish the game by passing the finish line or continue exploring

for no longer than five minutes, otherwise the game terminates automatically. Figure 3.1 illustrates the speedboat simulation game in the neutral scenario. The sub-scenarios of the game involve avoiding mines, firing torpedoes and negotiating a maze. The content (parameter) of the environment is manipulated to create scenarios that are capable of evoking different affective experiences within the circumplex model. This is achieved by generating a sudden event (e.g. a faulty timer, faulty controller or explosion) within a game, or by imposing conditions which persist for the whole duration of the game (e.g. time limitation, changing the input from mouse to joystick, and forced feedback by vibrations), see Table 3.1 for an example. Four affective games were selected for use in this research, to elicit different affective responses from players, in addition to the neutral game. Their selection was based on their mean and standard deviation values for arousal and valence, according to [10]. The total number of games was limited to five of the original ten so that players would not be fatigued by wearing the eye tracker for a long period of time. Table 3.1 lists the five selected games along with their content parameter settings and their arousal and valence labels, the 'ground truth'. Further details of the affective virtual gaming environment design can be found in [10].

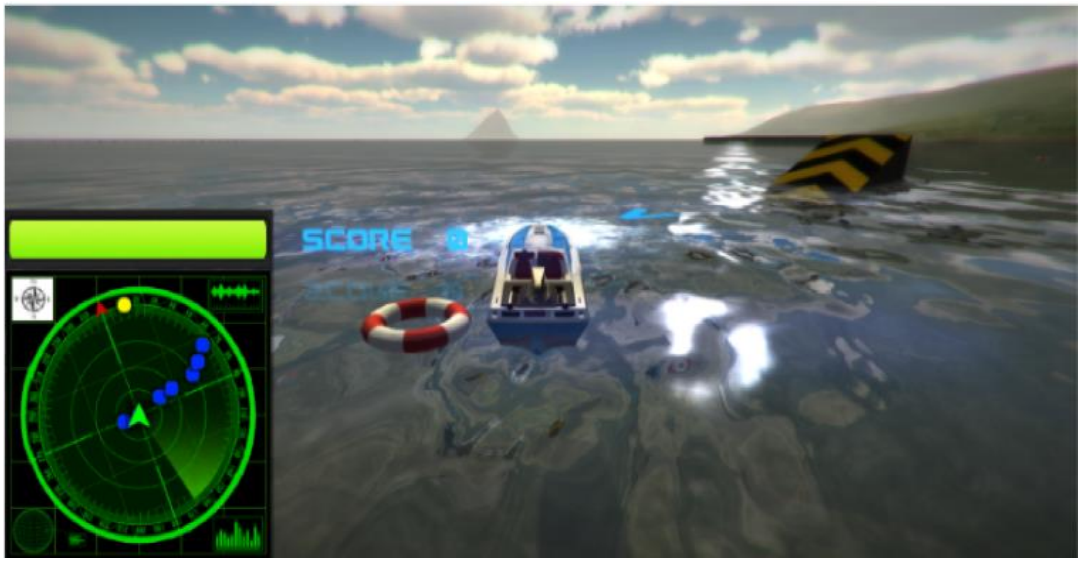


Figure 3.1: A snapshot of the affective virtual gaming environment, a neutral scenario, adapted from [10].

Table 3.1: Five affective games with their parameter settings and ground truth lables, according to Russell's circumplex model of affect, selected from [10].

Game/ parameter	**Affective Label 'Ground truth'**	**Environment**	**Time limitation**	**Timer**	**Controller**	**Barrier**	**Input**	**Camera**
Neutral (N)	Neutral valence Neutral arousal	Free exploration	No	Normal	Normal	Visible	Mouse	No shaking
PV/LA	Positive valence Low arousal	Free exploration	No	Normal	Normal	Visible	Joystick without force feedback	No shaking
PV/HA	Positive valence High arousal	Torpedo firing	Yes	Normal	Normal	Visible	Joystick with force feedback	Shaking and blurring
NV/HA	Negative valence High arousal	Avoiding mines	Yes	Faulty	Faulty	Invisible	Mouse	Shaking and blurring
NV/LA	Negative valence Low arousal	Maze	No	Normal	Normal	Visible	Joystick without force feedback	No shaking

3.3 Method of developing the aGame corpus

The experimental setup for collecting eye movement and speech data for the aGame corpus is depicted in Figure 3.2, which illustrates the layout of experiment, and the apparatus used for playing in the interactive virtual gaming environment, and capturing eye movement and speech data. In the following subsections the conduct of the experiment is described, detailing the participants, apparatus, experimental task and procedure.

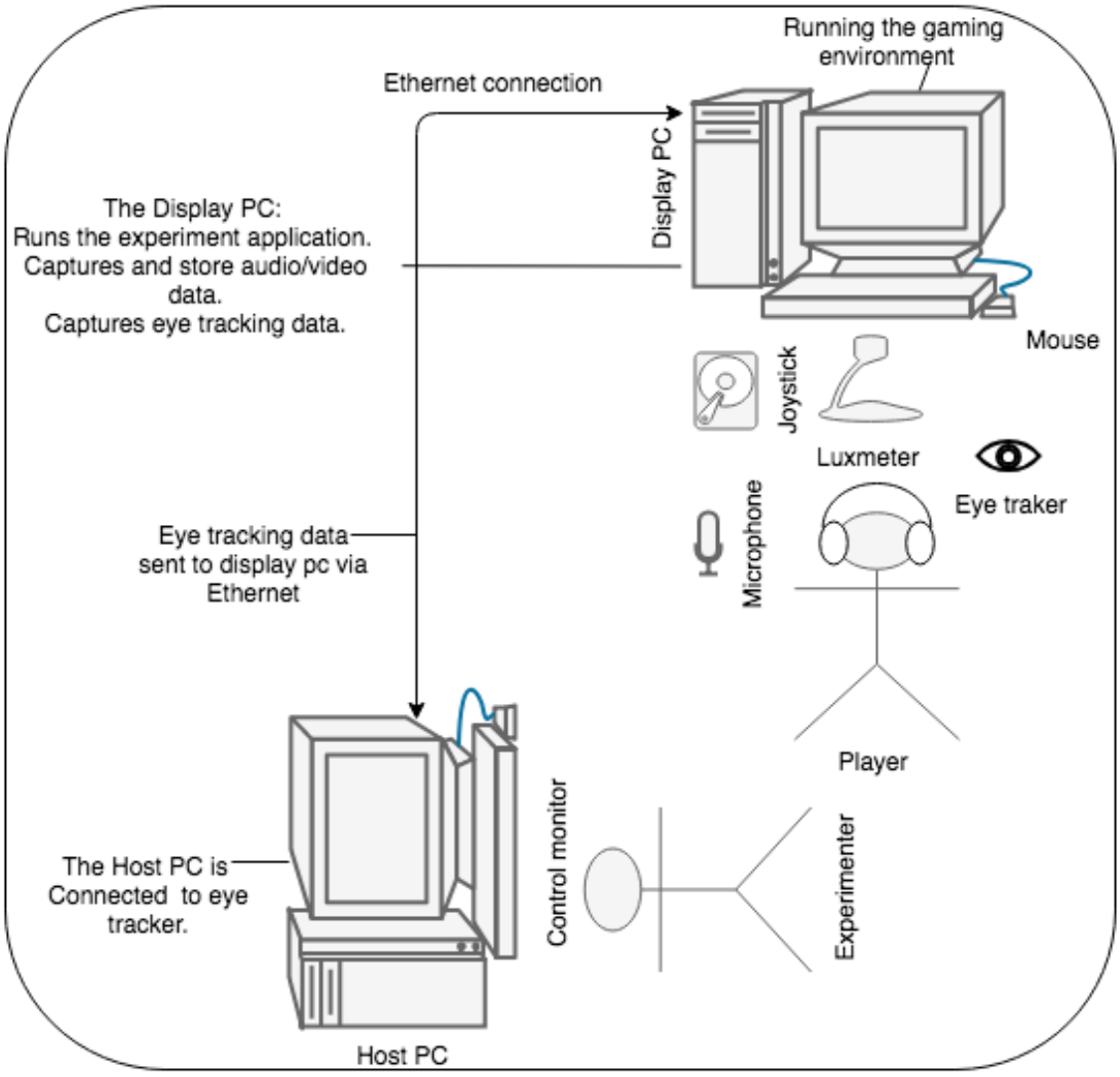


Figure 3.2: Experimental setup for eye movement and speech corpus collection.

3.3.1 Participants

Participants in the study were eighteen students with normal or corrected-to-normal vision at the University of Birmingham, of varied cultural and educational backgrounds, four of whom were female; the age range was between 19 and 35 years (M = 25.8, SD = 5.8). Recruitment was via advertising email. Participants were given a voucher as compensation at the University's standard rate. Three participants participated in the preliminary experiment and the remaining fifteen in the main aGame corpus collection experiment. All participants were gamers with more than two years of gaming experience using a PC and/or gaming consoles; three of them wore glasses. Their spoken English was sufficient for the task. Hereinafter, participants are referred to as 'players'. Ethics permission was granted in order to conduct the experiments, according to University of Birmingham ethics procedure[23], (refer to appendix B for the Consent Form).

3.3.2 Apparatus

The players' eye movements were captured using an EyeLink II eye tracker.[24] The EyeLink II is a head-mounted tracker operating at a 250 Hz or 500 Hz sampling rate. Its key features are: (i) the use of two video cameras, infrared illuminators and the corneal reflection methods to record eye movement events (fixation, saccade, pupillary responses and blinking) in both left and right eyes, (ii) accurate tracking results with average accuracy of 0.5º visual angle, (iii) fast and simple setup [27].

In accordance with the EyeLink II user manual [27], two PCs were used to set up the eye movement tracking system: a host PC for the system and a display PC (Figure 3.2). The host PC was responsible for performing real-time eye movement tracking, including the calculation of eye position and the online detection of eye events such as saccade and fixation. In addition, it was used to monitor and set up the eye tracker and the calibration process. All eye movement data was stored on the host PC and transmitted via Ethernet to the display PC, which hosted the applications required to run the software for the experiment and displayed the calibration targets and eye movement data.

[23] University of Birmingham research ethics procedure can be found here: https://intranet.birmingham.ac.uk
[24] SR Research Eyelink II, http://www.sr-research.com.

The players' speech was recorded using a USB external sound card (Edirol UA-25[25]) to ensure the high-quality audio recordings required for accurate transcription. All speech was recorded at a sampling rate of 44.1 kHz and stored in stereo WAV file format using a Shure WH20,[26] a high-quality headset-mounted studio microphone.

Equipment for the game included a 19-inch display screen with 1280 x 1024 resolution and 60 Hz refresh rate, a Microsoft mouse and a Logitech Wingman 3D force feedback joystick. The illumination of the game was measured using an Extech HD45 Luxmeter,[27] attached to a desk-mounted holder and placed as close to each player's nose as possible, in order to measure the amount of light influencing his/her pupillary responses. The distance of players' eyes to the screen was set to 50 cm, in order to maintain tracking resolution within the recommended range [27].

3.3.3 Recordings and software

The game simulation environment was a 3D Unity-based game running on the Microsoft Windows 7 operating system on a standard display PC. The eye movements and speech data were recorded using EyeLink Screen Recorder software[28], which allowed parallel recording of eye movement data (including fixation, saccade, pupillary responses and blinking), speech and on-screen activities on the EyeLink II host PC. An example of the video captured of on-screen activities while a player is playing in the virtual gaming environment is shown in Figure 3.3.

In addition to the data recorded by the eye movement tracking system, each player's experience was collected using a questionnaire and stored in a plain text file, in order to validate the elicited affective state associated with each game (refer to appendix C for the questionnaire). According to [10], the first section of the questionnaire used seven-point questionnaire scale to evaluate the player's affective state in terms of valence and arousal. The second listed eight discrete affects (relaxed, content, happy, excited, angry/annoyed, afraid, sad and bored) from Russell's circumplex (described in Chapter 2), asking the player to choose the closest affective label to his/her experience during the game. The illumination

[25] https://www.roland.com
[26] https://www.shure.co.uk
[27] A digital light meter: http://www.extech.com/
[28] A new release from SR research that provides integrated audio and video recordings (Beta version: 1.0.0.179).

of the game during the experiment was measured continuously in lux (lumen per square meter) and stored in a separate log file, to be used to account for the contribution of light reflex to the pupillary responses, as explained in Chapter 4. All data was marked with a timestamp to allow easy alignment.

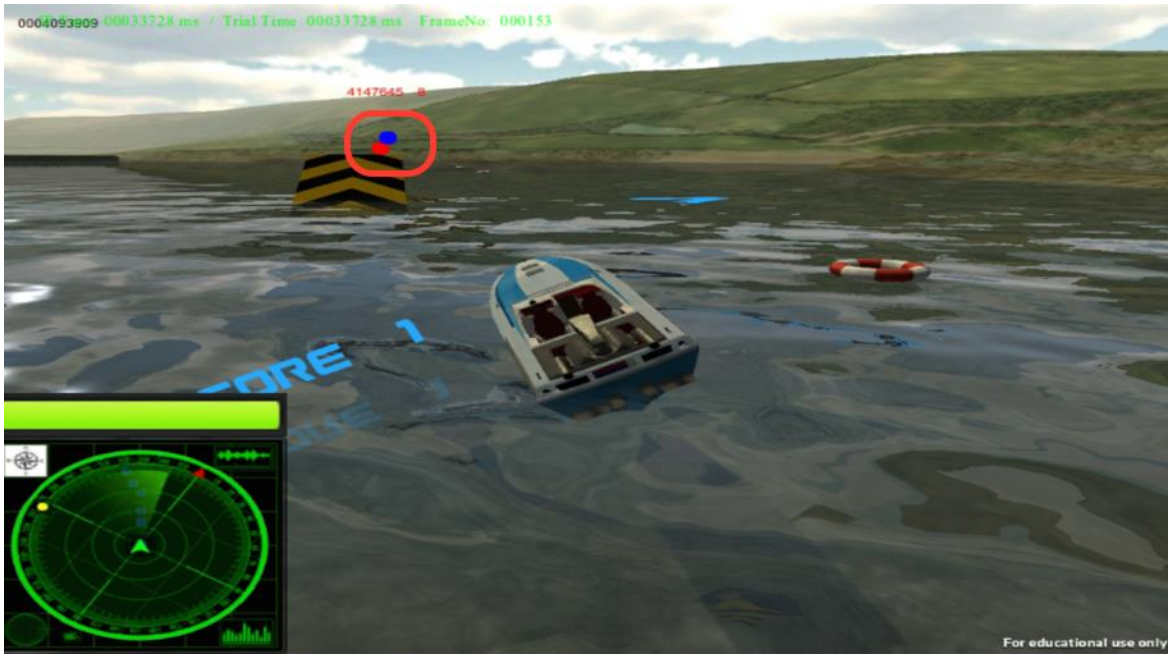


Figure 3.3: A screenshot of the playback video of a neutral game session (Player 2) recorded using the EyeLink Screen Recorder software. The red and blue circles represent the position of the player's eyes on the screen (fixation event). The numbers show the timestamp (system clock) for host and display PCs.

3.3.4 Experimental task and procedure for corpus collection

Collecting behavioural and physiological data from individuals is expensive. For the purposes of this research, two experiments were performed: a preliminary one and the main corpus collection experiment. The experimental task of both experiments utilises the affective virtual gaming environment described in section 3.2. Figure 3.4 illustrates the general experimental procedure followed in both experiments. They began by a training session to encourage familiarity with the gaming environment and introducing the task, then a standard EyeLink II calibration procedure to estimate the gaze position on the screen followed by the actual experimental task. Each player played the five affective games detailed in Table 3.1. S/he began the experiment by playing the neutral game, that is less likely to elicit a strong affective experience among players [10], then the next affective game

was selected at random. After each game the player self-report her/his affective experience by completing the questionnaire and select an affective keyword (Section 3.3.3), which appeared automatically on completion of the game. Each game lasted for a maximum of five minutes, after which 20-30 seconds were spent in completing the questionnaire. To keep them engaged in the game, they were given the goal of collecting as many life rings as possible.

The players' speech was captured by instructed them to speak out loud, the so called 'think aloud' technique [142], while playing the games. They were free to verbalise whatever they were looking at, thinking about or feeling. This behaviour is typical of in-game chatter and game commentaries webcast on video services such as YouTube. Collecting speech data using this technique is considered as a valid way to gain insights about individuals' underlaying cognitive process and states [142]. It is a commonly used approach in HCI usability research [143] and has been applied recently to evaluate users' affective experiences in online shopping [144]. It is believed that speaking out load while playing in the interactive virtual gaming environment allows spontaneous affective speech responses to be captured.

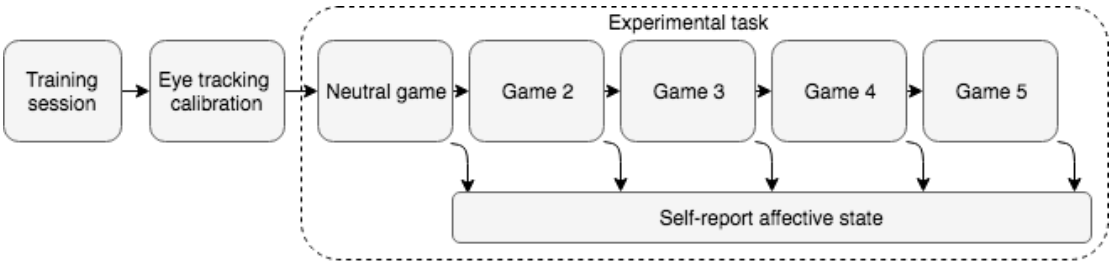


Figure 3.4: Experimental procedure for collecting the aGame corpus.

Training session

Before the experiment, players completed a training session, during which they were given an information sheet setting out instructions, a detailed description of the task and the purpose of the experiment. They were made aware of the EyeLink II tracking sensor and told that their speech would be recorded during the experiment. They were also shown an example of popular YouTube gamers' video commentary on games to help them understand the task. During the session players were invited to ask the experimenter questions and to

repeat the training task until they became familiar with the environment. The training task allowed them to explore the gaming environment and the use of the input devices and sensors. They were also asked to practice speaking out loud in order to avoid unfamiliarity in the actual experiment. This training procedure was followed in both the preliminary and the main aGame corpus collection experiments.

Experiments

The preliminary experiment, conducted before the main corpus collection experiment, was designed to test the experimental setting and to address how to record eye movement and speech data accurately in a gameplay scenario. Following previous research [139, 140], the experiment was performed in a sound booth, in order to collect high quality speech recordings and eliminate any background noise. The eye tracker was operated at a 500 Hz sampling rate, to ensure high resolution of eye movement data.

The preliminary experiment gave a general view of players' speech and vocabulary. Players were found to engage with the game and speak as expected. Their speech styles provided insight into their natural interaction with the gaming environment. The experiment also identified problems in the recording settings. It was observed that player's movements during the game adversely affected the recording of eye movements. In the main experiment, the eye movement tracking settings were therefore changed to pupil-Corneal Reflection (CR) mode operating at a 250 Hz sampling rate, because the pupil-CR mode is more resilient to drift caused by head movements.

Video recordings of gameplay activities captured during the preliminary experiment were assessed subjectively for players' behaviour during the game. In most game sessions, the recordings clearly showed that players were trying to finish the games as soon as possible by taking a straight path to the finish line and ignoring scores. An inspection of time spent playing and of scores achieved indicated that the players did not enjoy playing. The players informally reported that they found the sound booth environment uncomfortable to stay in. To control the effect of the environment, the main corpus collection experiments were therefore conducted in a more relaxed environment with controlled artificial illumination measured at 250 lux, as assessed by the Extech HD45 Luxmeter. The control of ambient lighting levels ensures having sufficient amount of light in the room to be able to measure

the pupillary responses accurately[29]. A minimum acoustic isolation was ensured in order to avoid any background noise during speech recording. Figure 3.5 demonstrates the setup for the main experiment.



Figure 3.5: Main experimental setup.

3.4 Post-processing of recordings

3.4.1 Data synchronization

In a multimodal corpus collection experiment, synchronisation of data is necessary. The output of the eye recording software used in this research included each eye movement event (fixation, saccade, pupillary responses and blinking) with both start and end times. In addition, audio start and stop timestamps were marked in the eye movement tracking data output file as soon as the recordings began and ended. The captured video of on-screen activities was also timestamped by the eye recording software using the same scheme. The EyeLink II maintained latency in the synchronisation between the recording below 4 ms at

[29] In case the room is too dark then the pupil may become very dilated and be difficult to track.

the eye tracker sampling rate of 250 Hz. This allowed easy alignment of eye movement and speech data, based on the timestamps of host PC system clock.

While the eye movement and speech data were recorded on the host PC, the game was running on the display PC and its illumination values were recorded on that PC (see Figure 3.6). Hence, it was essential to synchronise the game recordings and illumination values with the multimodal recordings of eye movement and speech. Note that both the start and end times of games and the illumination values were synchronised using the system clock on the display PC. Moreover, the luxmeter recordings started before and ended after each game session. The synchronisation between both PCs was achieved by sending key press messages[30] to the host PC to mark the start and end of each game session and illumination recording, as illustrated in Figure 3.6, according to host PC system clock. Synchronisation was performed manually for each game session during experiment and confirmand using the captured video of on-screen activities, which showed both the host and display PCs' system clock time (see Figure 3.3).

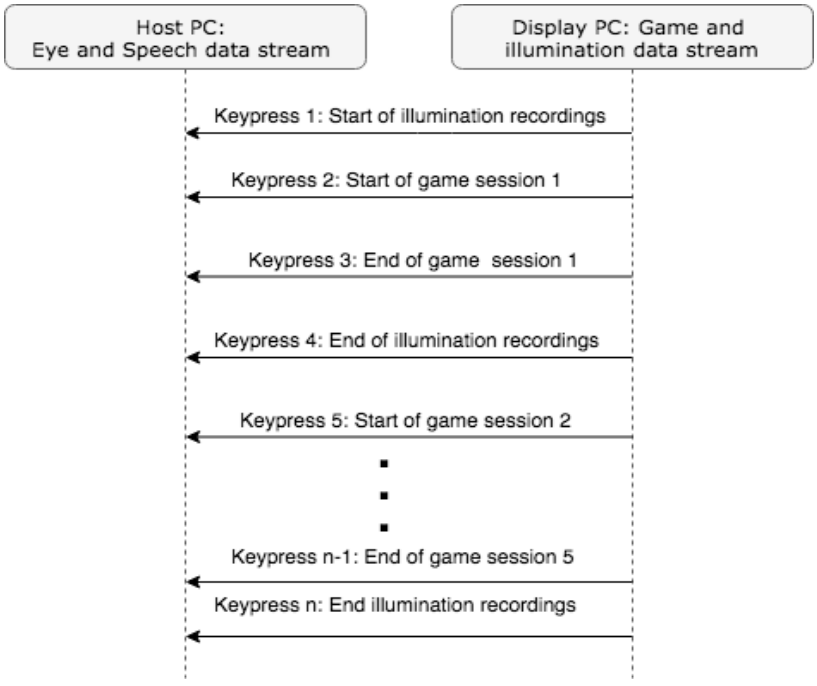


Figure 3.6: Scheme for synchronising recordings between PCs by sending keypress massages from the display PC to the host PC.

[30] A special key massage was coded for each event in the EyeLink Screen Recorder software.

3.4.2 Transcription of speech recordings

The audio recordings of players' speech were transcribed in two phases. First, manual transcription was performed by a transcriber and a reviewer, using Praat software[31]. Figure 3.7 show an example of the time-aligned transcription of a speech signal using Praat. The second phase used an ASR system implementing force alignment, in order to reduce human error in time alignment transcription. The ASR system used in this study was a general-purpose speech recognition system built using the Kaldi toolkit and trained on the LibriSpeech corpus [145]. Further details of the ASR system are provided in Chapter 6. All transcriptions from the first phase were output into plain text file format and transformed into a label file format recognisable by the Kaldi toolkit. Both manual transcription and ASR output were reviewed and validated by the reviewer.

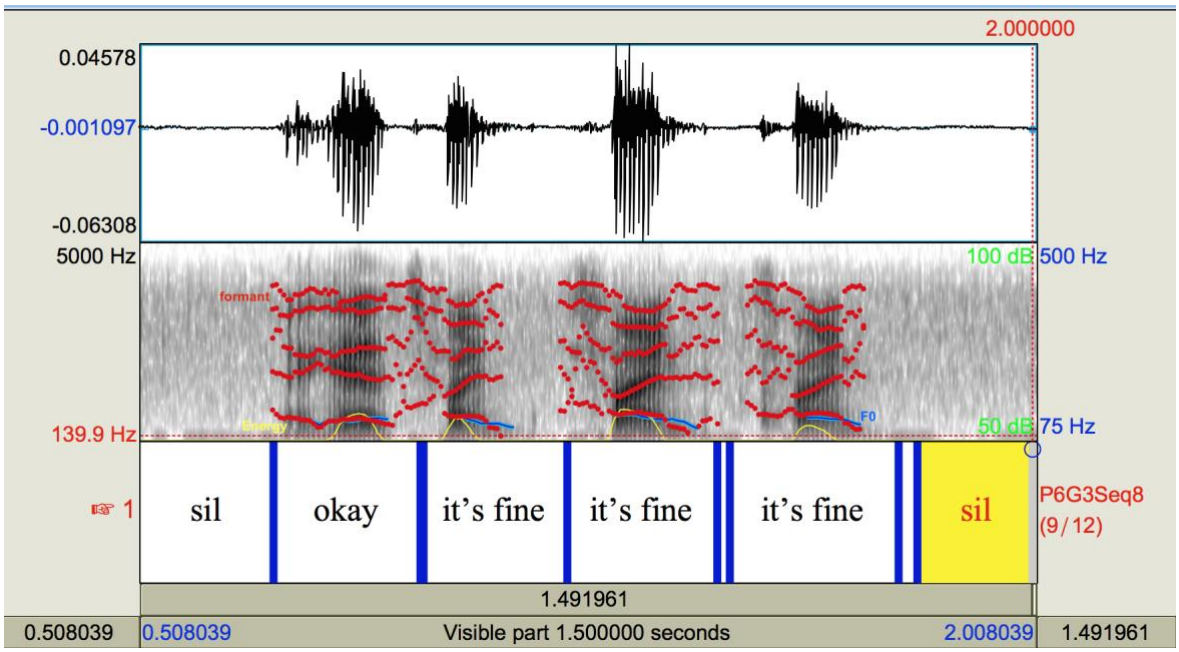


Figure 3.7: Example of time-aligned transcription (buttom) of a speech signal (top) using Praat software. The corresponding spectrum and acoustic featuers are shown in the middel.

3.4.3 Quality monitoring and assessment

The recording of eye movement data was heavily dependent on subjective settings, with the result that the quality of eye movement tracking data may have varied between players. The

[31] http://www.fon.hum.uva.nl/praat/

standard EyeLink II calibration procedure allowed the tracking of a player's eyes at nine different positions on the screen. The calibrated data was used as a reference to estimate the gaze position on the screen for any given eye fixating position. A validation procedure was performed after calibration to measure the accuracy of the calibrated data.

During calibration, the experimenter could reject the recording session if calibration results were poor, then recalibrate to ensure that calibration was adequate for the task. The experimenter was also able to monitor the eye movement data throughout the session and performed online drift correction before each game session if necessary. The recording of eye movement data using pupil-CR mode allowed any drift caused by the player's involuntary movements to be avoided. This minimised the need for drift correction which might have introduced error and caused the session to be rejected.

The measurement of pupillary responses is sensitive to changes in gaze position [146]. The pupil image was captured by the eye movement tracking cameras mounted on the player's head. When the player looked at the far corners of the screen, the pupil image would appear flattened rather than truly circular. The resulting error in pupil diameter measurements is referred to as 'gaze dependent' measurement error [146]. The 'gaze dependent' error is estimated at 10%, as pupil diameter is not calibrated during the EyeLink II standard calibration procedure [27]. To avoid this error, the EyeLink II user manual recommends that stimuli be placed near the center of the screen [27]. Eye movement data was inspected by playing the captured video in animation playback view,[32] which shows the player's fixation position (focus on the screen) overlaid on the game (Figure 3.3). It was found that the player's focus was mostly on the center of the screen, because the camera view changed with the location of the speedboat, keeping it at the center of the gaming environment. This can be clearly seen in Figure 3.8, which shows that the player's eye movement (fixation location) during a game session was at the center of the screen or surrounding area. Consequently, 'gaze dependent' measurement error could be discounted during data analysis, as the player's focus was at the middle of the screen or the surrounding area most of the time.

[32] Using SR Research Data Viewer software.

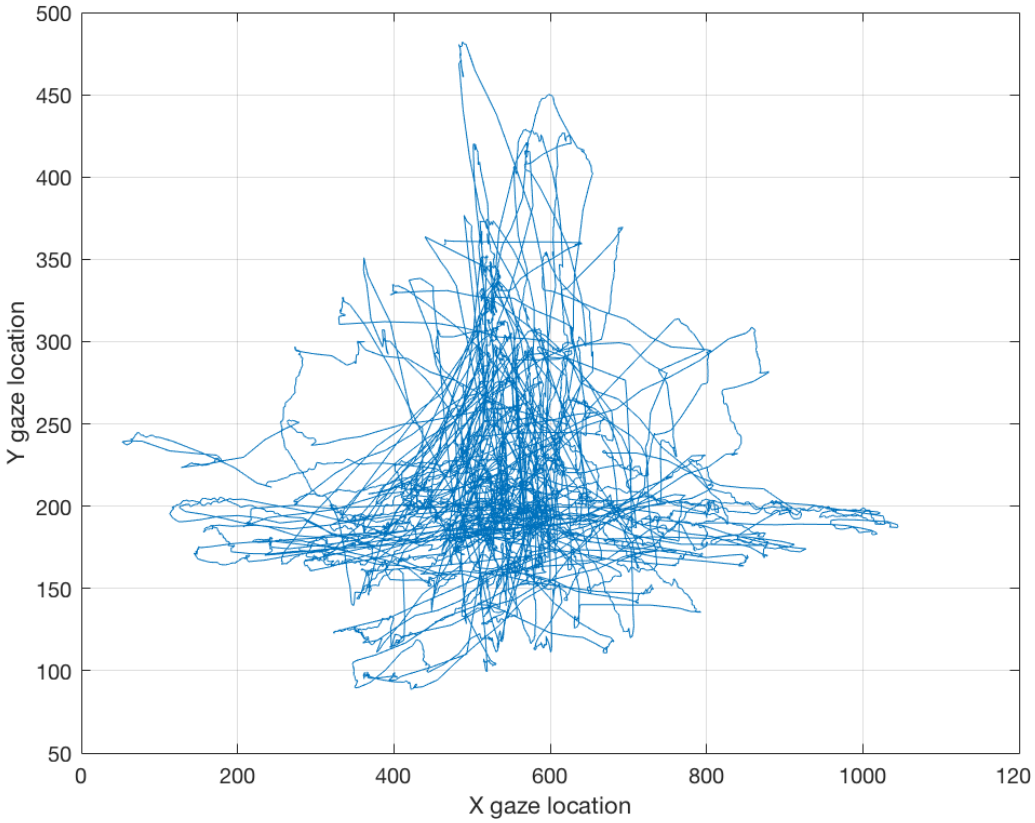


Figure 3.8: Plot of a player's X and Y fixation location during a game session over all captured video frames. The player's focus was mostly on the center of the screen and the surrounding area.

The quality of audio recordings was consistently good with respect to noise in the environment. The players' speech was recorded using a high-quality microphone able to isolate ambient noise (e.g. air conditioning). Because players were asked to speak out loud, they talked spontaneously throughout most of every game and their speech included hesitation and repetition. An example of a player's in-game monologue is shown in Figure 3.9.

The aGame corpus was created from the data collected from 14 of the 15 participants in the main experiment. The other player's data was discarded from the analysis because of technical errors and noisy signals (e.g. high blinking rate, signal loss), which is typical in eye movement tracking studies [9, 140]. This does not compromise the outcome of the corpus collection.

```
oh crap oh crap
crap crap crap crap crap crap
okay it is fine
oh you not know i can go backwards
and the haptic feedback is really useful
uh crap
okay it's fine it's fine it's fine dammit
okay it's fine i got it i got it
okay what's with all the torpedoes
okay okay i'm getting better at this
it's fine i got it
uh no no okay
oh my god where do these torpedoes <laugh> keep
coming from
uh crap uh no no no
```

Figure 3.9: Example of a player's speech in the aGame corpus.

3.5 Corpus size

A total of 2 (h) 55 (m) and 46 (sec) of raw eye movements and 1 (h) 51 (m) and 4 (sec) of raw audio data were recorded. Some recordings were discarded, as discussed in Section 3.4.3. In total, the aGame corpus contained data recorded in 70 sessions (5 games x 14 players), with an average duration of 2.5 minutes per session. The details of sessions recorded for each player are listed in Table 3.2.

Table 3.2: Details of sessions recorded per player in the aGame corpus.

Player	**Total recording duration (min)**	**% of speech**
1	14.9	56.4
2	13.4	82.6
3	10.4	70.2
4	11.8	53.1
5	13.8	77.7
6	10.6	49
7	13.2	82.8
8	15.0	43.7
9	13.0	68.7
10	13.4	66.1
11	12.3	47.4
12	11.7	64.9
13	9.8	50.7
14	12.7	67.5
Mean	12.6	62.9
Standard deviation	1.6	13.1

3.6 Summary

This chapter has described the experimental design and the procedure applied in the collection of the aGame affective eye movement and speech corpus. The task utilised a virtual affective gaming environment which allowed the collection of quasi-natural affective responses. Players were instructed to speak out loud during game sessions and their speech was recorded.

Recording a large amount of data from each player required extended session duration, which meant that players had to wear the eye tracker for long periods. To reduce consequent fatigue and discomfort, a reduced set of five games was selected (from [10]) so that each experiment session lasted no more than twenty minutes. The games elicit different affective states denoted by the three levels of the arousal and valence dimensions of Russell's circumplex model of affect.

A preliminary experiment was performed in order to test the apparatus and settings and to give a rough outline of players' behaviour. The main corpus collection experiment was conducted in a more relaxed environment. All data was subjected to quality assessment. Some sessions were rejected because of poor quality and technical errors, which is normal in eye movement tracking studies. Eye movements and speech data captured in the main experiment was analysed and used in the development of an affect recognition system, applicable for real-life HCI application such as speech-enabled interfaces, as described in the following chapters.

Chapter 4

Affect Elicitation Analysis and Detection from Eye Movements

Chapter 3 described the development of the aGame corpus. This chapter concerns the use of the collected eye movement signals to detect the affective experience elicited while players interacted with the virtual gaming environment.

Section 4.2 analyses the players' self-reported affective experience to investigate the affective states elicited. Section 4.3, sets out the detection methodology adopted throughout this thesis, including the machine learning algorithm, the feature selection method, the validation strategy for performance generalisation and the evaluation measures used when reporting the results. Section 4.4, explains the development of the affect detection system based on eye movement signals and discusses the results. New pupillary response features based on the Hilbert transform are introduced to improve affect detection performance. Section 4.5 compares the results obtained with start-of-the-art systems and the chapter concludes with a summary in Section 4.6.

4.1 Motivation

The affective computing community has paid considerable attention in the past decade to detecting individuals' affective states from physiological signals [3]. Most existing contributions have employed invasive sensing methods (e.g. EEG), which can be inappropriate for natural HCI applications because of low accuracy, the inability to use

remote sensing, in contrast to eye movements (Section 2.2.3). As mentioned in Chapter 2, eye movement signals have received considerable attention recently because they can be observed naturally (with remote cameras) and provide rich behavioural information. However, few researchers have narrowly attempted to detect affects by measuring individuals' eye movements, focusing on passive affect elicitation methods (e.g. [9]) in which the results may not be possible to generalise to real-life HCI applications (e.g. interactive gaming). This work reports the use of eye movement signals, collected in a more natural setting, while players played in the interactive virtual gaming environment, to develop an affect detection system that can be applied to real-life HCI application.

4.2 Affective Experiences Analysis

To develop an affect detection system capable of inferring players' affective experience from data in the aGame corpus, ground truth labels had to be established. These can be defined based on external expert observations/annotations. This approach is common in affective speech corpora (e.g. [69]). However, in this case, affects are defined based on how they are perceived by others rather than the felt affective states by individuals. Although the interactive virtual environment would be expected to evoke the affective experiences as defined and evaluated by [10] (see Table 3.1), one may assume that the ground truth in affective labelling would accord with the felt affective experience of those who had played the games. Therefore, players' self-reported gaming experiences were analysed to determine whether the interactive virtual environment had elicited the expected affective states as defined in [10], prior to their use as ground truth labels in the detection task and further analysis.

The pairwise Cohen's kappa analysis [147] was first applied in order to measure the amount of agreement among the players on their self-reporting experience. The results reveal that on average, players agreed fairly well on their felt affective experiences reported on a seven-point scale (range [–3 - 3]), with $mean(k) = 0.22 \pm 0.04$ and $mean(k) = 0.41 \pm 0.05$ on the arousal and valence dimensions respectively (refer to appendix C for the questionnaire). There was stronger agreement on valence reporting than arousal, a possible explanation being that players may have found it easier to interpret the valence dimension in comparison to the arousal dimension when reporting their experience [46], which is

consistent with previous findings [9, 13]. With regard to affective keyword-based reporting, weak agreement among players was observed on the eight keywords (see Section 3.3.3) with $mean(k) = 0.12 \pm 0.04$. There was less agreement between players on affective keywords than on arousal and valence reporting, possibly because interpretations of the keywords differed among the players [46]. This result suggests that when developing a player-independent affect detection system, it would be more reliable to represent the players' affective experience using arousal and valence dimensions rather than affective keywords. The correlation between the expected affective experience, the ground truth labels (see Section 3.2, Table 3.1), for arousal and valence and the players' self-reporting was computed in order to assess their agreement. Players' self-reported experience was found to be highly correlated with the ground truth labels on both dimensions, as Table 4.1 shows. Indicating that the affective gaming environment successfully elicited the expected affective experiences from the players, confirming the ground truth labels described in Table 3.1, despite the different experimental setting compared to [10].

Table 4.1: Signficant correlation between players' self-reporting and the expected affective experiences (the ground truth labels as defined in [10]).

	Correlation $\rho(70)$	***p-value***
Arousal	0.567	< 0.05
Valence	0.602	< 0.05

From the results, it can be concluded that the players' self-reporting was more accurate on both arousal and valence compared to keyword reporting, as the analysis shows better agreement between players' affective experience on both dimensions and close correlation with the ground truth (Table 4.1). Hence, it was decided to use the ground truth labels (Table 3.1) according to [10] when analysing the aGame corpus data and developing the affect detection systems in this study. The two affective dimensions were treated separately, as they are assumed to be independent [5]. Hence, the affective games were grouped into three states on each dimension: low (LA), neutral (N) and high arousal (HA); positive (PV), neutral (N) and negative valence (NV).

4.3 Detection method

The affect detection system (ADS) developed in this this work has two stages, as illustrated in Figure 4.1. The front-end stage involves pre-processing the input signal and transforming it into a feature vector containing appropriate affect-related features. At the second stage, a classification technique is applied in the feature space to detect the underlying affective state of the individual. It involves selection of the feature space, optimising the classification algorithm's parameters during training and evaluating system performance during testing, using 'leave one player out' cross-validation (which is described in Section 4.3.4). Several classification techniques have been used in affect detection studies (see Chapter 2). For the ADSs developed in this study, a support vector machine (SVM) was applied to detect the affective states experienced by players. The reason for using an SVM is that they have been shown to provide good performance in several studies (e.g. [9, 14, 101]) and have proven to yield satisfactory performance for a relatively small dataset [107]. They are also suitable when dealing with high dimensional data [148]. The following subsections detail the detection method including: the SVM algorithm, feature selection, system validation and performance evaluation methods.

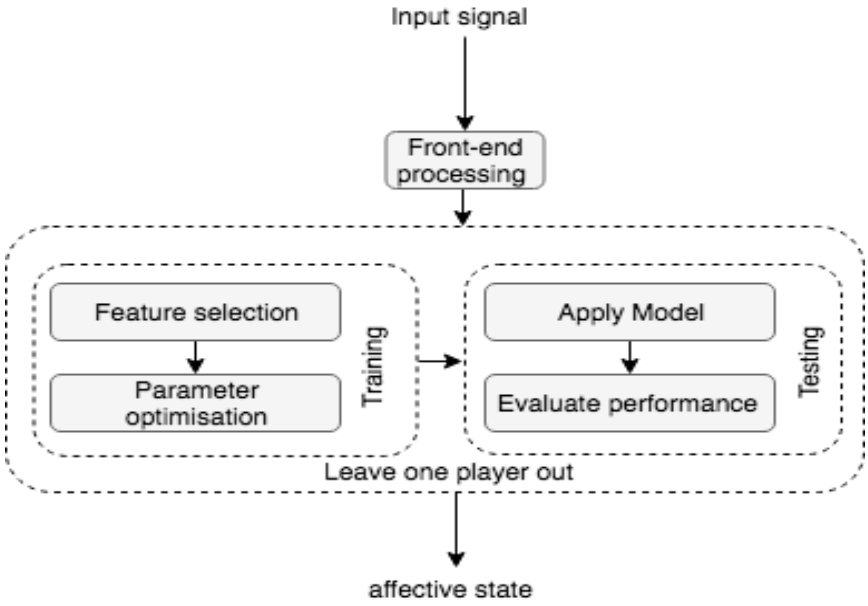
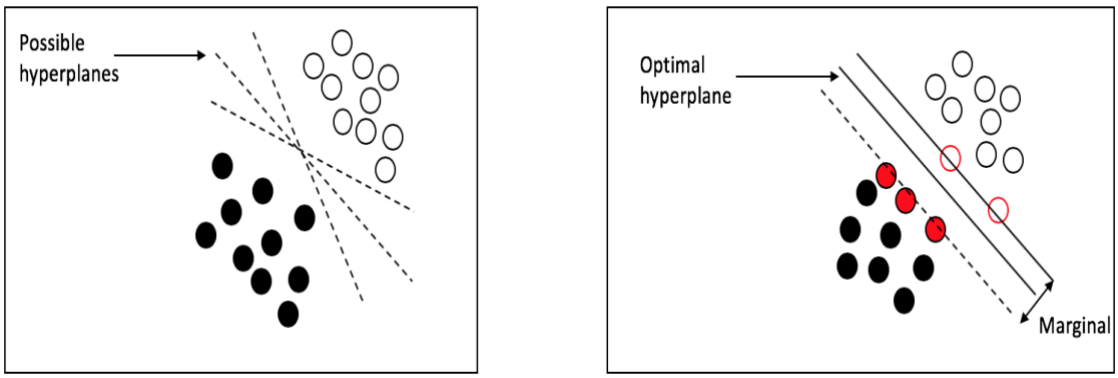


Figure 4.1: An overview of the framework for the affect detection process used throughout this study.

4.3.1 Support vector machine (SVM)

An SVM is a supervised discriminative machine learning algorithm in which the decisions are made by defining a separating hyperplane based on the characteristics of the observed data. This type of algorithm was introduced in 1995 by Cortes and Vapnik [149], since when SVMs have become commonly used in machine learning for their computational efficiency and good generalisation performance [150]. They were mainly designed for linearly separable binary class problems, but can also be used efficiently for linearly inseparable problems, using a 'kernel trick', which is described below.



(a) (b)

Figure 4.2: Two-dimensional binary classes, linearly separable; (a) shows possible separating hyperplanes and (b) shows the optimal separating hyperplane. Red circles indicate support vectors for the two classes.

Linear SVM

Given a dataset with *N* labelled *D*-dimensional training vectors $T = (\{x_1, y_1\}, \ldots, \{x_n, y_n\})$ corresponding to class labels $y_i \in \{-1, +1\}$, where $x_i \in \{R^D\}$, the goal of the SVM is to find the optimal hyperplane that separates the positive $(y_i = +1)$ from the negative $(y_i = -1)$ classes. Figure 4.2(a) illustrates an example of two-dimensional binary class data. Clearly, there are many possible ways to separate the two classes linearly. The optimal separating hyperplane (the so-called decision boundary) is one that maximises the margin around itself, as shown in Figure 4.2(b), which satisfies Equation 4.1. In other words, the optimal hyperplane has maximum distance from its support vectors and can separate the binary classes without errors, thus yielding a good generalisation of the observed data [107, 150].

$$w.x_i + b = 0 \tag{4.1}$$

where w is a vector normal to the hyperplane and b is the bias; their values are determined from the observed training data. For the linearly separable case, the optimal separating hyperplane with maximum margin is found when the training data satisfies the constraints in Equations 4.2 and 4.3. The complete SVM constraint can be defined by combining the two equations into one set of inequalities, as given by Equation 4.4 [150].

$$w.x_i + b \geq 1, \ \forall_i \in y_i = +1 \tag{4.2}$$

$$w.x_i + b \leq -1, \ \forall_i \in y_i = -1 \tag{4.3}$$

$$y_i(w.x_i + b) \geq 1, \forall_i \tag{4.4}$$

For points where the equalities in Equations 4.2 and 4.3 hold, this defines the support vectors for both the negative and positive classes corresponding to the lines parallel to the separating hyperplane. Their perpendicular distance from the separating hyperplane with the normal w is $\frac{1}{\|w\|}$. Therefore, the margin is equal to $\frac{2}{\|w\|}$ (the distance between the support vectors of the binary classes). Note that the two lines share the same parameters (w, b), as they are parallel. Thus, the parallel lines that maximise the margin can be found simply by minimising $\|w\|$ as given by Equation 4.5, subject to the constraint given by Equation 4.4.

$$min_w \frac{1}{2} \|w\|^2 \ \ subject\ to\ y_i(w.x_i + b) \geq 1, \forall_i \tag{4.5}$$

The SVM problem is a convex quadratic optimisation problem that can be solved by introducing a Lagrangian multiplier α_i to represent the constraints in Equation 4.4 [150]. Hence, the primal SVM formulation becomes that given by Equation 4.6, where N is the number of samples.

$$L_p = \frac{1}{2} \|w\|^2 - \sum_{i=1}^{N} \alpha_i (y_i (w.x_i + b) - 1), \quad \alpha_i \geq 0 \tag{4.6}$$

The goal is to minimise L_p with respect to parameters w and b and to maximise subject to $\alpha_i \geq 0$. This can be solved by making the derivative of L_p with respect to both w and b equal to zero, separately. This yields the following:

$$w = \sum_{i=1}^{N} \alpha_i y_i x_i \, , \; \sum_{i=1}^{N} \alpha_i y_i = 0 \tag{4.7}$$

The primal SVM formulation in Equation 4.6 can be written in the form of the Lagrangian dual L_D by substituting w as follows:

$$L_D = - \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j (x_{i,} x_j) \, \alpha_i \geq 0 \text{ and } \sum_{i=1}^{N} \alpha_i y_i = 0 \tag{4.8}$$

In case the observed data cannot be perfectly separated without errors, an SVM can allow a hyperplane to be found that separates the negative and positive classes with some misclassification errors ξ_i (slack variable) by introducing a penalty factor C, the so-called cost parameter. In this variant, referred to as a 'soft-margin' SVM, the cost C balances the tradeoff between minimising classification errors in the training data and maximising the hyperplane margin. The optimisation problem becomes:

$$min_w \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{N} \xi_i \tag{4.9}$$

$$subject\ to\ y_i (w.x_i + b) \geq 1 - \xi_i, \ where\ \xi_i \geq 0, \forall_i$$

Non-linear SVM

An SVM can be used in case of a linearly inseparable feature space, by applying a 'kernel trick' to map the original inseparable data into a high dimensional space in which the data points can be separated linearly. This is done using a kernel function which transforms the feature space by adding a new dimension to the data that allows a linear SVM (described

above) to be used in the new high dimensional space. The 'trick' in the kernel function implicitly computes the transformation by the dot product of the original feature space [150, 151]. The most common types of kernel functions are linear, sigmoid, polynomial and radial basis function (RBF). The choice of the kernel function depends mainly on the nature of the training data. The RBF kernel is typically considered a wise first choice [152]. It tends to perform like a linear or sigmoid kernel by simply adjusting the cost and kernel parameters [153, 154]. It also takes less parameters than the polynomial kernel, thus less computational complexity [152]. Moreover, it is good at handling the nonlinear relation between the features and the classes by nonlinear mapping [152]. Thus, the RBF kernel was selected for the detection process. The RBF kernel function is given by Equation 4.10, where γ is the kernel parameter (determined empirically).

$$k(x_i, x_j) = e^{-\gamma \|x_i - x_j\|^2} \tag{4.10}$$

In order to achieve the best performance using an SVM with RBF kernal, the model parameters (C , γ) must be optimised. This is typically done by performing a grid search within a selected range to find the optimal combination of C and γ that yields the best detection performance. In [152], it is suggested that the best parameters can be identified by performing the grid search on an exponentially growing range (e.g. [$2^{-15}, 2^{15}$]).

Decision function

Once the SVM parameters have been learnt from the training data the decision about a new observation x can be made by applying the decision function $f(x)$ defined by Equation 4.11, where x belongs to the positive class if $f(x)$ is positive and to the negative class if $f(x)$ is negative.

$$f(x) = \sum_{i=1}^{N} \alpha_i y_i k(x, x_i) + b \tag{4.11}$$

where $k(x, x_i)$ is the kernel function. Further details about SVMs can be found in [150].

Multiclass problems

Although SVMs were developed for binary class problems, their use can be effectively extended to multiclass problems. Several approaches have been proposed to solve such problems by dividing them into a series of binary class problems. The approaches most widely taken are one-against-all and one-against-one. The first of these to be applied was one-against-all [155]. For *c* classes, *c* binary SVMs are constructed, one for each class, in which positive labels are assigned to the target class training data and negative labels to the data belonging to the remaining classes. The decision on a new sample is made according to the highest SVM output score. In one-against-one [156], *c(c-1)/2* binary SVMs are constructed for all possible binary classes. A new sample is classified by the 'max-wins' voting strategy, where the sample is assigned to the class that receives the highest number of votes. It is difficult to decide which is the best approach to use. The disadvantage of one-against-all is its weakness in dealing with unbalanced training data. One-against-one can solves this problem, but requires training with more binary SVMs. Hus and Lin [155] show that the one-against-one approach is more practical and easier to use; therefore, it was selected for use in this study, considering the relatively small number of classes involved.

4.3.2 Feature selection

An important step in the development of an ADS is to select the features that can most effectively describe the underlying affective state of an individual and provide the greatest accuracy. Using a large number of features may reduce system performance, especially for relatively small corpora [157]. This is known as the 'curse of dimensionality'[33] problem [158]. In addition, not all features provide information useful to the system and some may be redundant. Incorporation of non-useful and redundant features can degrade the detection performance. Therefore, it is important to select the most suitable features which can provide the best performance.

Two types of feature selection methods have been used widely in affect detection systems: wrappers and filters [100]. In both types, the goal is to reduce the feature space by searching for a subset of candidate features based upon an evaluation function which tests

[33]As the dimensionality of the feature space increases, the samples become sparser. Thus, the required number of samples increases exponentially in order to describe that feature space.

their ability to discriminate between the classes. In wrapper-based selection, the performance of the detection algorithm is used as an evaluation function to select the candidate features. The disadvantage of this method is that it is computationally expensive and highly dependent on the detection algorithm used for evaluating these features. On the other hand, in filter-based selection, the candidate features are independently evaluated according to criteria obtained from the characteristics of the observed data. Using these evaluation criteria, the features are ranked by their performance in descending order. Finally, the best performing features above a predefined threshold are selected for use by the detection algorithm. This method is computationally inexpensive and performs well in practice. A review of different feature selection methods can be found in [100].

In order to identify the most suitable features for the ADS in this study, the researcher took a filter-based approach, using the Fisher discriminant ratio (FDR) [159] as evaluation criterion. Assuming a feature set $F = \{f_1, f_2, \ldots \ldots . f_n\}$ of *D* dimensions, the aim of the FDR is to find the most discriminating features $F_d \subset F$ such that $d < D$, where between-class variance is maximised and within-class variance is minimised. This is done by computing the ratio of between-class variance to within-class variance, as given by Equation 4.12, then the features are ranked by their FDR values and the highest ranking features are selected as being the most discriminative for the task. Hence, eliminating the irrelevant features. The FDR method is used mainly for binary class problems; however, it can be extended to multiclass problems in which the total FDR value for a given feature is divided by the number of comparisons made between binary classes.

$$RDF(f) = \frac{|\mu_1 - \mu_2|^2}{\sigma_1 + \sigma_2} \tag{4.12}$$

where μ_i and σ_i are the mean and standard deviation of the *i*th class for feature *f.*

4.3.3 Feature normalisation

Affective responses from physiological signals can vary between individuals [133]. In [160], it is argued that changes in affective response are due to changes in context specificity. In this study, the responses were expected to be stable across the players, as they had been

elicited in the same context, the virtual gaming environment. However, the extracted features $F = \{f_1, f_2, \ldots\ldots f_n\}$ were subjected to maximum-minimum normalisation, as given by Equation 4.13, in order to reduce the variability further and make the measures comparable across games and players. Normalisation was applied separately to each player's features.

$$f_{norm} = \frac{f - f_{min}}{f_{max} - f_{min}} \tag{4.13}$$

where f_{norm} is the normalised feature vector and f_{min} and f_{max} are the minimum and maximum values of a given feature f.

4.3.4 Validation methods

The main goal in developing an ADS is to detect the target affective state of an individual with the best performance that can be generalised to unknown data. The generalised performance on the aGame corpus can be obtained by dividing the data into two partitions: training and testing. The ADS is developed using the training partition of the corpus and tested on the testing partition, thus allowing the system to be tested on unseen test data and thereby to attain generalised performance. If the partitioning are done randomly, however, there would be a high chance that the training partition contained data from the testing partition and vice versa, resulting in overfitting and an invalid performance. This can be avoided by applying cross-validation, whose goal is to gain insight into the system's performance in practice. The validation methods most commonly employed to evaluate the performance of affect detection systems are *k*-fold and 'leave one subject out' cross-validations (e.g. [9, 14]).

***K*-fold cross-validation**

In *k*-fold cross-validation, the data is divided randomly into *k* non-overlapping partitions (folds), each of which contains the same proportion of samples from each class as in the original dataset. Supposing the size of the dataset is *n*, then each fold will have $\frac{n}{k}$ samples. During the development process the ADS is trained and evaluated *k* times. For each fold, the

system is trained on *k-1* folds of the dataset and evaluated on the remaining fold. The overall system performance is then computed as the average over the *k* folds.

'Leave one subject out' cross-validation

Partitioning the data randomly into *k* folds does not guarantee subject independence. Thus, it is likely that data from the same subject will appear in both training and testing partitions. In real-life applications, however, the system typically has no information about the test subject. In order to guarantee more realistic performance, the ADS should be evaluated in a subject-independent manner. In the 'leave one subject out' method, the data generated from each subject is pooled together as a fold. During development, one subject's data is left out for testing and data from the remaining subjects are used for training the system. This process is repeated for each subject. The overall system performance is then computed as the average over all subjects. To obtain generalised performance on the aGame corpus the 'leave one subject out' cross-validation method was applied and is hereinafter referred to as 'leave one player out'. Such that the data divided into fourteen-fold, one for each player, each containing five sessions (see Section 3.5). Accordingly, the ADSs are trained fourteen times, with a total of 65 sessions available for training and 5 sessions for testing in each iteration.

4.3.5 Performance evaluation metrics

A traditional evaluation metric that has been widely used to evaluate detection systems is detection accuracy, also called weighted average, which indicates how well the detection system is able to make correct decisions about the test samples. It is defined as the number of correctly detected samples over all samples in the test set, as given by Equation 4.14.

$$Accuracy = \frac{1}{N}\sum_{c=1}^{C} n_{cc} \tag{4.14}$$

where C is the number of classes, N is the size of the test set and n_{cc} denotes the diagonal elements of the confusion matrix as shown in Figure 4.3, indicating the number of correctly detected samples as belonging to class C. This measure of accuracy is highly dependent on the class distribution in the training data (prior probability $p_c = \frac{N_c}{N}$). In case of unbalanced

class distribution, the system's decisions will always be biased towards the class that has the maximum prior probability p_c, regardless of any other information provided. For example, if a system is trained on data of which 90% belongs to one class, all the test data belonging to that class will be detected correctly, achieving 90% accuracy. Hence, the accuracy measure in Equation 4.14 may provide a misleading evaluation of performance, especially for highly unbalanced data. Two measures widely used to represent true system performance on an unbalanced class distribution are recall and precision [161]. They provide measures of a specific class as follows:

$$Recall_c = \frac{n_{cc}}{\sum_{j=1}^{C} n_{cj}} \tag{4.15}$$

$$\text{Precision}_c = \frac{n_{cc}}{\sum_{j=1}^{C} n_{jc}} \tag{4.16}$$

Recall is the fraction of correctly detecting test samples of class *c*, while precision refers to the fraction of the total number of test samples detected as belonging to class *c* that actually do belong to class *c*. The unweighted average accuracy is another useful measure for dealing with unbalanced data, providing a more realistic representation of the system's performance. It combines detection accuracy, recall and precision. Hence, in this thesis the ADS system performance results are reported in terms of unweighted detection accuracy, as defined in Equation 4.17, hereinafter referred to as detection accuracy.

$$\text{Unweighted accuracy } = \frac{1}{c}\sum_{c=1}^{C} n_{cc}/N_c \tag{4.17}$$

Detected class

True class	C_1	C_2	C_3	.	.	C_c	Σ
C_1	n_{11}	n_{12}	n_{13}	.	.	n_{1c}	N_1
C_2	n_{21}	n_{22}	n_{23}	.	.	.	N_1
C_3	n_{31}	n_{32}	n_{33}	.	.	.	N_1
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
C_c	n_{c1}	n_{c2}	n_{c3}	.	.	n_{cc}	N_c
Σ							N

Figure 4.3: An illustration of a confusion matrix.

4.4 Development of an affect detection system based on eye movements

This section discusses the development of an affect detection system using eye movement signals, namely pupillary response, fixation, saccade and blinking, collected in the interactive virtual gaming environment (see Chapter 3). Figure 4.4, depicts the front-end processing of eye movement signals prior to the affect detection, which involves preprocessing the pupillary response signal and applying Hilbert transform, and the features extraction step. The following subsections describe the pupillary response preprocessing procedure, explain the features extracted from eye movements and present the affect elicitation and detection results.

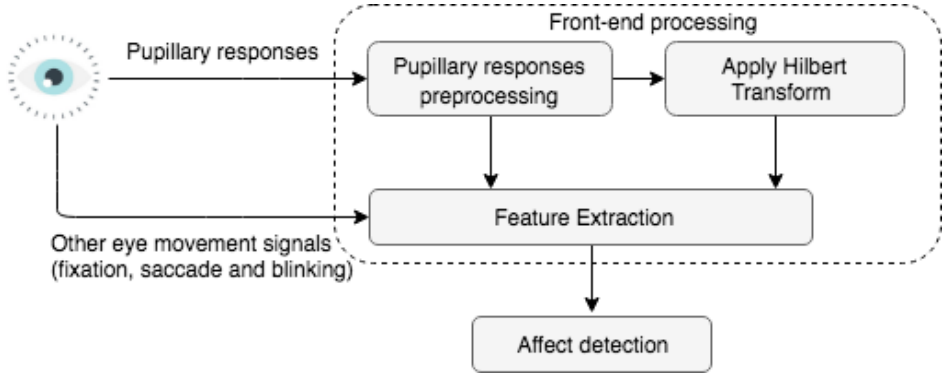


Figure 4.4: Overview of the front-end processing of eye movement signals.

4.4.1 Preprocessing of pupillary response signal

Variations in the pupillary response signal reflect the activity of the ANS [119]. One type of ANS activity is affective processes [118]. However, these variations are dominated by the response to environmental changes in luminosity, the so-called pupillary light reflex (PLR) [162], where the pupils respond to brighter and darker stimuli by rapid constriction and slow dilation respectively, under the control of the ANS [119]. The magnitude of the PLR is sufficient to conceal the influence of affective processes. The measurement of pupillary responses is also influenced by blinking events and noise. In most eye movement tracking systems, pupillary responses are measured by processing images of the pupil [19]. During blinking events the eyes are closed, albeit momentarily, preventing the image of the pupil from being captured and resulting in a loss of pupillary response information. It is necessary

to isolate and filter out all of the above artefacts unrelated to the affective processes of interest, prior to further processing of the signal for use in affect detection. Figure 4.5, illustrates the procedure followed in this study to preprocess the pupillary response signal, including unit conversation, blinking recovery, PLR removal and noise filtering. The following details each step.

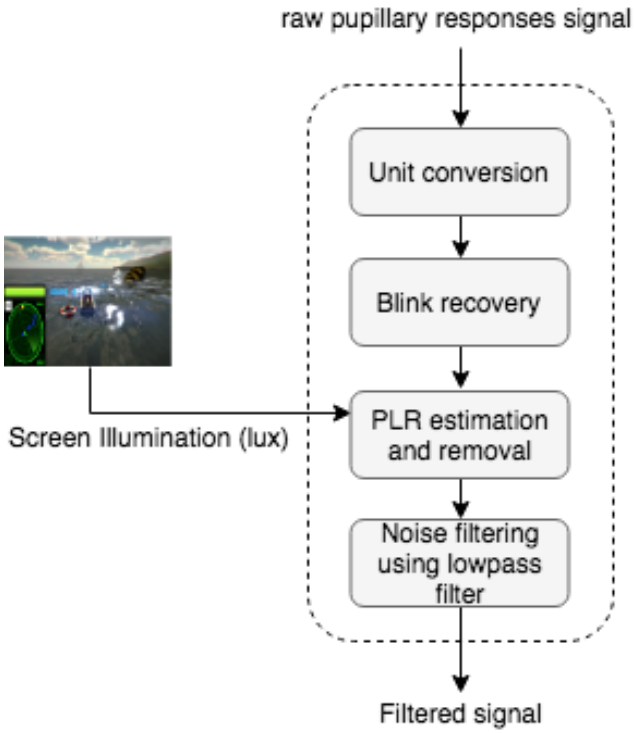


Figure 4.5: Overview of the procedure for preprocessing pupillary response signals. Screen illumination is measured continuously during the gameplay session using a luxmeter to compensate for PLR.

Unit conversion

The EyeLink II sensor measures pupillary responses in arbitrary units. These had to be converted to millimetres in order to compensate for PLR, as described later in this section. This linear conversion from arbitrary units to millimetres was performed as given in Equation 4.18.

$$PD_{mm} = PD_{Acutal}/f \tag{4.18}$$

where f is a conversation factor derived from the average pupil diameter estimated from the five-second benchmark recordings of an artificial pupil[34] divided by 5 mm, PD_{Acutal} is the pupil diameter of the actual eyes in arbitrary units recorded during the experiment and PD_{mm} is the diameter of the actual pupil in millimetres. Figure 4.6, shows an example of pupillary responses before and after unit conversion to millimetres. It can be seen that the unit conversion was successful as the arbitrary range of the artificial pupil is converted into 5 mm, although some variations in the recordings of the artificial pupil were observed due to noise. According to Eyelink II user manual [27], this noise is limited to 0.015 mm for 5 mm pupil. Therefore, it can be discarded, because the benchmark recording was averaged over five seconds to derive the conversation factor.

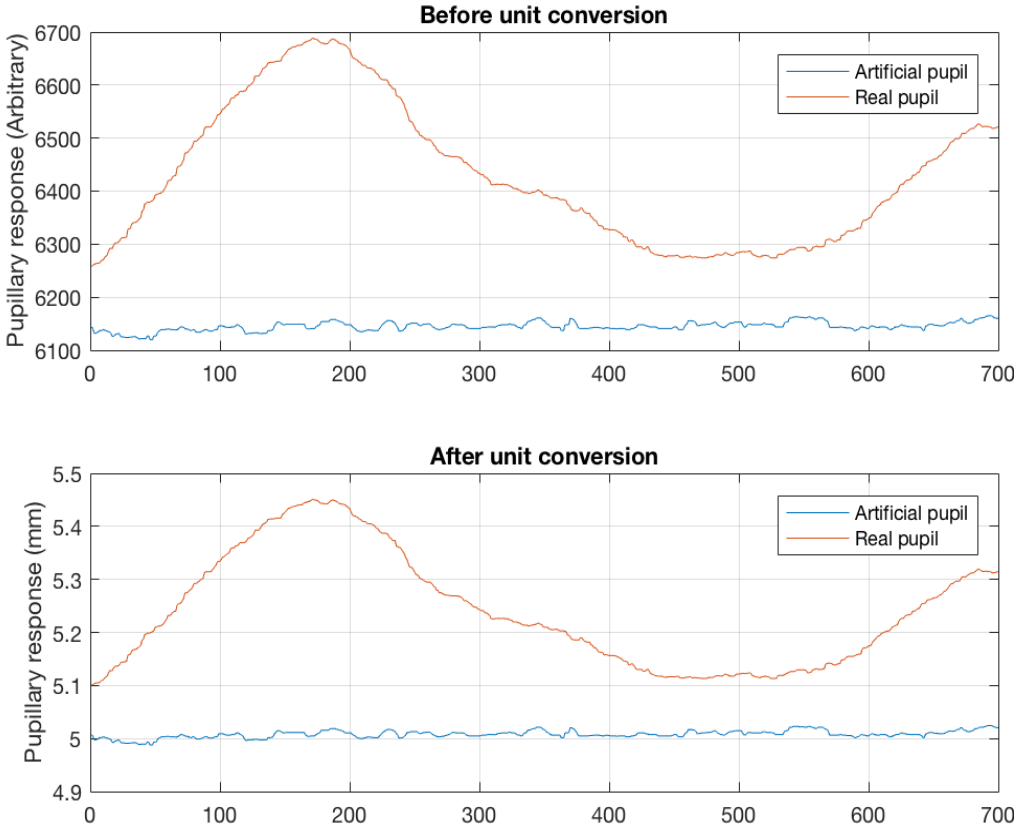


Figure 4.6: Pupillary response unit conversion using an artificial pupil.

[34]The artificial pupil is a printed black dot of 5 mm diameter that was placed close to the player's eye at the end of the experiment, so that it was as close as possible to the same depth as the player's eye.

Pupillary responses blink recovery procedure

Any blinking events in the pupil response signal were also treated as artefacts, liable to significantly impair the accuracy of measurement of variations in the signal in both the time and frequency domains [163, 164]. Once eye blinking events are detected, several methods for recovering pupillary responses can be applied, such as support vector regression, multi-layer perceptron, singular spectrum analysis, moving average and linear/cubic interpolation [163-167]. Cubic interpolation was employed in this work because of its simplicity and the nature of pupillary responses.

To recover pupillary responses during blinking, missing samples must first be identified. This information was provided by the EyeLink II online parsing algorithm [27]. The blinking onset is detected when the pupil constricts rapidly and the offset is detected when the eye reopens rapidly after a blink [165]. Since the algorithm detects blinking events in real time by calculating the velocity and acceleration of eye movements over a short time period (typically 8 ms), pupillary responses around blinking events had to be re-estimated in order to eliminate any artefacts during the recovery of the signal. To do this, once blinking events had been identified, their onsets and offsets were shifted forwards and backwards respectively, using a time window of 20 ms (selected empirically, based on visual inspection). Finally, the corresponding parts were recovered by cubic interpolation. Figure 4.7, illustrates an example of the recovery of pupillary responses during blinking. It can be seen that missing responses were recovered successfully.

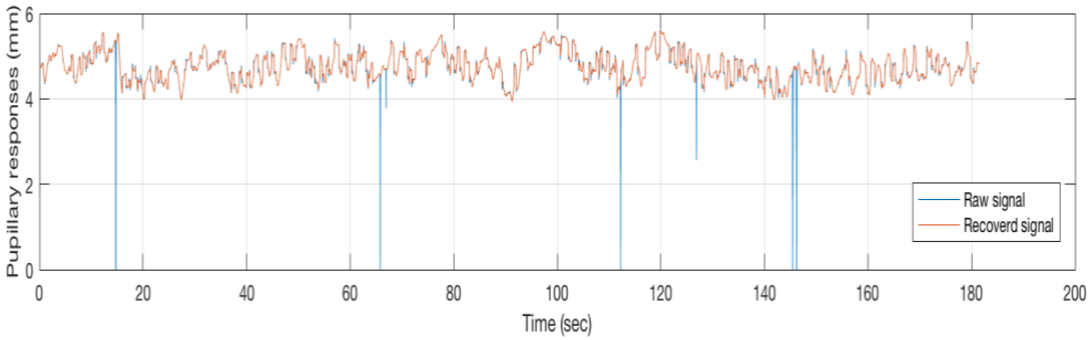


Figure 4.7: An illustration of pupillary response blink recovery. Missing samples due to blinking (blue) were recovered successfully using cubic interpolation.

Pupillary responses and illumination effect in a dynamic environment

It has been shown that task-evoked pupillary responses remain the same even if the illumination level of the stimulus changes [168, 169]. Therefore, the light reflex caused by changes in the level of illumination does not affect the task-evoked pupillary responses and can be isolated. This isolation is easy in static environments where there is prior knowledge of the illumination level. In this case, the pupil diameter can be calibrated to the illumination level of the stimulus, then subtracted from the overall pupillary responses measured during the experimental task. The remaining changes in pupillary response must now be purely task evoked [169], so pupillary responses are comparable across tasks. However, in dynamic environments such as gaming, the illumination level can vary and the effect can differ between players, making it difficult to use available signal source separation techniques such as principal component analysis to remove the influence of PLR (e.g. [9, 170]). Pamplona et al. [171] propose a physiologically-based model that extends the existing theoretical and experimental PLR models presented in [172, 173], to provide real-time estimation of the PLR and account for inter-subject variability. This model was adapted to estimate pupillary responses as a function of illumination level.

In order to estimate the light reflex using the model of Pamplona et al. [171], two input parameters are required: the average pupil diameter (mm) of the first second at the beginning of the experiment and the measured illumination level (lux) during the experiment. After applying the PLR model, the output light reflex was subtracted from the original pupil responses and the remaining changes were the result of affect elicitation plus noise. The noise was filtered using a Butterworth low-pass filter with a 4 Hz cut-off frequency, since the frequency of pupillary responses is approximately 1-2 Hz [167, 174].

4.4.2 Eye movements features extraction

The selection of eye movement features for the affect detection task is typically empirical. As described in Section 2.2.3, the evidence of correlation of affects with pupillary response, fixation, saccade and blinking-related measures has been presented in the literature. The majority of affective research on eye movements have used pupillary response features to detect affects (e.g. [9, 14, 131]), because of their well-established relation with the arousal

dimension [61, 63, 122]. Pupillary response has also been shown to correlate with the valence dimension [124]. Fixation is typically influenced by the content of interest in the visual field, while saccade comprises rapid eye movements closely associated with fixation. Blinking is a physiological response which has been linked to negative affects, anxiety and arousal [125-127]. These eye movement have also been shown to be useful for detecting affects (e.g. [9, 130]). In this study, features of the four eye movement signals were extracted and used in the affect detection task.

The conventional features of pupillary response in the time and frequency domains whose use has been widely reported in the literature were extracted to form the pupillary response feature vector. These features are mean, standard deviation, mean of derivative, spectral entropy and the power spectral density (PSD) features of four different frequency bands: 0.0-0.2 Hz, 0.2-0.4 Hz, 0.4-0.6 Hz and 0.6-1 Hz [9]. The PSD features were extracted using the Discrete Fourier Transform (DFT) [175], which estimates the short-time PSD. The power spectral features up to 0.4 Hz reflect the hippus effect, an involuntary variation in pupil diameter in the frequency range between 0.05 Hz and 0.3 Hz [171], which may appear during relaxation and resolve during mental activity (e.g. response to arousing stimuli) [9, 119].

Conventional pupillary response features are often insufficiently informative and may not provide accurate time and frequency resolution of pupil responses, because the signal is nonstationary and nonlinear, changing slowly over time. Time domain features may mask interesting information if they are not available across the session. Moreover, the use of Fourier-based methods may be unsuitable because they assume stationarity of the signal. It was therefore necessary to seek another method that would provide better time and frequency resolution, would not assume linearity and stationarity of the signal and would allow meaningful information to be extracted, to better distinguish between different pupillary responses. Variations in the analytic phase of the Hilbert transform (HT)[35] [176] of pupillary responses have recently been used to study cognitive overload [177]. The initial results indicate the potential for using the HT method to distinguish between cognitive overload activities. It therefore seemed appropriate to extract pupillary response features from the HT, as the variations in pupillary response signal are nonlinear and nonstationary.

[35]The Hilbert transform is explained in Appendix E.

Consequently, features based on this method were investigated to detect affects from pupillary responses.

The proposed features based on the HT analytic signal were extracted from the HT instantaneous magnitude (the smooth envelope) and unwrapped instantaneous phase, as detailed in Table 4.2. The instantaneous magnitude $A(t)$ describes the envelope of a given signal $x(t)$ over time and is computed by Equation 4.19. It can provide information about the activities in the signal more locally. The instantaneous phase angle $\theta(t)$ is computed by taking an inverse tangent function, as shown in Equation 4.20. The unwrapping corrects the radian phase angles by adding $\pm 2\pi$ whenever the phase angles equal or exceed the tolerance of π radians [178]. It has been shown that sharp changes in the unwrapped phase are related to cognitive overload [177]. Figure 4.8 shows an example of the HT unwrapped instantaneous phase across the affective games. Continuous changes during the course of the game may indicate cognitive dissonance [177], which involves a state of high arousal [179]. This was clearer during games designed to elicit high arousal compared to low arousal states. It can also be seen that games designed to elicit negative valence were characterised by strong fluctuation (sharp changes) that might be related to an increase in cognitive processing during the game [177, 180], while the fluctuations during games designed to elicit positive valence were smoother. These differences indicate that the features extracted using HT can be used to detect a player's affective experience.

$$\mathrm{A(t)} = |\sqrt{\mathrm{x(t)}^2 + \tilde{\mathrm{x}}(\mathrm{t})^2}\,| \tag{4.19}$$

$$\theta(\mathrm{t}) = \tan^{-1}\left[\frac{\tilde{\mathrm{x}}(\mathrm{t})}{\mathrm{x(t)}}\right] \tag{4.20}$$

where $x(t)$ is the real signal and $\tilde{x}(t)$ is the Hilbert transform. Several statistical features of fixation duration, saccade-related measures and blinking were extracted. Table 4.2 summarises the features extracted from eye movement signals. The HT analysis and all eye movement features were extracted using MATLAB[36]. In total, 30 features were extracted from eye movement signals.

[36] http://www.mathworks.com

Table 4.2: Features extracted from eye movement signals.

Signal	**Extracted features**
Pupillary responses	Mean, standard deviation (SD), mean of derivative, spectral entropy, PSD from four bands: 0.0-0.2 Hz, 0.2-0.4 Hz, 0.4-0.6 Hz, 0.6-1 Hz.
Pupillary response-based HT	Mean, SD, root mean square and maximum of the instantaneous magnitude and unwrapped instantaneous phase.
Fixation	Mean, SD and mean of derivative of fixation duration (ms). Fixation frequency.
Saccade	Mean, SD and mean of derivative of saccade duration (ms), amplitude (degree). Saccade frequency.
Blinking	Mean, SD of blink duration (ms). Blink rate.

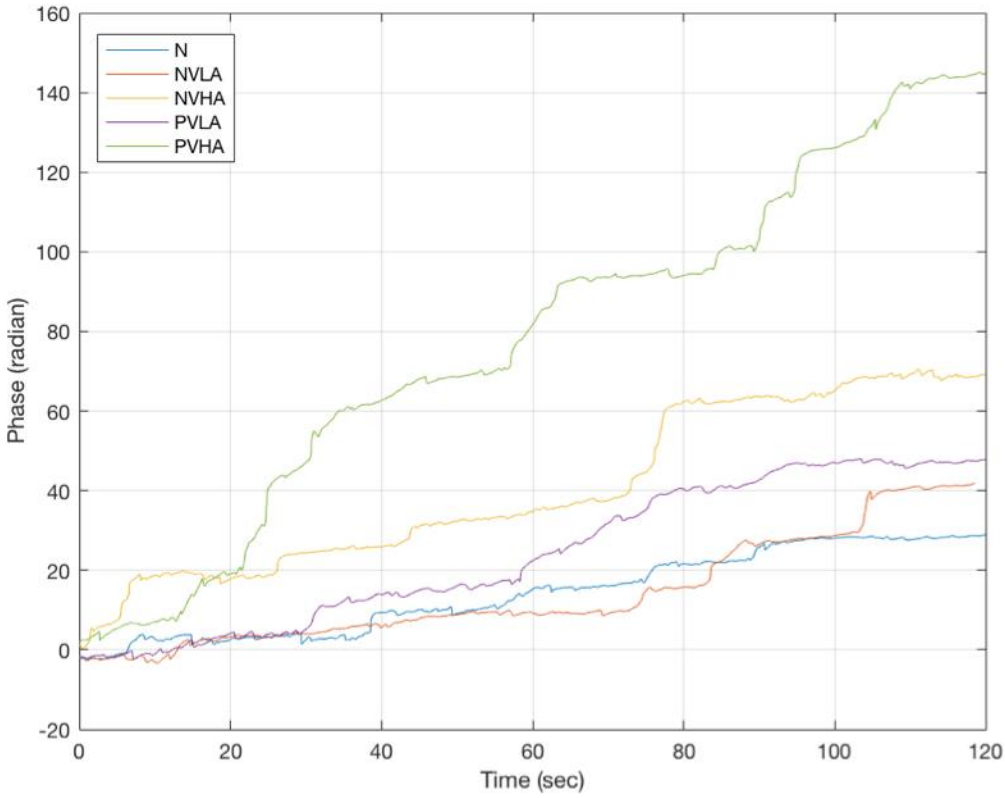


Figure 4.8: An illustration of the HT unwrapped instantaneous phase of pupillary responses during different affective games for player 15. N: neutral, NV: negative valence, LA:low arousal, PV: positive valence and HA: high arousal.

4.4.3 Analysis of affect elicitation

This section explains how the eye movement signals were analysed to determine whether playing games in the virtual interactive environment elicited different affective responses on the arousal and valence dimensions, according to the ground truth labels (see Section 4.2),

which would be reflected in pupillary responses, fixation, saccade and blinking features. One-way repeated measures analysis of variance (ANOVA)[37] was conducted to determine the significance of these affective responses on the arousal and valence dimensions, separately.

Figure 4.9 clearly shows that pupillary responses varied across the different affective games, on both dimensions. On average, it was found that the HA games evoked stronger pupillary responses than the N and LA games. Moreover, pupillary responses were observed to decrease during LA games compared to N games. On the valence dimension, playing both NV and PV games resulted in larger pupillary responses compared to the N game. These variations are consistent with previous studies using passive (noninteractive) stimuli [63, 130]. The variations in the mean across the three states of both dimensions were found to be statistically significant, $p < 0.05$. This finding reveals that the affective games with different levels of arousal and valence modulated the players' pupillary responses, consistent with results reported in the literature that pupillary responses correlate with affects when induced using passive (noninteractive) affect elicitation techniques (e.g. [9, 130, 131]).

Fixation duration was observed to be longest in LA games, followed by HA games, and the shortest duration was seen during N games (Figure 4.9). Similarly, longer fixation duration was observed during both NV and PV games in comparison to N games. These variations were found to be statistically significant on both dimensions, $p < 0.05$. A possible explanation for the observed increase in fixation duration is that the players' focus of attention increased while playing games that were designed to elicit a strong affective experience in comparison to the neutral ones. This is consistent with the findings by [128], which showed that affective movies evoke more focus and less likely to be related with the visual field.

Saccade amplitude was observed to be highest in PV games, followed by NV then N games, $p < 0.05$. However, no significant difference in saccade amplitude was observed on the arousal dimension, $p > 0.05$. This indicates that the saccade signal has the potential to be more useful for affect detection on the valence dimension than on arousal.

As to blinking, the rate tended to be higher during LA games and lower during HA games in comparison to N games, $p < 0.05$. A high blink rate has been shown to be

[37] The significance of ANOVA indicates the existence of a significant difference between a feature's mean across at least two of the three affective states.

associated with with low arousal [127]. Thus, it is not surprising that blink rate was observed to rise as the arousal level of the game decreased. In fact, it indicates that players were more relaxed while playing those games designed to elicit a low arousal state. On the valence dimension, on average blink rate variations were not significant, $p > 0.05$. This suggests that blinking is less useful for distinguishing affects on the valence dimension.

Overall, the results demonstrate that playing in the virtual interactive gaming environment successfully induced different eye movement responses at different levels on both the arousal and valence dimensions, consistent with findings for passive (noninteractive) affective stimuli reported in the literature [9, 130, 131]. It can be concluded that eye movement signals have the potential to be used to detect players' affective experiences while playing in an interactive virtual environment.

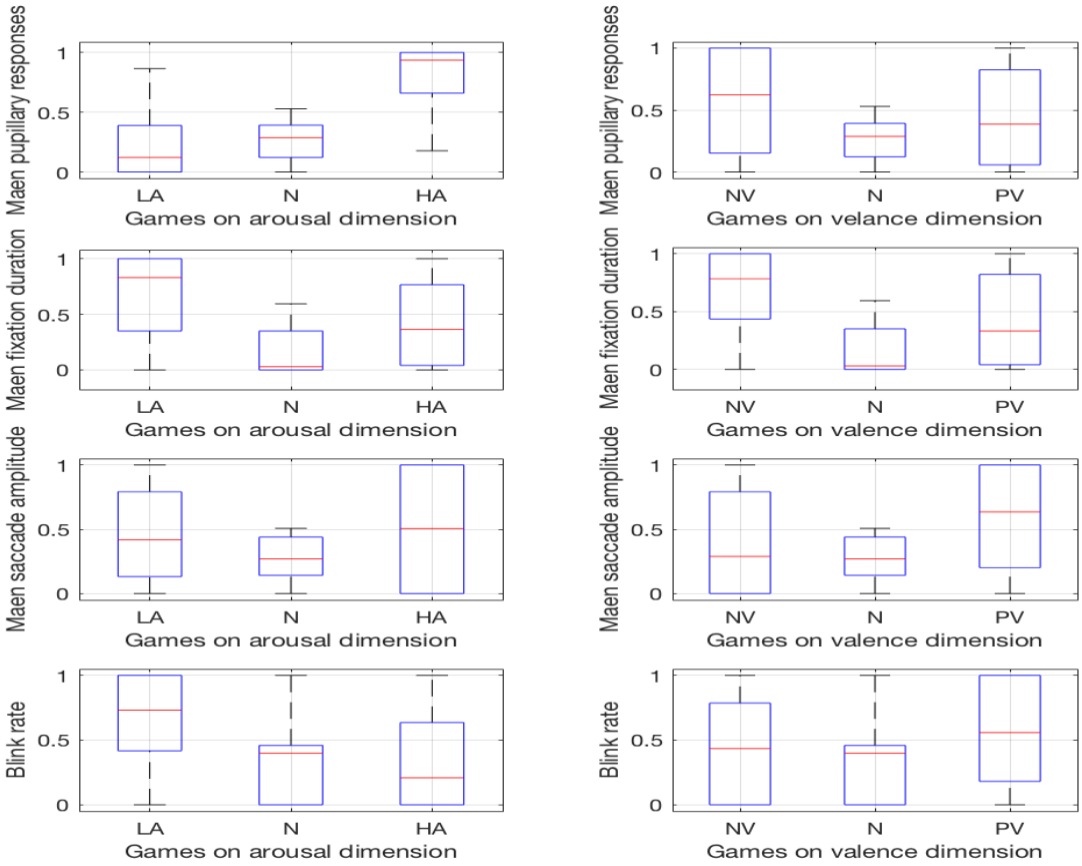


Figure 4.9: Box plots of different eye movement features of the three affective states on the arousal and valence dimensions. One-way repeated measures ANOVA results revealed significant differences between pupillary responses and fixation duration means on both dimensions ($p < 0.05$). Saccade amplitude and blinking rate means were significantly different on valence and arousal dimensions, respectively.

4.4.4 Experimental settings

The goal of the ADS was to use eye movements signals to automatically detect players' affective experience elicited while playing in the virtual interactive gaming environment as one of the arousal (LA, N or HA) and valence (NV, N or PV) states, according to the ground truth labels (see Section 4.2). The ADS was implemented using a multiclass support vector machine with RBF kernel and taking the one-against-one approach, where three separate models were trained on each pair of affective states (as described in Section 4.3). Due to the imbalanced affective state distribution in the aGame corpus, a uniform prior probability of 1/3 was assigned to each affective state (class). System performance was reported in terms of (unweighted) detection accuracy averaged over players.

The system was evaluated using different feature sets extracted from eye movement signals, as described in Section 4.4.2, and on the combination of all the features by fusing them together to form a single feature vector. The evaluation was conducted using the 'leave one player out' cross-validation strategy, where each cross-validation iteration contained testing data from one player and training data from the remaining players (see Section 4.3.4). For each testing player, the best model parameters (C , γ) that minimised the mean square error of the training data were determined by searching the parameter space in the range $[2^{-5}, 2^{15}]$ over a tenfold cross-validation using grid search. The FDR feature selection was applied to the training data in order to select the most relevant features with good discrimination across the states. The FDR threshold was set empirically during the experiment from values of [0.1, 0.15, 0.20, 0.25, 0.30, 0.35]. Since the selection was made at each cross-validation iteration, the features might be different across the players.

4.4.5 Results

Table 4.3 and Figure 4.10 present the detection results obtained using different feature sets on the arousal and valence dimensions. It can be seen that the detection accuracies of all feature sets were above the random accuracy of 33.3%. This reveals that eye movement signals can detect affective states elicited in an interactive gaming environment, on both dimensions. Pupillary response and blinking feature sets clearly provided better performance

on the arousal than the valence dimension, whereas the saccade and fixation feature sets yielded superior performance on the valence dimension.

Table 4.3: Affect detection results for different feature sets of eye movement signals. PRHT performed best on arousal dimension and SAC+FIX performed best on valence dimension.

Feature sets	**% Detection accuracy**	
	Arousal	**Valence**
Pupillary response	70.0	56.1
Pupillary response with HT (PRHT)	**76.0**	61.4
Saccade (SAC)	58.6	64.3
Fixation (FIX)	50.0	53.5
Blinking	42.8	41.4
SAC+FIX	60.0	**71.4**
Combined features (PRHT+SAC+FIX)	71.4	58.6

The use of pupillary response with the Hilbert transform features (PRHT) led to an improvement in detection performance over the conventional pupillary response (PR) feature set, with 76% accuracy on arousal and 61.4% on the valence dimension. It can be seen from Figure 4.11 that PRHT improved the detection accuracy of the LA and HA states by 22% and 5.3% respectively, with inferior accuracy observed only for the N state. Similarly, the detection of PV and NV states were improved, showing an increase in accuracy by 12.5% and 12.9% respectively. However, the detection accuracy of the N state remained unchanged. The performance of PR was compared with PRHT to discover whether the differences were statistically significant. The paired sample t-test was applied to compare the accuracy from each player's samples. The results show that using PRHT yielded a significant performance improvement on the arousal dimension, $p < 0.05$, while the improvement on the valence dimension was insignificant. Overall, detection accuracy on the arousal dimension was higher than on the valence dimension, because the pupillary response signal is strongly related to the arousal dimension [61, 63, 122].

The worst detection performance on the arousal dimension (42.8% accuracy, only slightly better than random) was obtained using the blinking feature set. As can be seen from Figure 4.11, blinking features provided good accuracy (better than random) only for LA and

N states while random accuracy was observed for HA state, which might be due to the high confusion with the other states. This indicates that blinking features do not seem to be suitable measures for detecting affective states on the arousal dimension. Similarly, the feature set did not perform well on the valence dimension, giving 41.4% accuracy. In fact, this result aligns with the finding reported in Section 4.4.3 that differences in blinking behaviour across the valence dimension were insignificant. Thus, blinking features were omitted from fusion with the other feature sets on both dimensions.

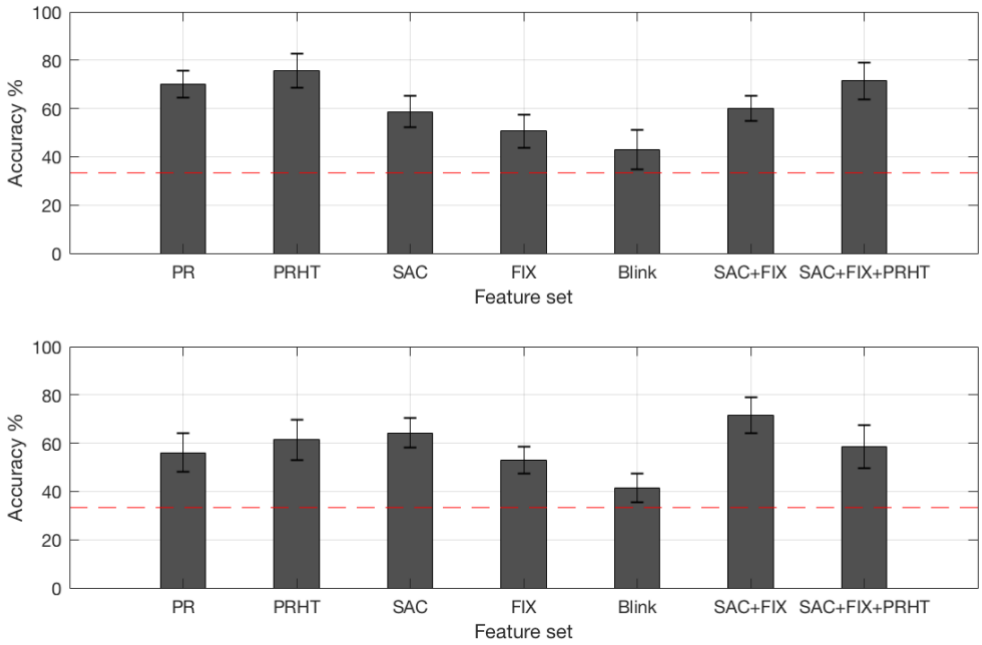


Figure 4.10: Bar graph of detection accuracies on the arousal (top) and valence (bottom) dimensions using different eye movement feature sets, with 95% confidence interval (n=14). Straight line (red) shows random accuracy.

Combining the SAC and FIX feature sets gave a detection performance of 71.4% on the valence dimension, a significant improvement compared with using them separately ($p < 0.05$ - paired sample t-test). Figure 4.11 shows that the detection performance of all valence states was increased, indicating that both feature sets contributed to enhancing detection performance across the valence dimension. However, the SAC+FIX combination did not outperform the pupillary feature set results on the arousal dimension, yielding only 60% detection accuracy. This indicates that these features are less related to the arousal dimension. The random accuracy of HA state using FIX feature set, as Figure 4.11 shows,

might be due to the high confusion with the other states. Combining it with SAC feature set clearly reduced this confusion, since SAC features provided better accuracy for HA state.

The combination of all three eye movement feature sets (SAC+FIX+PRHT) did not yield superior performance. In fact, it provided inferior performance compared to the best feature sets on both the arousal and valence dimension, namely PRHT and SAC+FIX respectively. This suggests that adding less accurate features to a well performing feature may not always yield an improvement in performance. It can be concluded that the PRHT feature models the arousal dimension effectively, and that the saccade and fixation feature sets are better used together in order to model the valence dimension effectively.

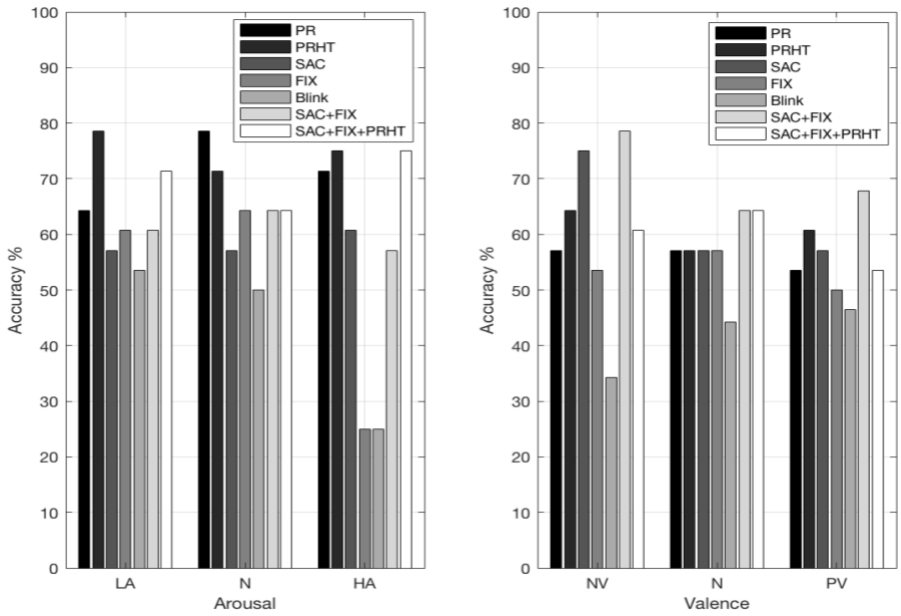


Figure 4.11: Detection accuracies per affective state on the arousal (left) and valence (right) dimensions using different eye movement feature sets.

As stated in Section 4.4.4, the features employed to detect affective states varied across the players. To gain an insight into the most important features in each set, a measure of 'feature significance'[38] was calculated, indicating the percentage of cross-validation iteration that used a particular feature from that set (see Table 4.2 for eye movement feature sets). Table 4.4 lists the most important features from each eye movement feature set that were highly significant. Among all feature sets, it can be seen that the significant features that were employed to detect affects differed across the arousal and valence dimensions. In

[38] A feature is considered significant if it is selected more frequently across the players (cross-validation iteration).

the PRHT feature set, for example, all instantaneous magnitude features were found significant on the arousal dimension, which indicates that these features provided important information in detecting the players' affective states on the arousal dimension, whereas on the valence dimension only the 'maximum' feature was found to be significant (see Table 4.4). In the saccade and fixation feature sets, more features were employed to detect affective states on the valence than the arousal dimension, indicating their stability in discriminating valence states across the players.

Table 4.4: A list of the most important features selected more frequently across the players.

Feature set	**Arousal**	**Valence**
PR	Mean SD Spectral entropy PSD1	Mean of derivative Spectral entropy PSD1 PSD4
PRHT	Instantaneous magnitude: all features Instantaneous unwrapped phase: Mean maximum	Magnitude: maximum Unwrapped phase: maximum
SAC	Frequency Duration: Mean SD	Frequency Duration: Mean SD Amplitude: Mean SD Mean of derivative
FIX	Mean Frequency	Frequency Mean SD
Blinking	Blink rate SD	Mean

4.5 Comparison with the state of the art

Directly comparing the detection accuracy results in Table 4.3 with the state-of-the-art results in the literature is problematic, because of differences in the signals employed and in experimental conditions. Nevertheless, the results obtained in the present study were

compared with those of the most similar studies employing gameplay scenarios and using physiological signals. Table 4.5—to the best of the author's knowledge—summarises the reported results of the most similar studies.

Chanel et al. [181] used Tetris (a classic 2D game) to elicit three levels of arousal (anxiety/engagement/boredom) by adjusting the difficulty level of the game to easy, medium and hard. They obtained a detection accuracy of 53.3% using an SVM with RBF kernel and player-independent cross-validation. The performance was improved further by combining information from EEG signals at decision level, giving 63.3% accuracy [182]. This inferior performance compared to the 76% detection accuracy obtained in the present study reveals the superiority of using eye movements over other physiological signals to detect affects on the arousal dimension in a gaming context. Friedrichs et al. [54] used a simple side-scrolling game to elicit two states of valence (positive and negative) using ECG and GSR signals. They obtained an accuracy of approximately 70% using an SVM classifier, which is relatively similar to the valence result obtained in this study. However, it was based on only two states of valence and obtained using player-dependent cross-validation. Hence, the system would be expected to perform well, as information about each player had already been learnt.

Moghimi et al. [55] obtained 98% maximum detection accuracy from EEG, heart rate and GSR signals, using an SVM classifier and utilising the same gaming environment employed in this study. One possible explanation of this high accuracy is the players' level of gaming experience, as they may have been more experienced than the participants in the present study, making their affective responses stronger and easier to distinguish. Another possible reason is that the system was evaluated using random folding cross-validation, which may have led to the results being overestimated, as the training data may have contained samples from the testing player. Therefore, it would be expected that the system would perform poorly when new player data was presented to it, given that the physiological signals employed, especially EEG signals, typically exhibit high variability among individuals [133]. Besides, it is well known that EEG signals are noisy, highly sensitive to physiological artefacts [3, 183] and have shown inferior performance in comparison to eye movement signals in many affect detection studies (e.g. [9]). The present study contributes to that area of work by investigating new pupillary response features based on the Hilbert transform to enhance the affect detection performance.

Other studies have evaluated players' affective experiences in various gameplay scenarios using physiological signals, but no automatic affect detection systems have been developed and evaluated. A review of these studies can be found in [184]. To the best of the author's knowledge, this thesis is the first to explore, in an interactive virtual gaming environment, the use of the eye movement modality in the automatic detection of affective states corresponding to the arousal and valence dimensions of Russell's circumplex model. The promising results obtained here (see Table 4.3) suggest the feasibility of using eye movement signals for affect detection in preference to other physiological signals, which are mostly invasive with no potential for remote sensing (e.g. EEG), especially scenarios where the objective is to make the interaction more realistic and enhance the the individual's experience. Moreover, the setup of such physiological sensors is not easy and can discourage individuals, in comparison with eye movement tracking sensors.

Table 4.5: State-of-the-art comparison of different affect detection studies in gameplay scenarios using physiological signals. Eye movement signals seem promising for affect detection.

Study	Signals	Game	Affects	Players	Method	Results
Chanel et al. [181]	Peripheral	Tetris	Arousal	20	CV P.indep	53.3%
Chanel et al. [182]	Peripheral, and EEG	Tetris	Arousal	14	CV P.indep	63.3%
Friedrichs et al. [54]	ECG, and GSR	Simple side-scrolling	Positive /negative valence	47	CV P.dep	70.0%
Moghimi et al. [55]	EEG, GSR, and Heart rate	Virtual interactive game	Arousal/ valence (2D)	30	CV random	98%
Current research	Eye movements	Virtual interactive game	Arousal Valence	14	CV P.indep	76.0% 71.4%

EEG: Electroencephalography, ECG: Electrocardiogram, GSR: Galvanic Skin Response. Peripheral signals include: GSR, blood pressure, heart rate, breathing, skin temperature. CV: cross-validation. P.indep: player independent. P.dep: player dependant.

4.6 Summary

This chapter has reported the development of an ADS capable of inferring game players' affective experience using eye movement data from the aGame corpus, collected while they interacted with a virtual gaming environment. The players' self-reported felt affective experience was first analysed and the results led to the use of the ground truth labels (as

described in Table 3.1), corresponding to the arousal and valence dimensions of Russell's circumplex model, to represent their affective states in analysing aGame corpus data and in the affect detection task in this chapter and subsequent chapters. The chapter next detailed the detection methods used throughout this study.

An affect elicitation analysis of the aGame corpus eye movement signals demonstrated that playing in the virtual interactive gaming environment successfully elicited the expected affective states from the players, according to the defined ground truth labels. A player-independent affect detection system was accordingly developed, using different features extracted from eye movement signals, namely pupillary response, fixation, saccade and blinking. New pupil response features based on the Hilbert transform were also proposed for affect detection.

The analysis of the detection results indicates the feasibility of using eye movement as a modality for affect detection in interactive virtual environment, a gaming context, with pupillary response features based on the Hilbert transform yielding superior performance on the arousal dimension, whereas the saccade and fixation features yielded superior performance on the valence dimension. This demonstrates the potential of using eye movement sensing to develop affect-aware interfaces suitable for HCI applications such as interactive gaming, in order to facilitate adaptive experiences and make the interaction more efficient and natural. Although a wearable sensor was employed in this study, due to its ability to track pupillary responses with high precision [27], advances in the development of remote eye movement tracking sensors offer the potential to enable players to interact more freely within the virtual gaming environment, allowing more natural and less invasive interaction. The contributions to knowledge outlined in this chapter would be valuable to the HCI community and to game designers and developers.

This chapter has evaluated players' affective experience over the whole duration of the game, but it may change during games in response to in-game events. Thus, it would be interesting to evaluate their affective experience continuously. However, establishing the ground truth for continuous detection is an open issue that requires further investigation.

The next chapter investigates the development of an ADS using players' speech and the multimodal fusion of eye movement and speech signals. The findings reported in both chapters (4 and 5) are then applied in Chapter 6 to investigate the improvement of ASR system performance during affective speech.

Chapter 5

Affect Detection from Speech and Integration of Modalities

Chapter 4 described how players' affective experiences elicited while playing various games in the interactive virtual environment were investigated and detected automatically using eye movement signals. It appears that using eye movement data to detect players' experienced affects performs well in comparison with more invasive measurement of physiological signals. This chapter discusses the development of an affect detection system using the spontaneous speech data of the aGame corpus, captured via the speak-aloud approach. It also investigates the integration of the eye movement and speech modalities to enhance the detection process.

5.1 Motivation

With the expanding use of speech-enabled interfaces in many HCI applications, detecting affects from individuals' speech is attracting considerable attention from the affective computing community. Despite advances in this field, speech-based ADS technology has not yet been fully integrated into real-life HCI applications. A key challenge is to identify the affective states that people actually experience, in other words to obtain realistic affective speech responses on which to base the development of a speech-based ADS that performs well enough to be used in real-life HCI applications.

As mentioned in Chapter 2, since most of the contributions to the field are based on acted speech data, which does not reflect realistic affective behaviour, the results cannot be generalised to real-life HCI applications. Alternatively, speech data can be collected in real-life situations (e.g. call centers [8]) or while interacting with artificial agent, but this may not contain affective responses and is typically annotated by external observers. A previous study by [13] captured affective speech data during dyadic conversations in multi-player gameplay scenario. This thesis develops this idea further by considering a player speaking aloud while playing in the interactive virtual gaming environment, which allows the simulation of real-life HCI. This approach has the potential to capture natural affective speech responses reflecting the affective states that players experience, which can be detected with meaningful results for real-life HCI scenarios. The previous chapter investigated the elicitation of players' self-reported affective experiences. This chapter investigates the degree to which these elicited affective experiences can be automatically detected from the captured speech.

Developing a speech-based ADS that can be generalised to real word is highly desirable for speech-enabled HCI applications. However, individuals may deliberately masked or faked their affective experience during spontaneous interactions, so detecting the actual felt affective states collected from speech data in the aGame corpus may not provide accurate results. Since affects are expressed via both external behaviour and internal physiological responses, represented by speech and eye movement signals in the aGame corpus, it is intuitively likely that information from the eye movement modality will complement speech data and that integrating the two modalities will allow a more robust detection of affects. The multimodal integration of information from eye movement and speech signals to enhance the detection process is investigated in Section 5.4.

5.2 Development of an affect detection system based on acoustic speech signals

As described in Chapter 3, speech data was captured by asking players to speak aloud while interacting with the gaming environment. It was hypothesised that speech thus captured could be used to detect the affective states that players experienced during the game. This section explains the development of an affect detection system based on acoustic speech

signals collected in the aGame corpus. It describes the signal processing procedure and the extracted acoustic features. Finally, it sets out the affect elicitation analysis and detection results.

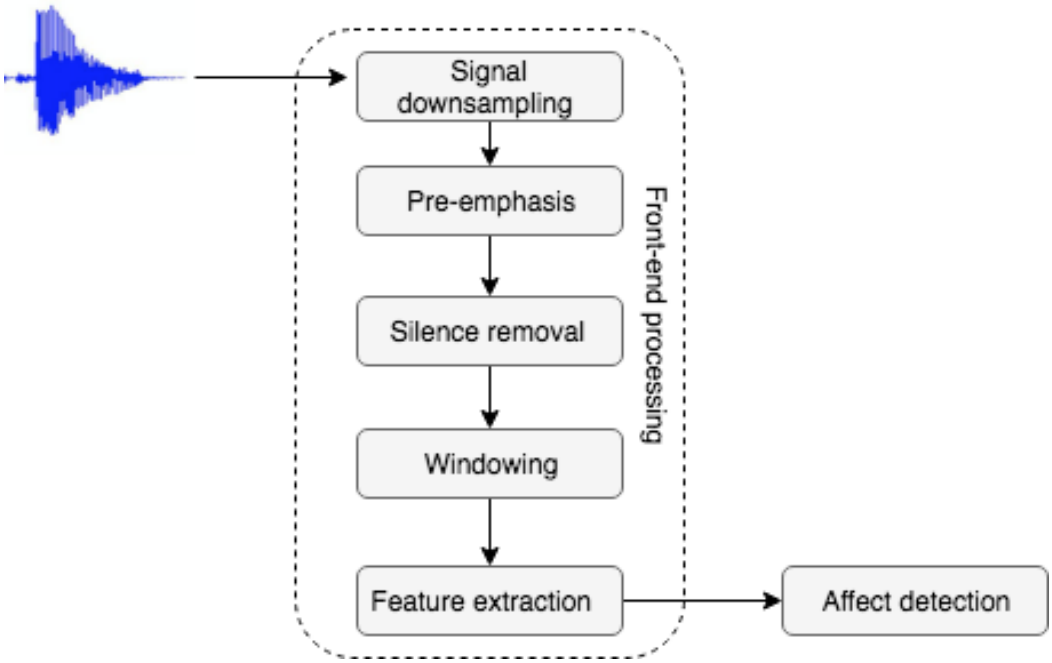


Figure 5.1: Overview of the front-end processing of acoustic speech signals.

5.2.1 Acoustic speech signal processing

The purpose of the acoustic signal processing step was to improve the efficiency of feature extraction. Figure 5.1 shows the front-end processing of speech signals prior to the affect detection process. The speech signal was first converted to mono by taking the average value of the two channels, then downsampled to 16 kHz from the original recording rate of 44 kHz, in order to match the baseline ASR system (described in Chapter 6), which was trained on 16 kHz speech data.

The speech signal was then filtered by applying a pre-emphasis filter in order to balance the spectrum, as given by Equation 5.1. In human speech signals, high frequencies usually have lower magnitudes than low frequencies, due to the lip radiation effect [185]. The pre-emphasis filter compensates for this by boosting the magnitude of high frequencies. It can also improve the overall signal-to-noise ratio.

$$y(x) = 1 - \alpha x^{-1} \tag{5.1}$$

where x is the speech signal and α is the filter coefficient, whose values range between 0.95 and 0.99. In this study, α was set to 0.97.

Speech segments were then extracted by detecting silences, a silence removal step. Several algorithms for detecting silences in a speech signal have been proposed [186]. The algorithm used in this study applies a simple thresholding scheme on two features: the signal energy and spectral centroid. The algorithm is described in [187] and the threshold was calculated as given by Equation 5.2. Since voiced units (non-silent speech) are normally characterised by high energy and spectral centroid values compared with unvoiced units (silent speech) in cases where there is a low level of background noise, the detection of silences using this method was deemed efficient for speech signals in the aGame corpus. The resultant segments were used for affect detection and later, as described in Chapter 6, for ASR system evaluation. Figure 5.2 illustrates an example of the detected speech segments.

$$T = \frac{W*(L_1+L_2)}{W+1} \tag{5.2}$$

where L_1 and L_2 are the locations of the first and second local maxima of the smoothed histogram of the two features and W is a weighting parameter determined empirically by trial and error. T was calculated separately for each feature sequence.

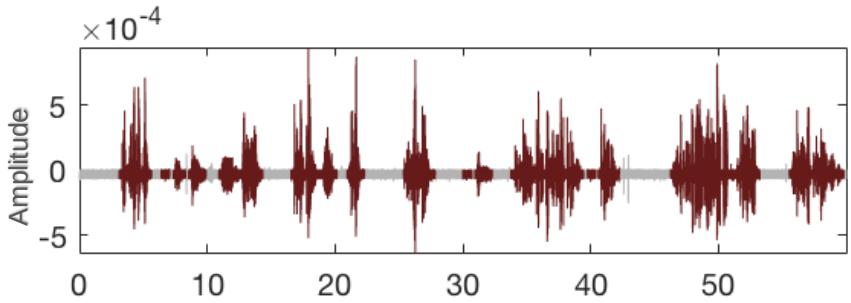


Figure 5.2: An example of detected speech segments/voiced units (red).

The speech segments were then divided into short-term windows, also called frames, such that within each frame the signal should be approximately stationary. Frame size typically ranges between 20 and 30 ms [188], where the vocal tract shape remains relatively constant during speech production and the frequency spectrum is thus relatively stationary. Approximately half of the window size is typically used as an overlapping interval in order to capture changes in the signal over time. In this study, the segments were divided into frames of 25 ms with a 15 ms overlap. The resulting frames were then multiplied by a hamming window in order to reduce the abruptness of changes in the signal by increasing the amplitude at the frame boundaries [189], which helped to avoid the problem of discontinuities during frequency domain analysis. The final step of the the front-end processing is the feature extraction which is described in the following section.

5.2.2 Extraction of acoustic features

The main step in speech-based affect detection is to extract the most informative feature set in order to effectively represent the speaker's affective state. As described in Section 2.2.2, various acoustic features have been investigated and used in affect detection. However, identifying the best feature set remains an open question. The choice of features depends mainly on the characteristics of the corpus (e.g. language, act/spontaneous). This thesis investigates three acoustic categories widely mentioned in the literature: prosodic, voice quality and spectral features. The following subsections describe the key LLDs from each category used in this thesis, which have proven to be useful for affect detection.

Prosodic features

Four prosodic features have been widely used in affect detection tasks: fundamental frequency, energy, speaking rate and zero-crossing rate [13, 83, 87].

- Fundamental frequency

The frequency of human speech is controlled by the vocal folds, which are the primary source of speech sounds. During speech production, the lungs produce a pressurised airflow which passes over the vocal folds [75]. These vibrate to modulate the airflow and produce sounds, with the rate of vibration defining the fundamental frequency f_0, which is perceived

as voice pitch. Speakers typically modulate their speech f_0 to convey meaning. It has been shown that the f_0 carries information about the speaker's affective state [83, 84]. To estimate f_0, several algorithms have been developed in both the time and frequency domains [190]. The most popular of these is based on the autocorrelation function (ACF), which measures the similarity between a given signal and itself [191]. According to [191], f_0 is estimated by first shifting the speech signal, then calculating the correlation between the original signal and each shift (time lag). The fundamental period T_0 of the periodic signal is then defined as the lag in which the autocorrelation is maximised and f_0 can be calculated as given by Equation 5.3. At a given time t, f_0 can be computed from the local maxima of the ACF over a short-term window centred around t. As speech signals are pseudo-periodic, the assumption of periodicity still holds, where the signal repeats itself after a specific period T_0. This algorithm is simple and considered an accurate way to estimate fundamental frequency.

$$f_0 = 1/T_0 \tag{5.3}$$

- Energy

The source of energy in the speech signal is the pressure in the lungs that impels air through the vocal tract. For a given speech signal, the energy over a short-term window can be computed by Equation 5.4, where N is the frame length. Typically, the energy is normalised to account for the dependency on the window length, which referred to as the power of the signal.

$$E = \sum_{n=1}^{N} |x(n)|^2 \tag{5.4}$$

- Speaking rate

The speaking rate is defined as the ratio of the total voice to unvoiced speech units' durations, as given by Equation 5.5.

$$Speaking\ rate_s = \frac{voice\ units\ duration\ in\ S}{uvoice\ units\ duration\ in\ S} \tag{5.5}$$

where S is the speech segment. Speaking rate is calculated per speech segment.

- Zero-crossing rate

The zero-crossing rate (ZCR) is the rate at which the amplitude of the speech signal changes from positive to negative values and vice versa. It is typically higher for unvoiced than voiced units. The ZCR is calculated as given by Equation 5.6 [190].

$$ZCR = 1/N \sum_{n=1}^{N} \left| sgn\big(x(n)\big) - sgn(x_{-1}(n-1)) \right| \tag{5.6}$$

where *N* is the frame length and sgn equals –1, 0, or 1.

Voice quality features

Voice quality describes properties of the speech signal such as stressed, harsh, whispery or breathy voice. The correlation between voice quality and affects has been studied extensively [82, 84]. Voice quality features include jitter, shimmer and harmonic-to-noise ratio (HNR). Jitter measures the periodicity variations of the fundamental frequency, while shimmer measures the peak-to-peak variations in the energy level of the speech signal. They are calculated as given by Equations 5.7 and 5.8 respectively. HNR measures the degree of periodicity of the speech signal [191], in other words, the amount of added noise in the signal. It can be estimated from the fundamental period T_0 defined using the ACF, as given by Equation 5.9.

$$Jitter\ = \frac{|f_{0i} - f_{0i-1}|}{f_{0i}} \tag{5.7}$$

$$Shimmer\ = \frac{|E_i - E_{i-1}|}{E_i} \tag{5.8}$$

$$HNR = 10 * \ log_{10} \frac{ACF(T_0)}{1-ACF\ (T_0)} \tag{5.9}$$

where the $ACF(T_0)$ value is the relative power of the periodic signal and its complement is the relative power of the added noise.

Spectral features

Spectral features are used to represent the speech signal in the frequency domain in addition to f_0. They can describe the vocal tract characteristics very well [188]. The spectral features most commonly used for affect detection are mel-frequency cepstral coefficients (MFCCs) [192] and formants.

- Mel-frequency cepstral coefficients

MFCCs are widely used in automatic speech recognition systems, due to their ability to detect *what* is being said (the spoken words) [192]. MFCC features are also commonly used in speech-based ADSs, as they can provide relatively good detection performance in term of identifying *how* something is being said (Section 2.2.2). They are usually robust to noise and environmental changes. Recently, MFCCs have been shown to complement the prosodic features in affect detection tasks [53, 86]. Figure 5.3 illustrates the main stages involved in extracting MFCC features, as described in [193].

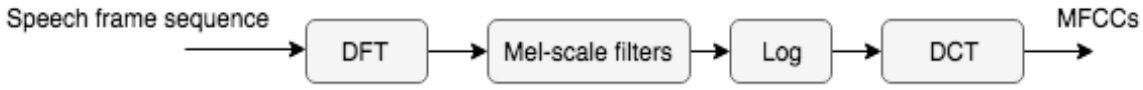


Figure 5.3: A block diagram of the MFCC feature extraction process.

To extract MFCC features, DFT [175] is first applied to convert the sequence of speech frames from the time to frequency domain and estimate the power spectrum of each frame. The power spectrum is then passed through a mel scale bandpass filter bank in order to produce mel scale power spectrum coefficients. Finally, the resulting spectrum coefficients

are converted into log domain and decorrelated by applying a discrete cosine transform (DCT) [194], in order to produce the MFCC features.

The mel scale [195] is a perceptually motivated frequency scale. It models human speech perception frequencies non-linearly, as given by Equation 5.10, where f is the speech spectrum in the frequency domain and $M(f)$ is the spectrum in the mel frequency domain.

$$M(f) = 2595log_{10}(1 + \frac{f}{700}) \tag{5.10}$$

The bandpass filter bank includes a set of triangular mel filters that are applied to each frame. The filters are equally spaced on the mel scale, which are approximately linear at low frequencies (≤ 1 kHz) and logarithmic at high frequencies (> 1 kHz) [192], in order to model the human hearing of high frequency sounds. Figure 5.4, shows an example of the mel scale bandpass filter bank.

The main advantage of using MFCCs is that they can represent the speech signal using a small number of coefficients. However, they are referred to as 'static' features because they describe the spectral characteristics of speech on a frame basis, where the signal is relatively stationary. To describe the dynamic behaviour of static MFCCs over time, delta (Δ) and acceleration (ΔΔ) coefficients are calculated for each MFCC by taking the first and second order time derivatives respectively. Typically, the first 13 MFCCs with their Δ and ΔΔ coefficients are used for affect detection [99], where the first coefficient represents the energy component and the remaining coefficients describes the vocal tract shape representing speech articulation [188].

- Formants

Formants are acoustic energies concentrated at specific frequencies in the speech signal, because of acoustic resonance in the vocal tract, and can be described by their centre frequencies and bandwidth [53]. Formant frequencies occur at approximate intervals of 1 kHz. They depend on the vocal tract shape, which is modified during affective speech, thus affecting the formants. The voice units of a speech signal can include more than four formants, the frequencies and bandwidths of the first and second formants being typically used for affect detection [83]. A simple method to extract formant frequencies and

bandwidths is based on linear predictive coding (LPC) analysis [196], where the vocal tract is modelled as a linear filter of *p*-order. The order refers to the number of LPC coefficients in the filter and is generally recommended to be equivalent to the speech sampling rate in kHz plus two [197]. For the 8 kHz bandwidth, for example, 18 LPC coefficients are used between 0 and 8 kHz, such that there are two coefficients for each 1 kHz of the speech frequency spectrum, while the remaining two coefficients are used to model the spectral slope. Once the speech signal has been subjected to LPC analysis, the formant frequencies and their bandwidths can be simply estimated by finding the peaks in the filter responses using the 'peak picking' method, which involves Fourier transform analysis: the peaks determine the formant centre frequencies and their spread determines the bandwidth.

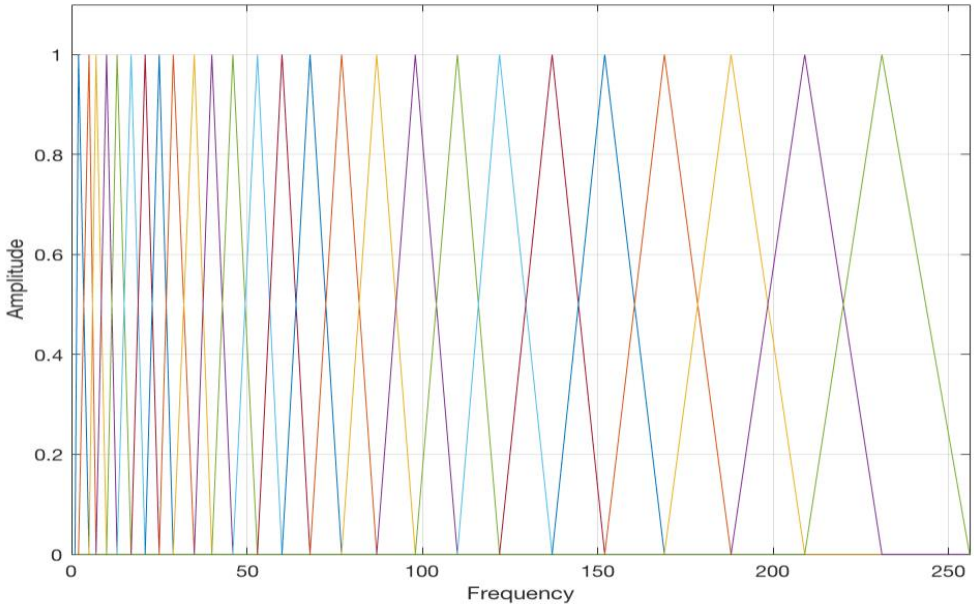


Figure 5.4: An illustration of the mel scale filter bank, with 26 triangular filters, for 8 kHz bandwidth. At the centre frequencies the filters' response is 1, then it decreases linearly towards 0. The filters have a response of 0 at the centre frequencies of adjacent filters, where they overlap.

Statistical features were computed over the extracted LLD feature sequences to generate game-level features, as shown in Table 5.1. A total of 300 features were extracted from each player's speech, comprised of 50 LLDs and six statistical features. Variability between players was minimised by scaling all features between 0 and 1 (Equation 4.13). Front-end processing including feature extraction was performed using MATLAB.

Table 5.1: Features extracted from speech signals.

Feature Category	**Extracted features**
Prosodic	Fundamental frequency. Energy Speaking rate and zero-crossing rate
Voice quality	Jitter, shimmer and harmonic-to-noise ratio
Spectral	13 MFCCs, with their **Δ** MFCC, and **ΔΔ** MFCC coefficients. The first two formant frequency and their bandwidth.
Statistical features	Mean, standard deviation, range, maximum, minimum, and root mean square.

5.2.3 Affect elicitation analysis

This section considers the influence of players' affective states, elicited in the virtual interactive gaming environment, on their speech signals. This is done by investigating the differences of acoustic features across the three different affective states, on the arousal (low, neutral and high) and on the valence (positive, neutral and negative), extracted from speech data in the aGame corpus. One-way repeated measures ANOVA was applied to determine whether the differences between the means of these features were significant on both dimensions, separately.

Due to affect elicitation, significant increases in average f_0 and energy were observed in HA games in comparison to LA and N games, $p < 0.05$ (Figure 5.5). However, f_0 and energy features did not yield significant changes across the valence dimension, $p > 0.05$. This is consistent with previous studies which report that the variations in f_0 and energy are correlated with the arousal dimension (see Chapter 2).

A significant increase in the average ZCR was observed in HA games, followed by LA then N games, $p < 0.05$. Similarly, on the valence dimension, the largest increase was observed for PV and NV games, then N games, $p < 0.05$. The variations in speaking rate across the different games on both the arousal and valence dimensions were found to be insignificant, $p > 0.05$. This indicates a consistency in terms of speaking rate across the

games on both dimensions, which means that players were speaking out loud as instructed during the games. In terms of affect detection, it seems that speaking rate features were less useful.

As to spectral features, on average several MFCC features were significantly different across both dimensions, $p < 0.05$. Table 5.2 provides a list of the most relevant MFCCs. It can be seen from the table that MFCCs were more indicative of the arousal dimension, as more features were found to be significant. However, the relevance of MFCC features over the valence dimension seems to be in agreement with [79], who argue that valence states influence speech articulation characteristics. In regard to formant features, the variations of the average formant frequencies and bandwidths were found to be significant on the arousal dimension, $p < 0.05$, whereas the changes in formant features mean on the valence dimension were insignificant, $p > 0.05$.

Table 5.2: List of the most relevent MFCCs ($p < 0.05$) on both the arousal and valence dimensions.

	MFCC	**Δ MFCC**	**ΔΔ MFCC**
Arousal	0,2,3,4,5,6,7,8,9,10,11,12	1,4,11	5,7,11,12
Valence	0,2,5,6,8,9,10	1,7	0,5,7,12

Regarding voice quality features, players in HA games were on average characterised by a significantly higher shimmer and HNR than in N followed by LA games, $p < 0.05$. Jitter features, however, did not yield significant changes across the arousal dimension. No significant variations corresponding to voice quality features were noted across the valence dimension, $p < 0.05$. This indicates that voice quality features are potentially useful for affect detection on the arousal dimension.

It can be concluded that the interactive virtual gaming environment successfully induced different affective responses from the players' speech. These responses were more indicative of the arousal dimension and were less likely to be associated with the valence dimension, which is consistent with previous studies (e.g. [13, 90]). Accordingly, acoustic features of speech have the potential to be used in detecting players' affective experiences on the arousal dimension, elicited in the interactive virtual gaming environment, but are less

useful in discriminating affects on the valence dimension. Overall, the results suggest that the players expressed their felt affective experiences while playing by utilising the speaking-out-loud paradigm as instructed.

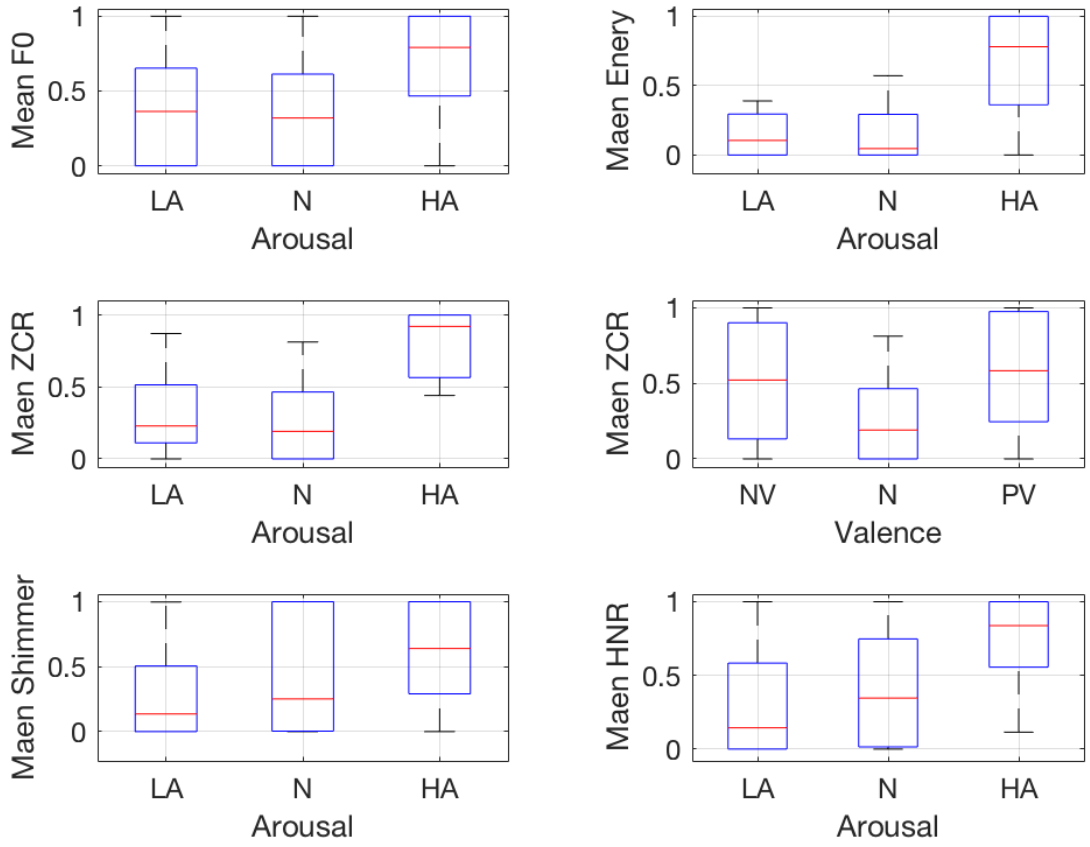


Figure 5.5: Box plots showing different acoustic features on different levels of the arousal and valence dimensions. One-way repeated measures ANOVA results showed signfiicant differences between featuers' means across the affective states of both dimensions ($p < 0.05$).

5.2.4 Experiment settings

In a similar way to that explained in Chapter 4 (Section 4.4.4) for the eye movement-based ADS, an SVM classifier with RBF kernel was employed to detect the affects that players experienced on the arousal and valence dimensions using acoustic speech signals. The acoustic feature sets whose extraction is described in Section 5.2.2 were used to create feature vectors for the ADS. A 'leave one player out' cross-validation strategy was employed to evaluate the system's performance. Accordingly, at each iteration, a feature vector of one player was left out for testing and the feature vectors of the remaining players were used to

train the system (Section 4.3.4). Prior to affect detection, the FDR was applied to each training set to select discriminative features across the states. The threshold was set empirically during the experiment. The SVM-RBF parameters (C , γ) were tuned using tenfold cross-validation on the training data over the range $[2^{-5}, 2^{15}]$ by grid search, and parameters that yielded the best performance are empirically determined. Detection performance using different features sets and the combined vectors of all the features is reported below in terms of detection accuracy, averaged over players.

5.2.5 Results

Table 5.3 and Figure 5.6 show the detection accuracies obtained using different acoustic feature sets separately and the combined features together, on both the arousal and valence dimensions. From the results acoustic features accuracies were higher than the random level of 33.33% on the arousal dimension, indicating their ability to detect the players' affective states elicited in the interactive virtual gaming environment. Regarding prosodic features, the best detection performance on the arousal dimension was 53.5% accuracy, achieved using the ZCR. Speaking rate features, with only 39.3% detection accuracy, do not seem to be suitable measures for detecting arousal states, as it is not significantly better than random (Figure 5.6). This supports the analysis results reported in Section 5.2.3. The performance of voice quality features, at 44.1% accuracy, was no better than the best prosodic features. As to spectral features, the best performance was obtained using MFCCs, with the highest accuracy of 58.3%. Comparing the accuracies obtained from each acoustic feature in Figure 5.6, the MFCC was the most important feature set in terms of detecting affective states on the arousal dimension. This result is consistent with other studies reported in the literature (e.g. [86, 90]).

On the valence dimension, prosodic features do not seem to provide stable information about valence states, as the accuracies were not significantly better than chance, especially for speaking rate, where accuracy was clearly below the random level (see Figure 5.6). This is not surprising, as the analysis in Section 5.2.3 reveals that acoustic features do not provide sufficient information to discriminate between states on the valence dimension. The best accuracy was achieved with ZCR at 44.1% accuracy, followed by MFCC at 40.5%. The

difficulty in detecting valence states from acoustic features seems to be in agreement with previous findings [13, 90].

Table 5.3: Affect detection results for different acoustic feature sets on the arousal and valence dimensions. Combined acoustic features performed best on arousal dimension, whereas ZCR performed best on valence dimension.

Feature sets	**% Detection accuracy**	
	Arousal	**Valence**
f_0	41.7	38.1
Energy	46.4	39.3
ZCR	53.5	**44.1**
Speaking rate	39.3	30.1
Voice Quality	44.1	36.9
MFCC	58.3	40.5
Formants	48.8	35.71
Combination	**65.5**	41.5

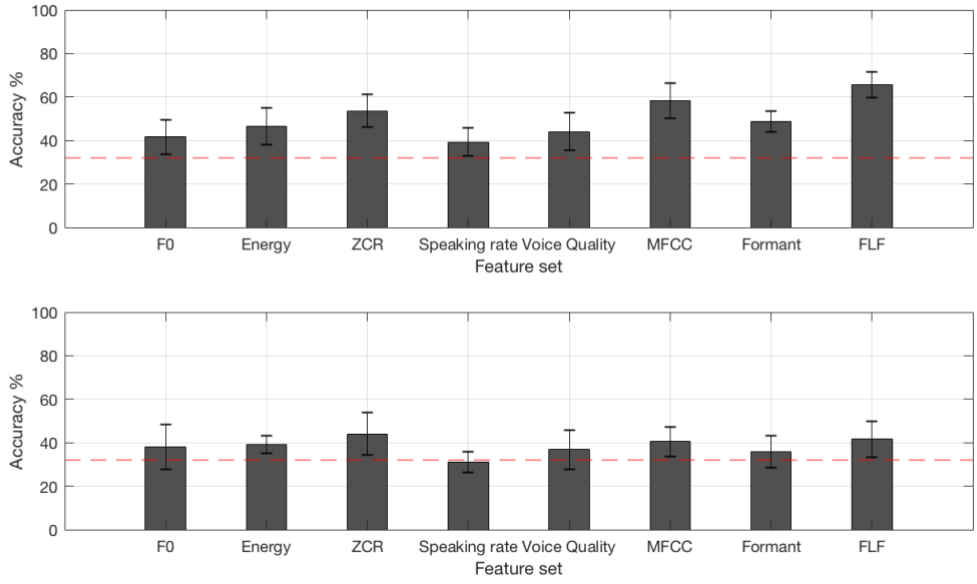


Figure 5.6: Bar graph of detection accuracies on the arousal (top) and valence (bottom) dimensions using different acoustic feature sets with 95% confidence intervals (n=14). Better accuracy indicates greater importance of the feature set for the affect detection task. Straight line (red) shows random accuracy.

On the arousal dimension, the combination of all acoustic feature sets yielded a performance of 65.5% accuracy, superior to the best feature set (MFCC) ($p < 0.05$ - paired

sample t-test). On the valence dimension, however, using all the feature sets together yielded an accuracy of 41.5%, a performance inferior to that of the best feature set, ZCR. This indicates that using all acoustic features did not enhance detection performance and that ZCR can effectively model the valence dimension. Note that speaking rate features were excluded due to their poor performance compared with other feature sets on both dimensions. As can be seen in Figure 5.7, all acoustic features except speaking rate seem to be more informative for the HA state than the other states. On the valence dimension, speech features seem to be more informative of the NV and N states than of the PV state. The low accuracies below chance observed for some states might be due to high confusion with the other states, where the feature set does not carry sufficient information to distinguish that state from others. Moreover, the ADSs were developed based on small dataset from fourteen players (Section 3.5), therefore the detection performance using these features could be improved with more training data.

Direct comparison with reported results in the literature is not possible (e.g. Table 2.1), because of differences in the dataset characteristics (e.g. languages, interaction method such as WOz scenario or natural human-human interaction). Nevertheless, the reported results are in line with studies of affect detection from spontaneous speech in gameplay scenarios (e.g. [13] with average accuracies of 59% and 56% on arousal and valence dimensions respectively), which confirms that this is a challenging task compared to detecting affect from acted speech. A possible reason is that speech is a behavioural signal produced by the vocal tract which can to a large extent be deliberately controlled and manipulated by the speaker. Hence, spontaneous speech captured in relatively natural settings, e.g. virtual interactive gaming environment, may not provide precise information about the felt affective state. In the case of acted speech, affects are typically portrayed by actors. Consequently, it is expected that the differences in speech characteristics across different affective states are clear and can be detected easily with high accuracy, in a way that cannot be generalised to real-life HCI applications. Overall, the results obtained emphasise the need for a more reliable indicator of the actual affective responses of individuals during natural HCI, in order to improve the performance of the speech-based ADS. The next section considers affect detection using linguistic information obtained from speech signals, then Section 5.4 reports an attempt to integrate eye movement as an additional modality to speech in order to improve the detection of players' affective experience.

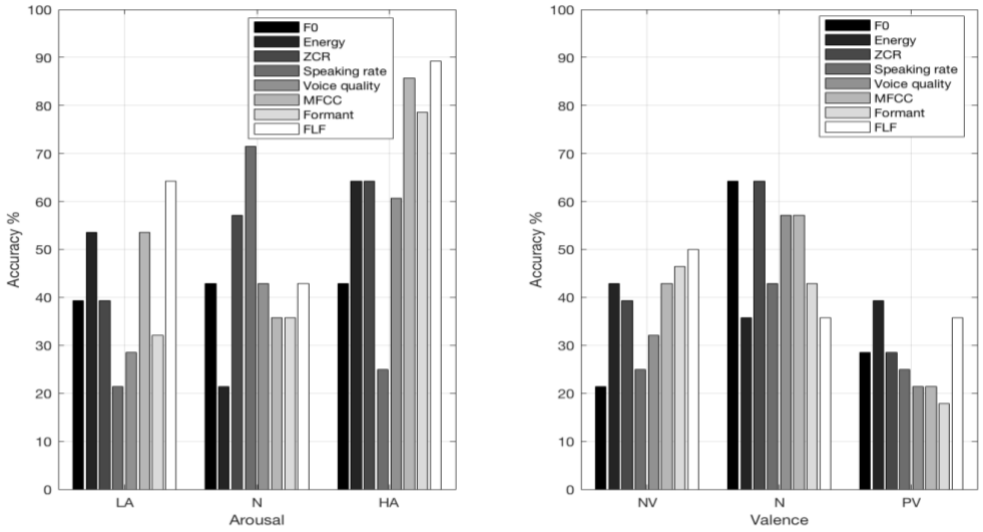


Figure 5.7: Detection accuracies per affective state on the arousal (left) and valence (right) dimensions using different acoustic feature sets.

5.3 Development of an affect detection system based on the lexical content of speech

The previous section considered the detection of players' affects from acoustic speech signals. As discussed in Chapter 2, the linguistic content of a speech signal can carry information about the affective state experienced by the speaker, typically expressed using certain words, e.g. 'no' to convey negative affect. The relation between spoken words and affects has been widely studied in psychology [105] and has been exploited for automatic affect detection (Section 2.2.2). The purpose of the work reported in this section is to determine whether the spoken words in the aGame corpus carry affective information and can be used effectively to detect players' affects on the arousal and valence dimensions.

The following subsections describe the development of an affective lexical dictionary, the representation of lexical features and the conduct and results of an affect detection experiment using different lexical feature representations.

5.3.1 Affective lexical dictionary

A common step in text-based detection problems is to identify the content that is relevant to the task. The aGame corpus consists of game-oriented words that are associated with the

visual field and the game task, as illustrated in Figure 5.8. Clearly, not all of the words in the corpus are related to affects (e.g. ship, maze, torpedo, etc). Thus, to detect affective states from lexical content it was necessary to identify the most relevant vocabulary associated with the affective states in question. Lee et al. [106] introduced the measure of 'affective salience' to give an insight into the affective meaning of spoken words in the application of affect detection. This measures the amount of information that each spoken word carries about affects, based on pairwise mutual information (PMI) [198], where a word is considered to be salient with respect to an affective state if it occurs more often within that state than other states. In this work, the PMI measure was adapted to create the affective lexical dictionary for each affective state in the aGame corpus.

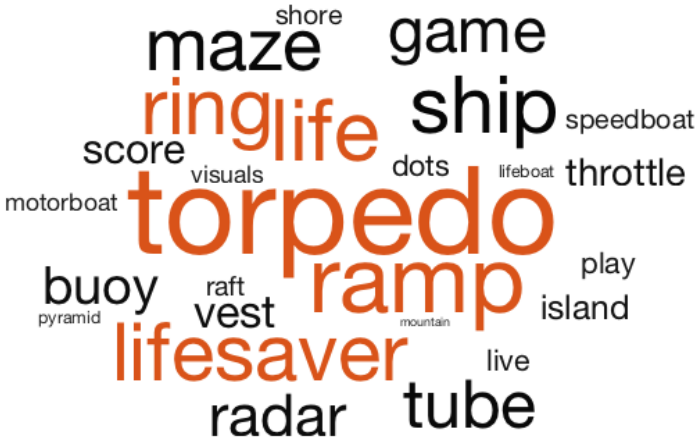


Figure 5.8: An example of Game-oriented words found in the aGame corpus.

The PMI weights the lexical content in the aGame corpus based on their pairwise semantic dependency on the affective states on both the arousal and valence dimensions. The PMI between a word w and an affective state a is determined as given by Equation 5.11.

$$PMI_{w,a} = \log\left(\frac{p(a|w)}{p(a)}\right) \tag{5.11}$$

where $p(a|w)$ is the estimated posterior probability that the word w implies affect state a and $p(a)$ is the estimated prior probability of affect state a in the aGame corpus. The PMI gives positive weight when there is a meaningful semantic dependency and zero when the

word w is unrelated to affective state a. A PMI weight can be negative, indicating that the association between the word w and affect state a is meaningless or random. Since higher PMI weights indicate more interesting semantic associations, only positive PMI (PPMI)[39] values are considered; otherwise the PMI value is zero, indicating an uninformative association between words and affective states, as given by Equation 5.12. It is suggested that PPMI can provide better semantic associations [199].

$$PPMI_{w,a} = \begin{cases} 0 & if\ PMI_{w,a} < 0 \\ PMI_{w,a} & otherwise \end{cases} \tag{5.12}$$

To create the affective lexical dictionary, the raw transcriptions in the aGame corpus were subjected to linguistic processing. The transcriptions were tokenised by extracting the vocabulary then removing stop words (e.g. 'a', 'the', 'and', 'that'), as they do not convey affective meanings. This is a general practice in text-based detection problems. It was also observed that game-oriented words were highly correlated with the games, which implies strong semantic association with the elicited affective states. Therefore, it was essential to remove these words, in order to avoid selecting words related to the game task rather than to the elicited affective state. Once the PPMI for the remaining vocabulary had been estimated, on both the arousal and valence dimensions, the lexical dictionary for each affective state was formed by selecting the words with maximum PPMI weight above a predefined threshold (determined empirically) for that state. Both the linguistic processing and the PPMI estimation were done using the 'leave one player out' strategy, where an affective dictionary for each player was created from the speech transcriptions of all other players. Thus, fourteen dictionaries were created, one for each test player. Table 5.4 lists examples of words having maximal association with the affective states.

[39] PPMI is a variant of PMI where negative values are replaced with zeros.

Table 5.4: List of words in the aGame corpus that maxiamully associated with affective states on both the arousal and valence dimention.

	N	PPMI	HA/PV	PPMI	LA/NV	PPMI
Arousal	Fairly	1.69	Fantastic	0.87	Sorry	0.92
	Pretty	1.40	Jesus	0.87	Hope	0.92
	Nice	1.28	Stop	0.66	Good	0.52
	Sweet	1.18	Crap	0.55	Whoa	0.48
Valence	Fairly	1.69	People	0.94	Jesus	0.86
	Pretty	1.40	Annoying	0.71	Weird	0.86
	Nice	1.28	Crap	0.68	Stop	0.80
	Sweet	1.18	Challenging	0.65	Frustrating	0.68

Creating a lexical dictionary using the PPMI measure can be viewed as a feature selection process by which words with low weight are removed. An alternative approach to creating the lexical dictionary is the term frequency-inverse document frequency (TF.IDF) [199] (see Section 5.3.2), which, in common with PPMI, assumes that words occurring more frequently in a given context than in other contexts have high information content related to that context. However, it has been shown that PPMI is better at measuring semantic information between words and contexts [199], and it has been successfully applied to affect detection [106, 107]. Hence, it was used in this work to create the affective lexical dictionary from the aGame corpus.

5.3.2 Lexical feature representation

Machine learning algorithms such as SVM naturally do not take vocabulary items as input features. Hence, to detect affects from linguistic content, lexical features must be captured. Each speech transcription t can be represented by a lexical feature vector $f_t = \{r_1, r_2, \ldots, r_n\}$, where n is the size of a predefined vocabulary V and r_i denotes the lexical representation of the i^{th} vocabulary item occurring in t. This is called vector space modelling (VSM), which was first developed to represent linguistic content in information retrieval systems [200] and has since been used successfully in text classification/semantic processing applications [199]. The type of VSM most widely employed in lexical-based ADS is the BOW model [53], which represents the linguistic content in a numerical feature space using various values. The BOW representations that were investigated in this study were binary vectors,

term frequency (TF) and TF.IDF. The binary vector consists of ones or zeroes to indicate the occurrence or absence of words from the predefined vocabulary in each speech transcription t. The TF vector takes the normalised word counts as values, instead of ones and zeroes. The TF.IDF value is calculated by weighting the TF with the inverse document frequency (IDF), as given by Equation 5.13. The IDF is calculated as the log ratio of the number of all speech transcriptions N in the training data to the document frequency DF_w, which denotes the number speech transcriptions that contain a given word w. It has been shown that TF.IDF weighting outperforms TF in information retrieval applications [199].

$$TF.IDF_{w,t} = TF_{w,t} * IDF_w = TF_{w,t} * \log(\frac{N}{DF_w}) \quad (5.13)$$

BOW representations are high dimensional, depending on the size of the vocabulary, and most elements have zero values for a given speech transcription. Another low dimensional approach to lexical feature representation that has been shown to outperform BOW representations is based on pairwise mutual information [107], (PPMI in this study). In this approach, the feature space is represented by the average of the PPMI values of the predefined vocabulary that occurs in the speech transcription for each affective state. For a given speech transcription t= $\{w_1, w_2, \ldots, w_n\}$ and affective state a, the $PPMI_{a,t}$ is calculated as given by Equation 5.14, where the PPMI value of each word occurring in the vocabulary is estimated from the training data.

$$PPMI_{a,t} = \frac{1}{n}\sum_{i=1}^{n} PPMI_{w_i} \quad , where\ w_i \in V \quad (5.14)$$

5.3.3 Experimental settings

The ADS for detecting the players' affective experience, see Section 4.2, based on the lexical content of the aGame corpus, using different representations, was developed using an SVM classifier with RBF kernel (Chapter 4, Section 4.3.1). The evaluation was conducted using 'leave one player out' cross-validation, as explained in Section 4.3.4. For each iteration, an affective lexical dictionary (Section 5.3.1) was used to extract the lexical

features (described in Section 5.3.2) from players' manually transcribed speech[40] in both the training and testing sets. The SVM parameters (C , γ) are optimized over ten iterations on the training data by grid search over the range $[2^{-5}, 2^{15}]$. The result is reported in terms of detection accuracy averaged over players.

5.3.4 Results

Table 5.5 and Figure 5.9 present the affect detection performance results, based on different representations of the lexical content in the aGame corpus on the arousal and valence dimensions. From the results, lexical features performed better on the valence dimension and that the PPMI-based lexical representations clearly outperformed the BOW representations, yielding 56.4% and 62.8% accuracy on the arousal and valence dimensions respectively. This feature considers the PPMI weight contributed by all vocabulary items appearing in the speech transcription. The results suggest that spoken words in the aGame corpus have less predictive power on the arousal than the valence dimension. Furthermore, lexical features provide better results than acoustic features (Section 5.2.5) on the valence dimension, which is in line with previous findings (e.g. [13, 90]). However, the lexical data performed less well than the best eye movement features on both dimensions (Section 4.4.5).

Table 5.5: Affect detection results using different lexical features. PPMI feature performed best on both dimensions.

Features	**% Detection accuracy**	
	Arousal	**Valence**
BOW-Binary	46.2	52.6
BOW-TF	47.4	53.8
BOW-TF.IDF	50.0	56.4
PPMI	**56.4**	**62.8**

[40] Speech transcription here refers to the aggregated transcription of all speech segments during the whole duration of the game.

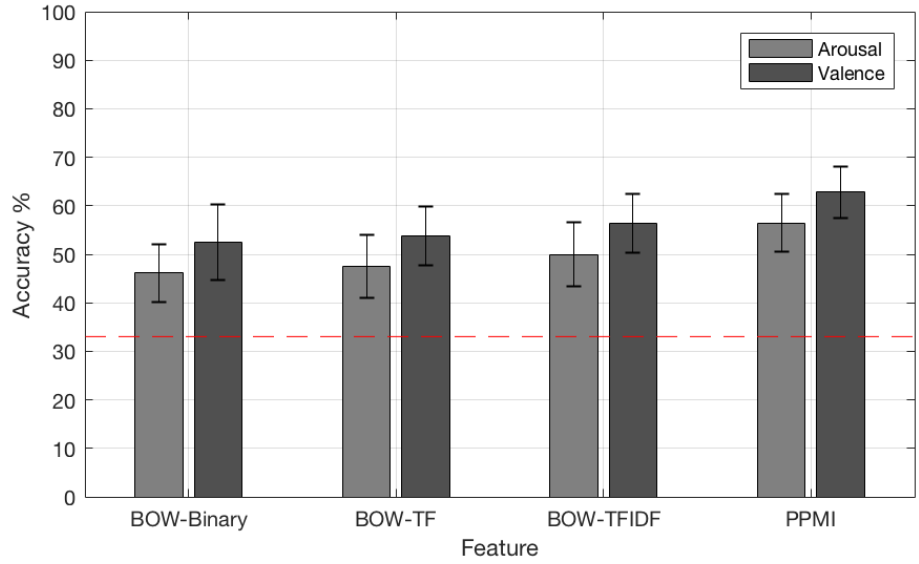


Figure 5.9: Bar graph of detection accuracies on the arousal and valence dimensions using different lexical features, with 95% confidence intervals (n=14). All features provided better accuracy than the random level of 33.33% (straight line in red).

Overall, the results show that the linguistic content of the aGame corpus carries information about the affective states that players experienced on both dimensions. In the real world, however, an accurate transcription of players' speech is not available and affect prediction from spoken words would have to rely on ASR output as an alternative to manual transcription. Several studies have explored the use of ASR output to detect affects from spoken words and have recorded a poorer detection performance (e.g. [90, 107, 109, 112, 114]). This is generally because of the poor performance of the ASR systems during affective speech. Hence, it is believed that the absence of accurately transcribed speech in real-life HCI applications makes it impractical to use lexical information to detect affects. For example, using affective information predicted based on spoken words to improve ASR system performance during affective speech would result in a close loop as the transcription has to be obtained from ASR system first. However, as spoken words in the aGame corpus carry affective information, this data can be exploited to model affective language in an ASR system and improve its performance accordingly, as discussed in Chapter 6.

The following section explains how information from both eye movements and speech modalities was combined to enhance the detection of players' experienced affects. Information based on lexical content was not included, on the assumption, discussed above, that in the real world accurate transcription would not be available for use in the ADS; the

goal was to use as much available information as possible in the affect detection task and for use later to develop affect-robust ASR system, without depending on unreliably transcribed ASR output.

5.4 Multimodal-based Affect detection system

According to the results reported in Section 5.2.5 and in Chapter 4, Section 4.4.5, it can be less reliable to detect affective states from spontaneous speech than from eye movements, because individuals can manipulate and control their speech more readily than the physiological responses reflected in their eye movements. However, since eye movement and speech signals provide affective information from two different sources, the ANS and the vocal tract, it is hypothesised that they are complementary and that they can be used together multimodally to detect players' affective states more robustly. This section explains how a multimodal affect detection system using eye movement and speech signals was developed and evaluated.

5.4.1 Modality integration methodology

To integrate eye movement and speech modalities, two approaches were taken: feature-level fusion (FLF) and decision-level fusion (DLF) (Figure 5.10). These are the methods most commonly used in multimodal affect detection [4].

In FLF, the multimodal system is built by combining the feature sets from both modalities into a general feature vector. The fused feature set is then subjected to feature selection and detection. In the present work, the feature sets from the respective modalities used in FLF integration were the combined acoustic features and PRHT feature sets on the arousal dimension, and the ZCR and SAC+FIX feature sets on the valence dimension. These were selected on the basis of the results reported in Chapter 4, Section 4.4.5 and in Section 5.2.5.

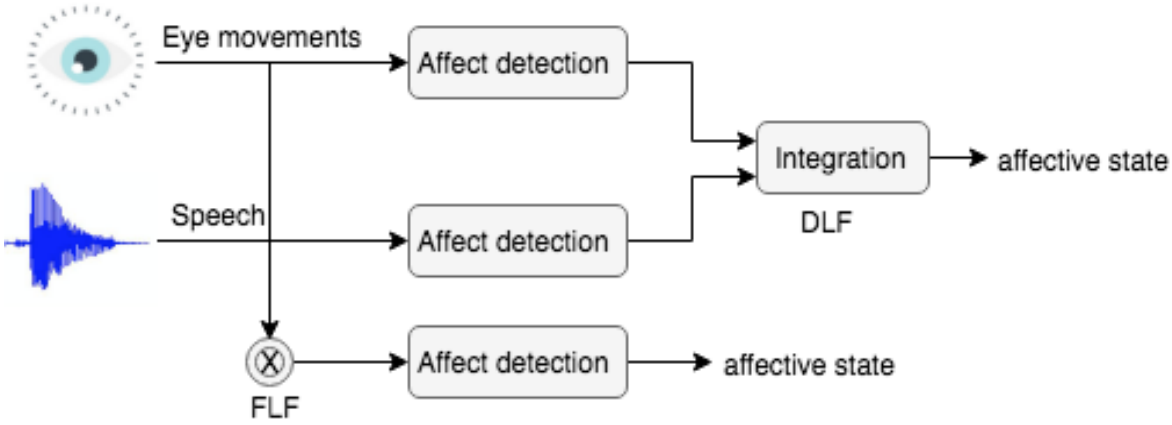


Figure 5.10: Overview of the modality fusion framework using both feature-level (FLF) and decision-level fusion (DLF).

In the DLF approach, detection is performed separately on each modality, then the output scores from the unimodal systems are combined to obtain the final detection results. Accordingly, the scores of the ADS system which performed best on each modality were selected for integration (Section 5.2.5 and Chapter 4, Section 4.4.5). The sum rule was used as the integration method to combine the scores of the unimodal ADSs, as given by Equation 5.15. It was chosen for its simplicity and successful use in affect detection systems [9, 130]. Other integration methods such as bayesian belief integration, product/maximum rules and weighted sum were also investigated and found not to yield better results.

$$a_{mm} = \sum_{q \in Q} \frac{1}{|Q|} S_q(a, x) \tag{5.15}$$

where a_{mm} is the sum of output scores from the unimodal systems for the affective state a, Q is the ensemble of the unimodal systems used for fusion and $S_q(a, x)$ is the score when detecting a given sample x by unimodal system q. The affective state with the highest a_{mm} score was then selected as the final decision on the player's state. The scores represent the posterior probabilities of the detection system for classifying samples as one of the predefined affective states.

5.4.2 Results

Table 5.6 shows the detection results for the multimodal-based ADSs using each integration approach, along with the relative improvement over the best unimodal ADSs. Figure 5.12 illustrates the confusion matrices for the best unimodal ADSs and the modality fusion approaches, showing the detection accuracies of different states of arousal and valence dimensions.

Table 5.6: Affect detection performance using both eye movement and speech as input modalities, based on feature and decision level fusion approaches, on the arousal and valence dimensions. *Significant improvement (p <0.05) over the best unimodal ADSs.

Modality integration	**FLF**		**DLF**	
	Arousal	**Valence**	**Arousal**	**Valence**
% Detection accuracy	71.3	64.3	85.7*	68.6*
% Relative improvement - Eye	-6.2	-10.0	13.8	-4.0
% Relative improvement - Speech	8.9	46.0	30.8	55.7

For the FLF approach, the multimodal-based ADS achieved superior performance compared to the best ADS results of 65.5% and 44.1% using speech modality alone (Section 5.2.5), yielding 8.9% and 46% relative improvement on the arousal and valence dimensions respectively. When compared with the eye movement modality, the multimodal-based ADS did not outperform the best performing ADS on either of the dimensions (Chapter 4, Section 4.4.5). The observed improvement in performance over the speech modality suggests that eye movement provides complementary information, which demonstrates the importance of using it as an additional modality to speech at feature level to detect affects more robustly.

The integration approach based on the unimodal ADS decisions yielded further improvement compared to feature-level integration. The performance obtained was superior to the best eye movement and speech-based ADS results on the arousal dimension, yielding 85.7% accuracy. This is a relative improvement of 13.8% and 30.8% over using the respective single modalities of eye movement and speech, which indicates that integrating their output scores can better model players' affective states on the arousal dimension and effectively enhance detection performance over the single modality approach. DLF also outperformed the best speech-based ADS result on the valence dimension, yielding a relative improvement of 55.7%. However, it did not outperform the best eye-movement-based ADS.

The comparative performance of the three ADSs on the arousal and valence dimensions is shown in Figure 5.11. The repeated measures ANOVA with Bonferroni correction for pairwise analysis was applied to determine whether the differences in the average performance between the best unimodal (Section 4.4.5 and 5.2.5) and multimodal-based ADSs, using DLF, were statistically significant. Significant differences in average accuracy were found on both the arousal and valence dimensions, $p < 0.05$. The pairwise post-comparison test showed that the improvement in performance between DLF and speech-based ADS on average was significant on both dimensions, $p < 0.05$. Significant improvement was also found on the arousal dimension between decision fusion and eye movement-based ADS utilising the PRHT feature set, $p < 0.05$, which suggests that the two modalities provided complementary information on players' affective states on the arousal dimension. No significant difference was found between eye movement and multimodal-based ADS performance on the valence dimension, however, $p > 0.05$.

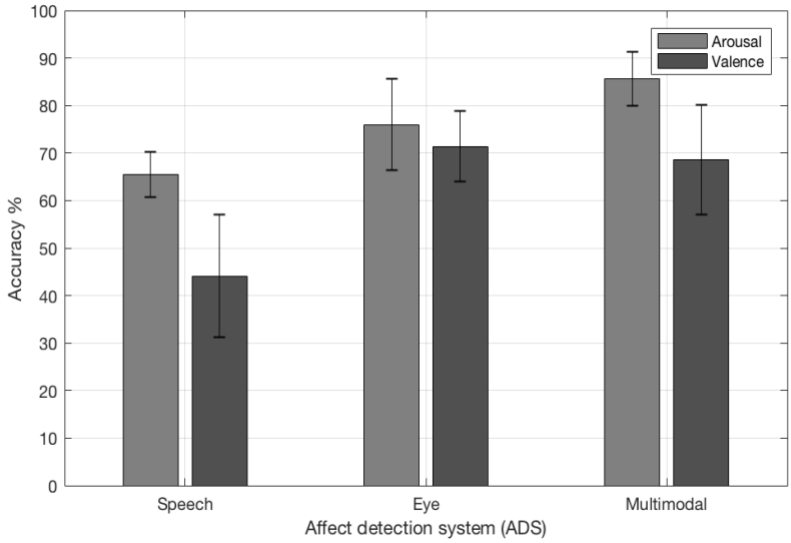


Figure 5.11: A bar plot comparing the best unimodal systems performances with multimodal-based ADS uaing DLF approach, on the arousal and valence dimensions, with 95% confidence intervals (n=14).

The complementary nature of information from the eye movement and speech modalities was investigated further by analysing the confusion matrices shown in Figure 5.12. It can be observed that on the arousal dimension, the HA state was detected very well using the speech modality, at up to 89% accuracy, while the N and LA states were well

detected using the PRHT feature set of the eye movement modality. Supplementing the speech modality with affective information from eye movements at feature level reduced the confusion between LA and the other states, improving detection accuracy to 71.1%. Similarly, the confusion between the N and HA states was reduced, yielding 50% detection accuracy. On the other hand, information from speech contributed to improving the accuracy with which the HA state was detected to 93%. The accuracy of detection of both LA and N states was improved further by decision-level fusion, achieving superior accuracy on all arousal states in comparison to the best unimodal ADSs. This supports the hypothesis that both modalities carry important complementary information regarding the players' affective states on the arousal dimension and that combining them can enhance affect detection performance.

On the valence dimension, the NV and N states were detected with high accuracy, up to 75% and 71% respectively using eye movements, while the PV state was detected with the lowest accuracy in both modalities. Combining information from eye movements helped to reduce the uncertainty in detecting the players' affective states from speech alone and improved the detection accuracy of both NV and PV states, but neither fusion approach improved the accuracy with which the N state was detected. Combining eye movement information with that from speech at decision level improved the accuracy of detection of the NV state by 4%. This gives some insight into the possible complementarity of the two modalities on the valence dimension, as each represents a different source of affective responses: the ANS and the vocal tract.

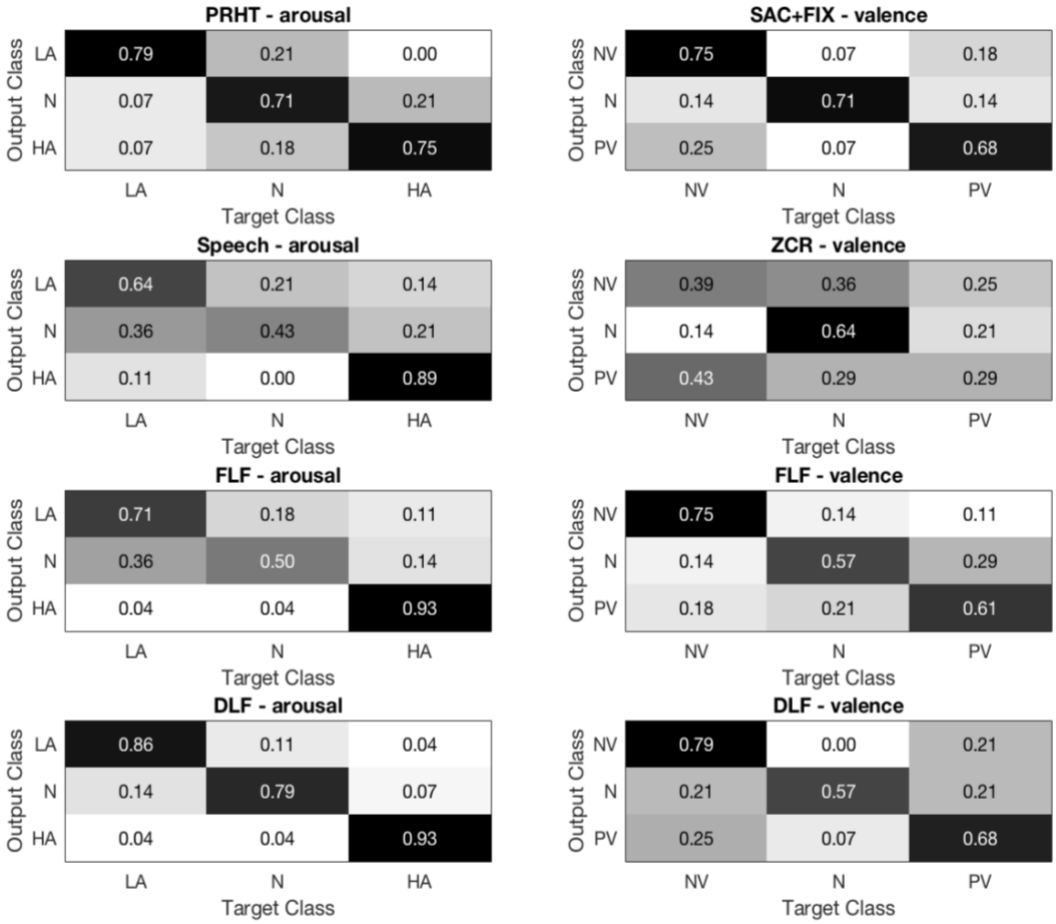


Figure 5.12: Confusion matrices for the best unimodal ADS and different modality integration approaches on the arousal and valence dimensions.

5.5 Summary

The ultimate goal of research in the speech affect detection field is to develop detection systems that can identify the actual affective state experienced by individuals from their speech. Chapter 3 reported on the collection of spontaneous speech data in an interactive gaming environment, using the speak-aloud approach. This chapter has considered the use of this data to infer the affective states experienced by players, by analysing both acoustic and lexical characteristics, that most widely used by studies reported in the literature. It has also investigated the performance improvements arising from the integration of speech and eye movement data in a multimodal system.

The analysis of acoustic features in Section 5.2, shows that the expected affective states were successfully elicited from the players. This result provides some interesting insights

into the effectiveness of evaluating the affective states elicited from the spontaneous speech of players in an interactive virtual gaming environment using the speak-aloud approach. A player-independent ADS based on the acoustic features of speech was developed to detect the affects experienced by players on the arousal and valence dimensions. The results show that information on speech acoustics can be used to detect players' affects on the arousal dimension, with MFCC being the most important feature. However, it did not perform well on the valence dimension. Section 5.3, investigated affect detection based on the lexical content of players' speech. The PPMI measure was used to obtain the lexical content related to the elicited affective states on both dimensions. The results show that the lexical content of the aGame corpus could be used effectively to model players' affective states on the valence dimension, but it did not outperform the acoustic features on the arousal dimension. Although the speech-based ADSs developed here yielded good results on the arousal and valence dimensions by utilising speech acoustic and lexical information respectively, their performance was inferior on both dimensions to the best eye-movement-based ADS reported in Chapter 4.

Section 5.4 explored the integration of the eye movement and speech modalities using feature- and decision-level approaches, in order to develop a more robust ADS on both the arousal and valence dimensions. The results reveal that the two modalities carry important complementary information and suggest that integrating them at decision level can model the players' affective states on the arousal dimension more effectively than either single modality. Moreover, information from eye movements reduced the uncertainty in detecting affects from the speech modality, revealing the potential for using eye movements to aid speech in detecting players' affects multimodally on the valence dimension.

This chapter has highlighted the use of eye movements as an additional modality to speech to enhance the accuracy of affect detection and facilitate the development of robust affect-aware speech-enabled interfaces. Although the aGame corpus of speech data is relatively small, the results open a prospect for developing a large spontaneous speech corpus representing actual experienced affective responses, elicited in an interactive virtual gaming environment, which would be valuable for the HCI and speech recognition communities.

Chapter 6 demonstrates the benefit of using eye movement signals as an additional modality to speech in a speech-enabled interface, by investigating its application to ASR to improve the recognition of affective speech.

Chapter 6

Multimodal Affect-Robust ASR System

The recognition of affective speech is a challenging task and it is well known that affects tend to downgrade the performance of automatic speech recognition (ASR) systems. Previous studies (as discussed in Section 2.3) have attempted to extract affective information from speech in order to detect affects and adapt ASR systems accordingly. However, relying on affective information from the speech modality to improve ASR performance might not be robust enough, especially in the case of spontaneous speech.

The previous two chapters have reported investigations involving the detection of affective states from eye movement and speech modalities that were elicited in an interactive virtual gaming environment. Work on the proposed affect detection systems suggests that combining both modalities is beneficial and that eye movement provides better accuracy than the speech modality, in terms of detecting the affective state experienced by the players. Building on these results, it is believed that using eye movement as additional modality in ASR can enhance the recognition of affective speech. To the author's knowledge, no studies have attempted to use eye movement in ASR systems to enhance the recognition of affective speech. Previous attempts have purely focused on speech modality (e.g. [33-35]).

This chapter investigates the benefits of integrating affective information from eye movement and speech, either unimodally or multimodally, into an ASR system to enhance the recognition of affective speech. It contributes to the previous work in this research area by considering information from eye movements rather than relying on speech modality

only. The following section provides a brief overview of ASR systems and the next describes the proposed integration architecture for building an affect-robust ASR system. Subsequent sections report the various experiments which were conducted to evaluate the proposed system using the aGame corpus.

6.1 ASR system overview

Automatic speech recognition systems enable machines to understand human speech by converting acoustic signals into readable text. This can be seen as a search process, where the goal is to find the word sequence corresponding to a given acoustic signal. A standard ASR system consists of five basic components, as shown in Figure 6.1. The input speech waveform is transformed into a sequence of acoustic feature vectors $Y = \{ y_1, y_2 , \dots . , y_n\}$ during the feature extraction process. The acoustic model provides a probabilistic representation of the acoustic features Y of a given word sequence W= ($w_1 , w_2 , \dots \dots . w_N$). The language model is used to estimate the prior probabilities of the possible word hypotheses during the decoding process. The pronunciation dictionary defines the mapping from a sequence of acoustic features into words. The decoding process uses information from the acoustic model, language model and pronunciation dictionary to search for the most likely word sequence $\widehat{W}$ that matches the given sequence of acoustic features y using Bayesian inference, as defined by Equation 6.1 [201]:

$$p(\widehat{W}|y) = arg\ max_w\ p(\mathrm{y}|W).p(W) \tag{6.1}$$

where the likelihood $p(\mathrm{y}|W)$ of the observed acoustic features is estimated by the acoustic model and the prior probability $p(W)$ is estimated by the language model. The following subsections describe each component briefly. For more details of ASR system can be found in [201-203].

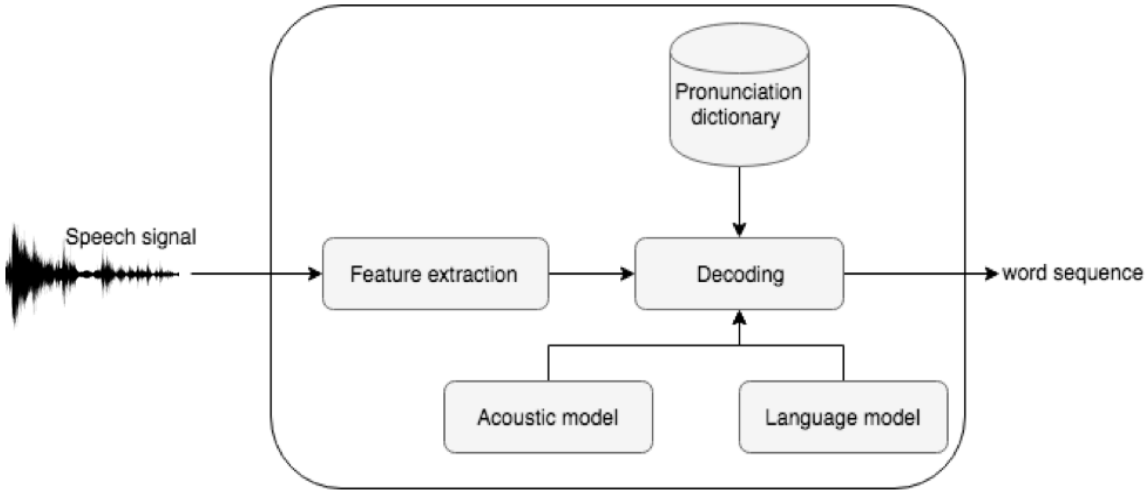


Figure 6.1: Basic framework of an automatic speech recognition system.

6.1.1 Acoustic modelling

Acoustic modelling plays an important part in improving the accuracy of recognition and is the core component of an ASR system. The underlying structure of the speech signal is modelled by a hidden Markov model (HMM) [30], where the sequence of word units (phones) represents the hidden state and the observed acoustic features represent the visible state, which is dependent on the hidden state. Typically, the HMM represents the time-varying acoustic features of the speech signal and the probability distribution of acoustic features associated with HMM states. These probability distributions are modelled by a Gaussian mixture model (GMM) [202].[41] Currently, a state-of-the-art ASR system employs a Deep neural network (DNN) to model the observed likelihood of the context-dependent HMM states [204]. The temporal structure of speech signals can be modelled sequentially; hence HMM models are considered an efficient framework for the task and are employed in all current speech recognition systems [30, 202].

6.1.2 Language modelling

A language model estimates the prior probability distribution of a given word sequence $W = (w_1, w_2, \ldots\ldots, w_n)$ that occurs in the vocabulary [203]. The best-known language model,

[41] A detail of HMM-based GMM acoustic modelling is provide in Appendix F.

one that has been used successfully in state-of-the-art ASR systems, is the *n*-gram model, which estimates the probability of each word *w* occurring in a sequence by conditioning the word w_i on the preceding *n*-1 words (Equation 6.2) [202, 203]. The concept of the *n*-gram language model is based on the Markov assumption, whereby the prediction of the next word is dependent only on the recent history [203]. Longer word histories tend to produce more accurate predictions. However, the word history *n* usually ranges between two and four words [193, 202], in order to avoid data sparsity problem.

$$p(W) = \prod_{i=1}^{n} p(w_i | w_{i-l+1}, \dots\dots, w_{i-(l-1)}), \text{ where } l < i+n \tag{6.2}$$

The *n*-gram conditional probabilities are determined using maximum likelihood estimation based on counting *n*-gram frequencies in the training data [202]. The *n*-gram probability for a word w_i given the word history $(w_{i-l+1}, \dots\dots, w_{i-1})$ can be estimated using Equation 6.3. In case of bigrams (n = 2), the word probability is conditioned on the immediately preceding word, $p(w_i | w_{i-1})$.

$$p(w_i | w_{i-l+1}, \dots\dots, w_{i-1}) = \frac{Count(w_{i-l+1}, \dots\dots, w_{i-1}, w_i)}{Count(w_{i-l+1}, \dots\dots, w_{i-1})} \tag{6.3}$$

For *n*-grams not seen during training, smoothing and backing-off techniques can be applied for robust estimation [205, 206], whereby probabilities of shorter word histories are used to estimate the probability of unseen *n*-grams. In a bigram back-off language model, for example, infrequently observed bigrams are replaced by their unigram probabilities multiplied by the back-off weight, based on a defined cut-off.

6.1.3 Pronunciation dictionary

The pronunciation dictionary, also called the lexicon, provides pronunciation modelling for the vocabulary items. It consists of a phonetic transcription of each word in the training data. During the training stage, the pronunciation dictionary is used for training the acoustic models on different word units, so called phones. In the recognition process, it is used to map a possible sequence of word units into the corresponding word. Vocabularies of interest

should be modelled in the pronunciation dictionary; otherwise they will not be recognised, even if they are present in the language model.

6.1.4 Feature extraction

In order for speech to be recognised, a feature vector representation of its waveform must be created. MFCCs are standard vector representations for modelling acoustic information in ASR systems [202]. Typically, the speech waveform is first transformed into a sequence of MFCC feature vectors, consisting of 13 MFCCs with their dynamic features (the delta and acceleration coefficients). Capturing the dynamics of MFCCs over time can improve speech recognition performance [207]. Cepstral mean subtraction (CMS) [208] is usually applied to MFCCs in ASR systems for noise compensation and to account for variations in the speech recording environment. It is calculated by simply subtracting the average value of each MFCC feature, computed across all speech frames, from the MFCC feature in each frame.

6.1.5 Decoding

The goal of a speech recognition system is to find the most likely word sequence corresponding to a given speech signal, as in Equation 6.1. This is achieved using the Viterbi decoding algorithm [209]. Such that for a given sequence of acoustic features $Y = \{y_1, y_2, \ldots, y_n\}$ of speech signal, the Viterbi algorithm identifies the optimal HMM state sequence $\hat{q}$ that maximises the joint probability of the observed acoustic feature y and the HMM state sequence Q of a given model M, as given by Equation 6.4. The optimal sequence $\hat{q}$ corresponds to the best word sequence for the speech signal.

$$\hat{q} = \arg max_Q \, p(y, Q|M) \tag{6.4}$$

The decoding process usually generates the first best (1-best) hypothesis for the reference hypothesis. If there is a need for multi-pass processing over the data, it is relatively efficient to generate multiple hypotheses in order to reduce the computational cost of Viterbi decoding [202]. The hypotheses are stored in a word lattice structure. The word lattice is a weighted directed acyclic graph composed of nodes which correspond to time points,

connected by arcs that represent word hypotheses weighted with their acoustic and language model likelihood scores [202, 210]. The word lattice contains all alternative word hypotheses that are possible for a given speech signal. It can be converted into an N-best list, which contains the N most probable word hypotheses ordered according to their acoustic and language model scores. Word lattices can be rescored using higher-order or domain-specific language models in order to improve their accuracy in term of estimating the best word sequence, or to account for in-domain context in the latter case.

The use of acoustic and language model scores during the decoding process can affect the performance of an ASR system significantly [211]. There are two parameters which are important during the decoding process and should be tuned for optimal system performance: the word insertion penalty (WIP) and the language model scaling factor (LMSF).

The WIP is a constant weight added to word probability each time the decoder transits between words to account for the word length dependency of the language model by balancing between word insertion and deletion errors. Adding a small WIP value makes insertion errors more likely and deletion errors less likely to occur, and vice versa. The LMSF multiplies all word probabilities by a constant factor, hence affecting the transition between words. It controls the influence of the language model probabilities on ASR output. A large LMSF increases the effect of the language model on the word sequence. The optimal value for WIP and LMSF can be determined empirically using grid search.

6.2 ASR model adaptation techniques

6.2.1 Acoustic model adaptation techniques

Although acoustic models are typically trained using large amounts of speech data, poor performance can occur when the model does not adequately represent the recognition conditions (e.g. a new speaker or environment). To tackle this problem and yield better performance, model adaptation procedures can be applied to reduce the mismatch between training and recognition conditions. This involves updating the acoustic model parameters to better represent the adaptation data. The most widely used acoustic adaption methods are maximum a posteriori (MAP) adaptation and maximum likelihood linear regression

(MLLR). Hybrid approaches combining both methods (e.g. MLLR followed by MAP) can be used for further performance improvement [139, 140, 193].

Maximum a posteriori adaptation

MAP adaptation [212], also called Bayesian adaptation, updates the existing trained model parameters (Gaussian densities mean μ), in which the probability of the model is maximised given the observed O adaptation data, as given by Equation 6.5:

$$p(\hat{\mu}|O) = arg\ max_{\mu} p(O|\mu)p(\mu) \tag{6.5}$$

where $p(\mu)$ is the prior probability of the existing model mean distribution. In speaker adaptation, a speaker-independent model is used to estimate prior probability. The new mean $\hat{\mu}$ is a function of the weighted sum of the existing trained model and the adaptation data, such that for each HMM state j and the mean μ_{jm} of the GMM component m, the new mean $\widehat{\mu_{jm}}$ is calculated by Equation 6.6 [193]:

$$\widehat{\mu_{jm}} = {}^{N_{jm}}\Big/_{N_{jm}+\tau}\ \overline{\mu_{jm}} + {}^{\tau}\Big/_{N_{jm}+\tau}\ \mu_{jm} \tag{6.6}$$

where τ is a weighting factor for balancing between the prior mean and the maximum likelihood estimate of the observed mean from the adaptation data $\overline{\mu_{jm}}$ and N_{jm} is the probability that the GMM component m of HMM state j occupies that adaptation data. The value of τ typically ranges between two and 20 (determined empirically) [213].

MAP adaptation updates every mean parameter of the model locally and only those parameters observed during adaptation will be updated. Thus, MAP adaptation requires a large amount of adaptation data to achieve the best performance. MAP is typically applied to account for speaker differences between training and testing data [213], where a speaker-independent model is adapted to create a speaker-dependent model, rather than accounting for recording and environmental differences [139]. MAP adaptation can also be used for domain adaptation, where speakers are pooled together in one group [214].

Maximum likelihood linear regression adaptation

MLLR adaptation [215], also called transform-based adaptation, maximises the likelihood of the adaptation data using a set of linear transformations to map the parameters of an existing trained model to a new adapted model. The linear transformation matrices can be computed for both the mean and variance parameters of the model. MLLR updates the trained model means as given by Equation 6.7:

$$\hat{\mu} = W\mu + b \tag{6.7}$$

where W is the linear transformation matrix, μ is the model mean vector and b is a bias vector. A similar formula can be derived for MLLR adaptation of the model variance parameters using the variance linear transformation matrix. The transformation matrix is estimated using the adaptation data such that the likelihood is maximised [213, 215].

The two types of MLLR adaptation are global MLLR and regression MLLR. In global MLLR, all model parameters are transformed using a global linear transformation, whereas in regression MLLR different linear transformations are applied to different subsets of the model parameters that are clustered according to their acoustic similarity, using a binary regression class tree [193]. The regression tree is constructed using the existing trained model and the number of regression classes is determined empirically. Where there is a limited amount of adaptation data, global MLLR can be applied by creating a regression tree with a root node only [193, 215].

While MAP adaptation requires prior knowledge of the trained model parameters to modify them directly, MLLR adaptation estimates the linear transformation matrix from the adaptation data, then applies the transforms to update the trained model parameters. Hence, MLLR adaptation is more effective when the amount of available adaptation data is small compared with MAP adaptation. However, when more adaptation data is available, MAP adaptation tends to perform better in comparison to MLLR, due to the detailed local updates of each component parameter of the model [193]. MLLR adaptation typically performs better to account for differences in the speech recording and environment settings [139]. It can also be used for speaker adaptation [215, 216].

6.2.2 Context-based language modelling adaptation

Context-based language modelling adaptation refers to using information about a given context in order to modify the word probabilities of an existing language model and capture the specific features of that context [217]. Various proposals have been made as to the contextual information to be used in modifying word probabilities in language models and enhance the word estimation. For example, Chueh and Chien [218] present a topic cache language model based on topic information derived from words occurring in the recent history. Their model was used to modify the word probabilities according to topic information that had occurred in cache instead of words, leading to observed improvements in perplexity and word error rate (WER) (Section 6.4.1). Visual information (e.g. focus of attention) has also been utilised to boost the word probabilities that are associated with visual objects [37, 38, 219, 220]. In [221, 222], several speech prosodic and timing features of spoken words were used to modify word probabilities dynamically and improve the prediction of words. Similarly, affective information as inferred from acoustic speech features was used to redistribute word probabilities in the language model [35]. However, perplexity reduction was limited and no improvement in WER was reported (Section 2.3). The technique proposed for integrating affective information into language modelling in this chapter is built on pervious work by considering the association between the spoken word and affects using the PPMI measure (Section 5.3.1) in order to adapt the language model during affective speech.

6.3 Integration of affects into ASR system architecture

Figure 6.2 illustrates the proposed integration approach to building an affect-robust ASR system, showing how affective information from eye movement and speech modalities are used in ASR system to enhance the recognition of affective speech. As has been discussed in Section 2.3, ASR system performance significantly degrades during affective speech. Hence, the aim of this integration is to improve the recognition of affective speech based on the state inferred from eye movement and/or speech modalities. It is hypothesised that using the inferred affective state as prior knowledge can enhance ASR system performance

especially when affective information from eye movements are considered. Accordingly, the most likely word sequence $\widehat{W}$ in Equation 6.1 can be expressed by the following equation:

$$p(\widehat{W}|y,a) = arg\ max_w\ p(y|W,a).p(W|a),\ \ where\ a\ \in A \tag{6.8}$$

where a is the affective state inferred from the proposed affect detection systems described in Chapters 4 and 5, which are (low, neutral and high) on the arousal and (positive, neutral and negative) on the valence dimension. The inferred affective state a influences the ASR system models as follows:

- *Acoustic model adaptation:* adapts the acoustic model of the ASR system according to the detected affective state a from eye movement and/or speech modalities, using the acoustic model adaptation techniques explained in Section 6.2.1.

- *Affect-adaptive language modelling:* models the affective information of spoken words in language model by providing high probabilities for those words related to the affective states (Section 5.3.1). The adapted models are used in the ASR system to estimate spoken word probabilities in response to the affective state a provided by eye movement and/or speech modalities.

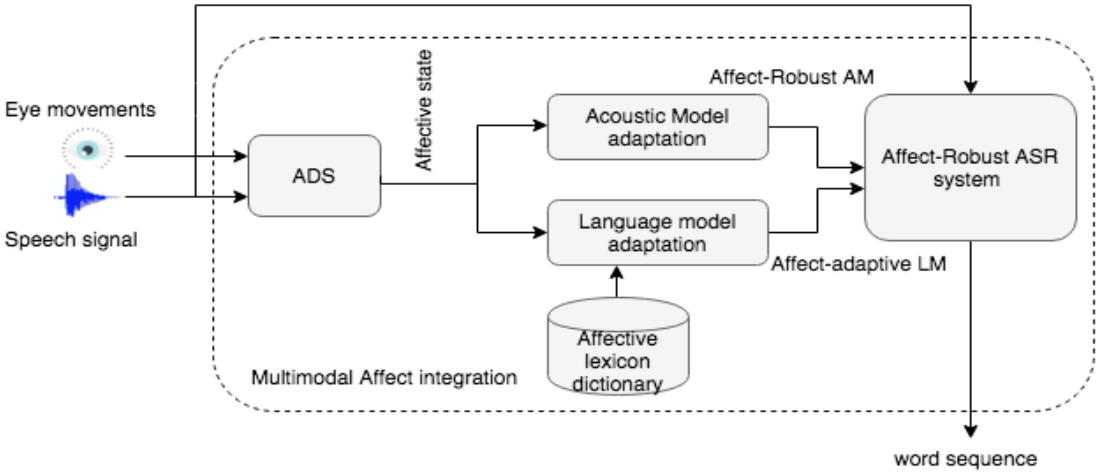


Figure 6.2: Integration of affective information into the speech recognition process. The affect detection system ADS serve as a front-end process for the ASR system.

6.3.1 Affect-adaptive language modelling

As has been discussed in section 5.3.4, the analysis of the linguistic content of the aGame corpus, using PPMI, indicates that players tend to utter certain words during a given experienced affective state and that these words can be used successfully to predict a player's experienced affective state. However, it is discussed that information about spoken words is not available in real-life HCI applications. To be able to detect the experienced affective state from spoken words, the affect detection system must rely as a source of information on the ASR word sequence output, which typically has low word accuracy during affective speech [33, 113]. As the spoken word choices are influenced by the player's affective state, in the aGame corpus, this information can be modelled to enhance the recognition of these words in the utterance, instead of using them for affect detection.

The affective information conveyed by spoken words is incorporated into the ASR system by means of the dynamic marginals adaptation technique [223]. This involve modifying the baseline language model to create affect-adaptive language models (hereinafter ALMs) in which the probabilities are shifted towards words that are related to the corresponding affective state. Such that for each affective state, the ALM is realised by scaling the baseline language model to boost the probability of words that are associated with the corresponding state vocabulary, selected by PPMI (section 5.3.1), as given by Equation 6.9:

$$P_a(w|h) = P_{bs}(w|h) * S(w|a)^{\beta}, \ where\ a \in A\ and\ w \in V^a \tag{6.9}$$

where $P_{bs}(w|h)$ is the estimated probability of w from the baseline language model and h is the preceding word history. As defined in [221, 223], the scaling factor S is measured as the proportion of the affective word probability in a given affective state relative to the word's overall probability in the aGame corpus raised to the power β, given in Equation 6.10:

$$S(w|a) = \left[\frac{p(w|\,A=a)}{p(w)}\right]^{\beta} \tag{6.10}$$

where $p(w)$ is the baseline unigram probability for w in the aGame corpus, $p(w|\ A = a)$ is treated as a dynamic marginal for adapting the baseline language model to the affective state a and β indicates how much the affective state modifies the baseline word probabilities, ranging from 0% (no influence) to 100% (full influence). To modify the baseline language model word probability the following assumptions are made:

1. The probability of words that are related to other affective states should be reduced, because these words have low information content about the target affective state a. This approach is similar to the context-based language modelling proposed by [139].

2. Affective states should not influence common words in the aGame corpus (e.g. stop words and game task-oriented words[42]). Typically, these words are not associated with any affective state, so they should not be modified. Reducing their probabilities may result in an undesirable increase in WER.

The above assumptions are implemented, as given by Equation 6.11, such that if a word belongs to the affective state's vocabulary V^a[43], its probability mass is boosted towards affective state a by β= 99%. For words associated with other affective states, their probability masses are shifted away by $\beta - 1$, that is 1%, in order to avoid reduction in the vocabulary and unrealistic language modelling [139]. Otherwise, affective information has no influence on word probabilities, hence the scaling factor is 1. All word probabilities across the vocabulary are normalised.

$$S(w|a) = \begin{cases} \left[\frac{p(w|\ A=a)}{p(w)}\right]^{0.99} & if\ \mathrm{w} \in V^a \\ 0.01 & if\ \mathrm{w} \in V^{\neg a} \\ 1 & otherwise \end{cases} \tag{6.11}$$

[42] Section 5.3.1
[43] Defined using PPMI as explained in Section 5.3.1.

6.4 Evaluation method

This section discusses the evaluation of the proposed integration architecture for the affect-robust ASR system described in the previous section. The benefit of this integration for the automatic affective speech recognition task is evaluated using speech data of the aGame corpus collected in an interactive virtual gaming environment, as described in Chapter 3. The following subsections describe the evaluation metric and the training corpora, then explain the development of the baseline ASR system and the baseline language model.

6.4.1 Performance evaluation metrics

ASR evaluation metric

Word error rate is the standard metric for evaluating the performance of ASR systems. The word sequence output is compared with the reference hypothesis at word level in terms of deletions, insertions and substitutions, as derived from the Levenshtein distance [224], then WER is calculated as given by Equation 6.12. The WER metric has been used previously to evaluate the use of eye movements in improving ASR system performance (e.g. [37, 38, 220, 225]). In this evaluation, WER is used to compare the performance of the ASR system after the integration of affective information from different ADSs.

$$WER = \frac{D+I+S}{N} \tag{6.12}$$

where *N* is the total number of words in the reference hypothesis, *D* is the number of deleted words, *I* is the number of extra recognised words and *S* is the number of wrongly recognised words.

Language model evaluation metric

Perplexity is a performance metric used for evaluating language models. It measures how well a language model can predict the words in a given word sequence $W = \{w_1, w_2, \ldots w_n\}$. It is computed as the normalised inverse word probability $p(W)$ of the word sequence, as given by Equation 6.13. Perplexity values range between one and infinity.

Perplexity can be used to compare the performance of different language models constructed with the same vocabulary. Lower perplexity ideally indicates that the language model is superior at predicating the word sequence probabilities. However, it may not necessarily indicate an improvement in the performance of the ASR system [226]. Thus, perplexity will be used to evaluate the performance of the ALMs, while their benefit to the ASR system will be compared in term of WER reduction.

$$Perplexity\ (PP)\ =\ p(\mathrm{W})^{-1/N} \tag{6.13}$$

where *N* is the number of words in the test set.

6.4.2 Training corpora

The aGame corpus, as described in Chapter 3, contains a total of 1171 utterances (speech segments), 10882 words and 1049 unique words. Table 6.1 illustrates the distribution of words and the size of the vocabulary across affective states on both arousal and valence dimensions. In order to build a baseline ASR system, supervised training using large corpora is required. The aGame corpus does not provide enough speech data in order to train an ASR system on neutral speech, thus it was necessary to use other existing corpora. In this evaluation, the LibriSpeech corpus [145] was used to train the acoustic model of the baseline ASR system. The speech transcripts of the aGame corpus were used to create the baseline language model and ALMs. Details of the LibriSpeech corpus are given in the next section.

Table 6.1: Word distribution and vocabulary size of the aGame corpus on arousal and valence dimensions.

		Arousal		**Valence**	
	N	**HA**	**LA**	**PV**	**NV**
Number of words	2190	4224	4468	4231	4461
Number of utterances	244	440	487	462	465
Vocabulary size	449	653	673	683	626

6.4.3 LibriSpeech corpus

The LibriSpeech corpus [145] is a large-scale publicly available corpus suitable for general purpose speech recognition systems. It contains 1000 hours of English speech collected from LibriVox[44] recordings, made at a 16 kHz sampling rate. The corpus training set consists of two 'clean' subsets and one 'other' subset. The 'clean' subsets comprise high quality recordings and accents close to the US English accent. The 'other' subset consist of 500 hours of more challenging speech and speakers from different backgrounds. The LibriSpeech corpus also includes a pronunciation dictionary containing 200k words. The pronunciations of most words were derived from the Carnegie Mellon University (CMU) pronunciation dictionary, while the remaining words that do not occur in the CMU dictionary were derived using the Sequitur G2P toolkit [227]. Finally, the corpus includes two test sets ('clean' and 'other') with trigram and 4-gram language models.

Data from the 'other' subset was selected to train the acoustic model of the baseline ASR system, because the speakers in the aGame corpus are from varied backgrounds, although they speak English fluently. In addition, it consists of challenging speech (see Chapter 3).

6.4.4 Baseline ASR system

An ASR system that has been successfully used for decades is the GMM-based system. More recently, DNN has begun to be used in ASR systems because it performs significantly better than conventional ASR systems. However, a DNN-based ASR system requires a large amount of training data. Moreover, there has been no efficient method for adapting an existing system to a new domain with limited in-domain data [228]. Thus, for this evaluation, a GMM-based ASR system was built using the well-established adaptation methods to create affect-robust ASR systems with limited in-domain data.

The baseline ASR system was built using the Kaldi[45] toolkit [229]. Kaldi is an open-source speech recognition toolkit originally developed in 2009 at Johns Hopkins University and freely available to the speech recognition community.[46] Because of its support and

[44] LibrVox is a collection of free public domain audiobooks from volunteer speakers, https://librivox.org/.
[45] Version 5.0 on OS X Yosemite operating system.
[46] http://kaldi-asr.org

development of different speech recognition techniques (e.g. GMM and DNN), it has become a standard tool for ASR researchers.

The implementation of the baseline ASR system was based on the LibriSpeech recipe[47], following a standard maximum likelihood acoustic model training [193], starting with the initialisation of a monophone HMM model and ending with tied-state triphone HMMs with GMM state output densities. The system was trained speaker independently on 10,000 utterances from the 'other' speech subset of the LibriSpeech corpus. The phonetic transcriptions of every word in the aGame corpus, with their multiple pronunciations, were added to the lexicon of the baseline ASR system. The transcriptions were extracted using the CMU dictionary.

Using the Kaldi toolkit, 13 MFCC features[48](with their delta and acceleration coefficients) were extracted from all speech data using a 25 ms hamming window with 15 ms overlap, and CMS normalization was applied to all MFCCs, for use in training the baseline ASR system and for further experiments. The performance of the baseline ASR system on the two LibriSpeech test sets in terms of WER are reported in Table 6.2. Further details of the LibriSpeech corpus and the baseline ASR system can be found in [145].

Table 6.2: Baseline ASR performance on the LibriSpeech test sets.

Test set	% WER
Clean	20.6
Other	41.4

6.4.5 Baseline language model

Estimation of the baseline language model was based on the speech transcription of the aGame corpus using a 'leave one player out' strategy. A bigram language model was built for each player's testing data from the speech transcription of all other players in the corpus. A total of 14 bigram language models were created. Witten-Bell smoothing was applied, as it has been shown to perform well for a relatively small corpus [206]. Due to data sparsity,

[47] http://www.openslr.org/12/

[48] Note that MFCC features used in ADS was extracted using MATLAB. This was done for simplicity and the pipeline was already setup outside Kaldi. Nevertheless, the implementation for extracting MFCCs in both tools was the same.

bigram language models were deemed sufficient for the task. The average perplexity of the baseline language model using all players' speech data with no affective information provided was 91. The language models were built using the SRI Language Modelling toolkit [230].

6.4.6 Lattice rescoring

To evaluate the affect-adaptive language models in the ASR system, the recognition process is performed in two passes, as illustrated in Figure 6.3. During the first pass, the ASR decoding process generates the word lattice, which contains the acoustic and language model likelihood scores for all alternative word hypotheses, using the baseline language model. In the second pass, the output word lattice is rescored with the selected affect-adaptive language model. This is done by subtracting the baseline language model score from each word hypothesis in the word lattice, then adding the affect-adaptive language model scores to the word lattice for rescoring [210]. Rescoring the word lattice is a standard method in ASR research to evaluate the effect of different language modelling on ASR output without the need for the additional computation and processing time of Viterbi decoding.

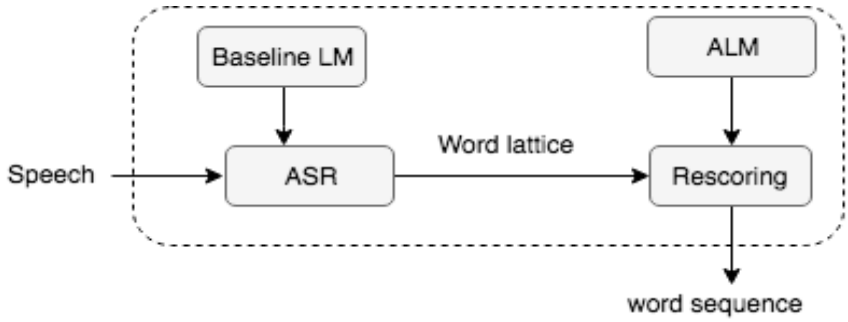


Figure 6.3: Lattice rescoring framework.

6.5 Experiments

Several experiments were performed to investigate the benefit of integrating affective information from the proposed affect detection systems, as described in the two previous chapters, into the acoustic and language models of the baseline ASR system to improve the recognition of affective speech.

Experiment 1: Affect integration performance via acoustic model adaption

The purpose of the first experiment was to demonstrate the effectiveness in integrating affective information using acoustic model adaptation techniques for the automatic affective speech recognition task. The speech data used to train the baseline acoustic model differed from the data in the aGame corpus, which had different speakers and recording characteristics. It was believed that adapting the ASR system to all speech data in the aGame corpus might yield better performance regardless of the players' affective state. Thus, the baseline acoustic model (described in Section 6.4.4) was adapted to the whole aGame corpus and to the speech data for each affective state in both arousal and valence dimensions, separately.

Three acoustic model adaption approaches were investigated in order to determine the optimal adaptation settings for aGame corpus speech data. These were MAP adaptation, regression MLLR and hybrid of both by applying regression MLLR first then MAP adaptation. The adapted acoustic models were tested on each player's speech in the aGame corpus and the experimental results were compared in terms of WER reduction against the baseline ASR system. It was hypothesised that the recognition performance of affective speech would be enhanced when the corresponding adapted acoustic model was selected for the recognition task. The acoustic model adaption was tested and evaluated in the optimal setting using the ground truth labels to provide an optimum upper-bound measure of expected affect-robust ASR system performance. In this experiment, the baseline language model (described in Section 6.4.56.4.5) was used, because the focus was on the influence of affective speech on the acoustic model.

Experiment 2: Perplexity performance evaluation of affect-adaptive language models

The second experiment was performed to determine whether the affective-adaptive language models described in Section 6.3.1 could be used to predict the linguistic content of the aGame corpus in terms of perplexity improvement. An ALM was created for each affective state based on the affective vocabulary associated with that state on both the arousal and valence dimensions, using the 'leave one player out' strategy. Each ALM was tested using the speech transcription for the corresponding affective state. The resultant perplexity is compared against the baseline language model in Section 6.6.2.

Experiment 3: Affect-adaptive language model evaluation in ASR system

The aim of the third experiment was to evaluate the potential benefit of using the ALMs for automatic recognition of affective speech in terms of WER reduction, as perplexity improvement does not guarantee WER reduction [226]. The implementation of the ALMs in the ASR system was performed by late integration via lattice rescoring, as described in Section 6.4.6. For each affective state, the ALM was used to rescore the word lattice generated by the ASR system, implementing the best-adapted acoustic model, identified in Section 6.6.1, corresponding to that affective state and the baseline language model. The comparative performance of the ASR system before and after rescoring the word lattice using ALMs is reported in Section 6.6.3.

Experiment 4: Affect-robust ASR performance evaluation

The final experiment was performed to evaluate the affect-robust ASR system using affective information from the proposed affect detection systems described in the two previous chapters, based either on eye movements, on speech or on both modalities. The automatic affective speech recognition task was performed by selecting the acoustic model (adapted offline) and ALM based on the player's affective state as inferred from the ADSs. The results were based on evaluating the affect-robust ASR system using all of the players' affective speech data. The baseline ASR system adapted to a neutral affective state using the optimal adaptation setting (MLLR then MAP adaptation), resulting from Experiment 1, was used for performance benchmarking. Random affective information was used for lower bound performance estimation, in the form of random uniformly distributed labels generated using the 'randi' function in MATLAB. The performance results were compared in order to determine the best ADS to use in the affect-robust ASR system, in terms of WER reduction.

6.6 Results and discussion

The following subsections report the experimental results for integrating affective information into the ASR system. Regarding acoustic model adaptation, MLLR regression trees were created with 32 terminal nodes (determined empirically). The value for MAP weighting factor τ is determined empirically from values of $\tau = [2, 5, 10, 15, 20]$. The ASR

system performance was optimised by finding empirically the best WER given WIP and LMSF using grid search. The affect-robust ASR system was evaluated by the 'leave one player out' approach, whereby in each iteration one player's data was used for testing and the remaining players' data for adaptation.

6.6.1 Acoustic model adaptation - Experiment 1

The acoustic model adaptation performance results using speech data in the aGame corpus are shown in Table 6.3. They demonstrate that MAP was better than regression MLLR adaptation, providing 38.5% compared to 25.1% relative reduction in the baseline WER. This indicates that there was enough data in the aGame corpus to perform MAP adaptation and account for speaker variations. A further improvement of 44.5% over the baseline was achieved by applying regression MLLR followed by MAP adaptation.

Comparing the best adaptation result of 43.3% WER against the benchmark performance on the LibriSpeech test data as shown in Table 6.2, indicates that adapting the acoustic models clearly reduced the variations between the baseline training data and the recognition of aGame corpus speech data, giving relatively similar result to the LibriSpeech benchmark of 41.4% WER on the 'other' test dataset. This justifies the use of that dataset to train the baseline ASR system. The results are based on adaptation to all aGame corpus speech data with no prior knowledge of the underlying affective state of the players.

Table 6.3: Acoustic model adaptation performance on aGame corpus speech data.

Experiment	**% WER**	**Relative WER reduction**
Baseline	78.0	---
Regression MLLR adaptation	58.40	25.1%
MAP adaptation	48.0	38.5 %
Regression MLLR + MAP adaptation	**43.3**	44.5 %

Table 6.4 and Table 6.5 show the results for acoustic model adaptation to speech data from each affective state in the aGame corpus on the arousal and valence dimensions respectively. The same acoustic model adaptation approaches as described in Section 6.5

were applied to each affective state and the adapted models were tested using the corresponding affective speech data in the aGame corpus. The results show a decrease in WER for all affective speech on both arousal and valence dimensions when the corresponding acoustic model was selected for the affective speech recognition task. The best recognition performance was achieved with regression MLLR and MAP adaptation for all affective states, giving 31.5% for the natural state N, 34.8% for high arousal HA, 37.5% for low arousal LA, 33.1% for positive valence PV and 38.6% for negative valence NV. Interestingly, the adaptation to N yielded the lowest WER, possibly because players' speech was most clearly articulated and contained least noise (e.g. hesitation and laughter).

Table 6.4: Acoustic model adaptation performance, showing the influence of affective speech on the arousal dimension.

Experiment	**% WER**		
	N	**HA**	**LA**
Baseline (no adaptation)	78.8	75.9	77.8
Regression MLLR adaptation	53.6	52.4	54.6
MAP adaptation	43.9	39.7	46.6
Regression MLLR + MAP adaptation	**31.5**	**34.8**	**37.5**
Total relative WER reduction	60.1%	54.1%	51.9%
Adapted to N affective state	---	57.8	59.9

Table 6.5: Acoustic model adaptation performance, showing the influence of affective speech on the valence dimension.

Experiment	**% WER**		
	N	**PV**	**NV**
Baseline (no adaptation)	78.8	76.8	77.38
Regression MLLR adaptation	53.6	51.5	55.44
MAP adaptation	43.9	39.2	46.83
Regression MLLR+ MAP adaptation	**31.5**	**33.1**	**38.6**
Total relative WER reduction	60.0%	56.9%	50.1%
Adapted to N affective state	---	56.5	61.0

Adapting the baseline ASR system to each state of affective speech in the aGame corpus clearly compensated for the variations between the baseline acoustic characteristic and the recognition of affective speech, resulting in an improvement in WER. Using the acoustic

model adapted to the N state to recognise the other states of affective speech, on both the arousal and valence dimensions, significantly degraded the recognition performance. This result, in line with the previous studies [33, 134, 135], demonstrates a potential benefit of integrating the proposed affect detection systems with ASR, where the players' affective state is used as prior information to select the corresponding adapted acoustic model for affective speech recognition. The system adapted to the neutral affective state was used as a benchmark in Experiment 4.

6.6.2 ALM performance – Experiment 2

For each player's recording session, the linguistic content was estimated using the baseline language model and the corresponding ALM. The perplexity results for both language models, presented in Table 6.6, show that on average perplexity was observed to decrease significantly when the ALMs were used for word prediction compared to the baseline language model ($p < 0.05$ - paired sample t-test[49]), for both the arousal and valence dimensions. This suggests that the prediction of the linguistic content in the aGame corpus can be enhanced when the affective information obtained from spoken words are adapted in the language model.

It can be noted that the ALM adapted to the N state yielded relatively similar perplexities of 66.8% and 67.2% ($p > 0.05$) on the arousal and valence dimensions respectively. This is unsurprising, as the selection of neutral affective words did not differ significantly between the two dimensions. It is also interesting to observe that the perplexity reduction resulting from using ALM correlates with the number of corresponding affective words in the aGame corpus. For example, when the language model was adapted to the NV state, a greater reduction in perplexity was observed, with the language containing 32.3% of affective words. This is supported by the finding in [231], suggesting that perplexity reduction is related to the proportion of affective words in the corpus.

[49] Under the assumption of normality.

Table 6.6: Perplexity scores for the adapted ALMs, averaged over players. * Significant improvement ($p < 0.05$) from baseline language model.

	Affects	Baseline LM	ALM	Relative reduction	Affective words
Arousal	**N**	81.6	66.8*	18.1%	18.8%
	HA	99.2	79.0*	20.4%	25.9%
	LA	93.5	74.9*	19.8%	24.1%
Valence	**N**	81.6	67.2*	17.6%	17.4%
	PV	105.7	85.6*	19.2%	20.3%
	NV	87.2	67.9*	22.9%	32.3%

To further investigate whether the observed reduction in perplexity can be attributed to the hypothesis that the linguistic content of the aGame corpus can be better predicted using an ALM adapted to the corresponding affective state, the ALMs adapted to each affective state were evaluated using data from other affective states and the results were compared. Table 6.7 presents the cross-affect results. It can be seen that the ALMs evaluated on data from other affective states yielded lower perplexity than the baseline language model. This result is unsurprising, as the selected affective words are not disjoint. Where affective words that are associated with a certain affective state appear in the test set of another affective state, the ALM that adapted those words will certainly provide more accurate predictions. However, lower perplexity was achieved when the ALMs were used to estimate data from the corresponding affective state. This finding supports the hypothesis and suggests that information about the underlying affective state is required when selecting the corresponding ALM during word prediction.

Table 6.7: Cross-affect testing of affect-adaptive language models on arousal and valence dimensions.

Arousal

	N	**HA**	**LA**
N	66.8	71.9	72.1
HA	88.6	79	88.6
LA	84	83.3	74.9

Valence

	N	**PV**	**NV**
N	67.2	71.6	73
PV	94.3	85. 5	96.1
NV	78.3	78	67.9

6.6.3 Evaluation of ALM in ASR system – Experiment 3

Table 6.8 shows the performance of the ASR system in terms of WER after the integration of ALMs via lattice rescoring. It can be seen that the use of ALMs provides a slight improvement in the ASR performance, in terms of WER reduction, on both the arousal and valence dimensions ($p < 0.05$ - paired sample t-test). These results demonstrate that the recognition of affective speech can be further improved by considering the linguistic aspect of affects throughout language model adaptation. The small reduction in WER is possibly because that the average percentage of affective words in the aGame corpus is relatively low (see Table 6.6). Inspection of the number of affective words correctly predicted before and after rescoring shows that affect-adaptive language models increase the prediction of affective words (Figure 6.4). It also indicates that the ASR system implementing the baseline language model was robust enough to detect most of the affective words. This suggests that affective speech has more influence on acoustic models than on language models in ASR system.

Table 6.8: ASR system performance before and after lattice rescoring with ALMs. *Significant improvement ($p < 0.05$) from the baseline.

%WER		Baseline	ALM	Relative reduction
Arousal	**N**	31.5	30.6*	0.92%
	HA	34.8	33.9*	1.02%
	LA	37.5	36.3*	1.15%
Valence	**N**	31.5	30.4*	1.1%
	PV	33.1	31.9*	1.2%
	NV	38.6	37.4*	1.3%

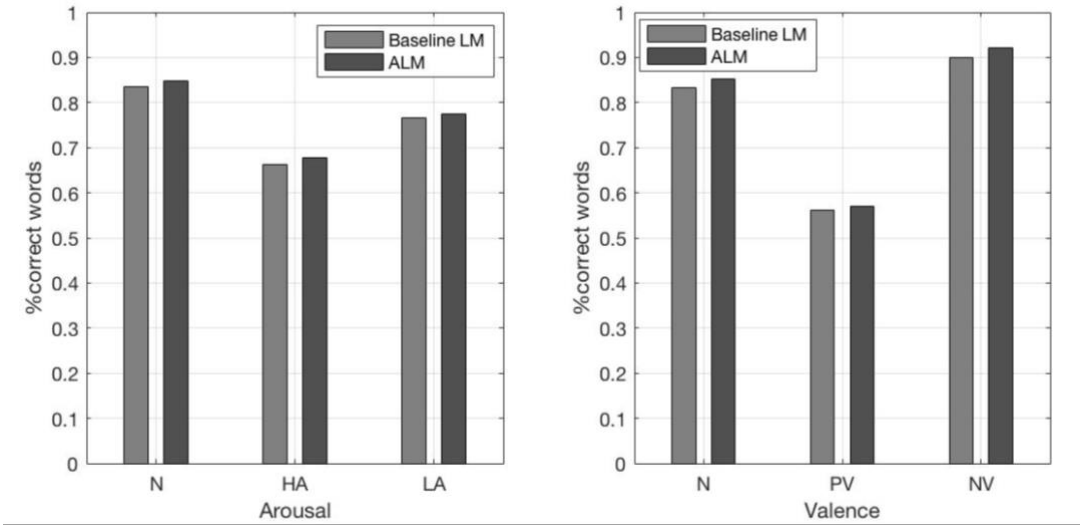


Figure 6.4: Percentage of correct affective words recognised before and after lattice rescoring with ALMs. Rescoring with ALMs increase the prediction of affective words.

6.6.4 Performance of affect-robust ASR-based ADS system – Experiment 4

Table 6.9 and Figure 6.5 present the results of integrating affective information from eye movement and speech unimodally and multimodally, for both arousal and valence dimensions. For each recording session, the adapted acoustic model was selected according to the player's detected affective state. Figure 6.5 shows that the selection of an affective acoustic model based on the affective state inferred from the proposed ADSs yielded a better WER than the baseline. One-way repeated measures ANOVA[50] showed a significant difference between the mean WER of different ASR systems, $p < 0.05$, for both the arousal and valence dimensions.

For arousal, the baseline WER of 54.7% was reduced significantly by 18.3% (relative WER reduction) using affective information from speech, $p < 0.05$. A greater reduction of 25.8% was observed when affective information from eye movement was used, $p < 0.05$, compared to speech-based ADS. Using the multimodal-based ADS yielded better results than either modality alone, with a relative WER reduction of 30.4%, $p < 0.05$. Similar results were achieved on the valence dimension when speech or eye movement-based ADS were used, giving relative WER reductions respectively of 9.6% and 22.9%, $p < 0.05$. On this dimension, multimodal-based ADS significantly improved the recognition of affective

[50] Bonferroni correction was used for pairwise analysis.

speech, giving 20.3% relative WER reduction compared to speech-based ADS, $p < 0.05$. However, there was no significant performance difference between multimodal and eye movement-based ADS, $p > 0.05$. This is simply because affect recognition results on the valence dimension using multimodal-based ADS did not significantly outperform eye movement-based ADS. Nevertheless, using affective information in the ASR system from eye movement as an additional modality outperformed speech modality, in term of reducing WER. This supports the finding reported in Chapter 5: that eye movement provides complementary affective information to speech, especially on the arousal dimension, suggesting that the use of eye movement is beneficial for the automatic affective speech recognition task, as eyes are always in operation. Moreover, it can be clearly observed that the performance improvement of the affect-robust ASR system on the arousal and valence dimensions depends on the accuracy of the ADSs, considered in Chapters 4 and 5. Interestingly, the random system gave poor results close to the baseline, $p > 0.05$. This demonstrates the effectiveness of using the proposed affect detection system in ASR, in order to enhance the recognition of affective speech.

Table 6.9: Affect-robust ASR system WER performance on aGame corpus using different affect detection systems, averaged over players.

Experiment	**Arousal**		**Valence**	
	%WER	**Relative WER reduction**	**%WER**	**Relative WER reduction**
Baseline	54.7	---	54.7	---
Random	54.9	---	55.7	---
ADS-Speech	44.7	18.3%	49.5	9.6%
ADS-Eye	40.6	25.8%	**42.2**	22.9%
ADS-Multimodal	**38.1**	30.4%	43.6	20.3%

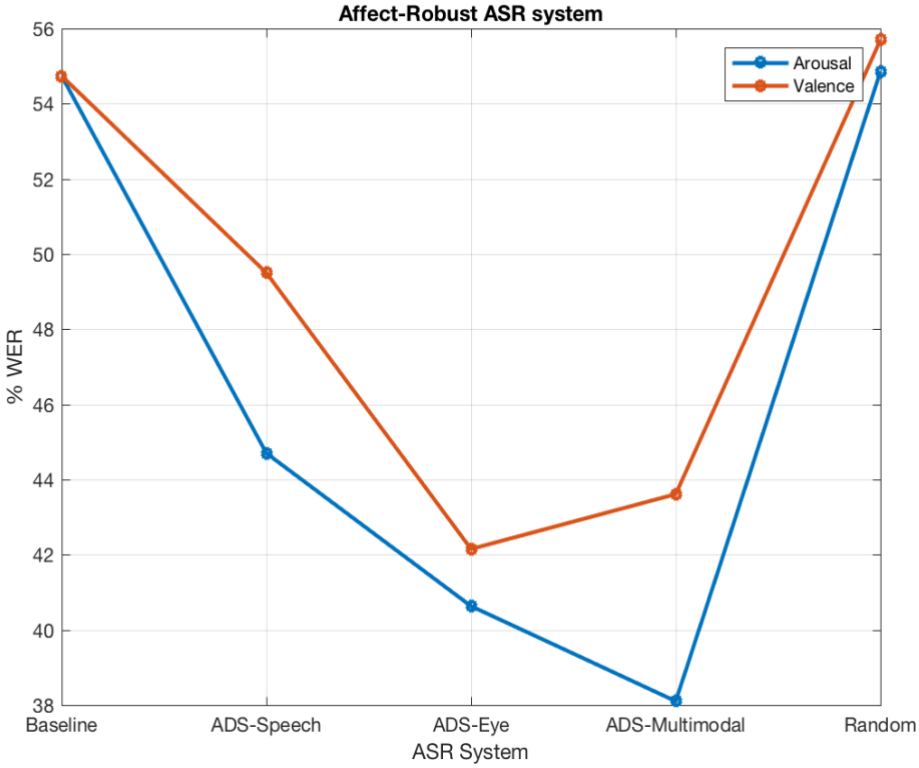


Figure 6.5: WER performance of affect-robust ASR system using affective information from different modalities on arousal and valence dimensions. Incorporating information from eye movement as additional modality to speech reduced the WER on arousal dimension significantly ($p < 0.05$).

Figure 6.6 shows the performance of the affect-robust ASR system for all players in the aGame corpus in terms of WER. Because the baseline system was adapted to the neutral state, it was expected that the system would behave worse than the baseline for some players, where the ADS misclassified that state. For instance, this was observed when speech-based ADS was used to recognise the affective speech of player 1, on both the arousal and valence dimensions. Comparing the use of the multimodal-based ADS for the affect-robust ASR system with the use of eye movements and speech-based ADS, on the arousal dimension, it was observed that the multimodal-based ADS performed better for most players. In cases where the affective information from speech was not accurate, adding information from eye movements lowered the WER (e.g. player 4). These observations confirm that eye movement and speech data are complementary when used in ASR. In some cases (e.g. for player 10 and 13), using multimodal-based ADS did not reduce the WER, because the affective state of those players was detected better unimodally, giving 100% accuracy. It is

also interesting to observe that for some players, the use of multimodal-based ADS provided better WER than unimodal-based ADS on the valence dimension. This indicates that eye movements provide complementary affective information on the valence dimension for some players (e.g. player 3 and 7); however, this did not reach a significant level.

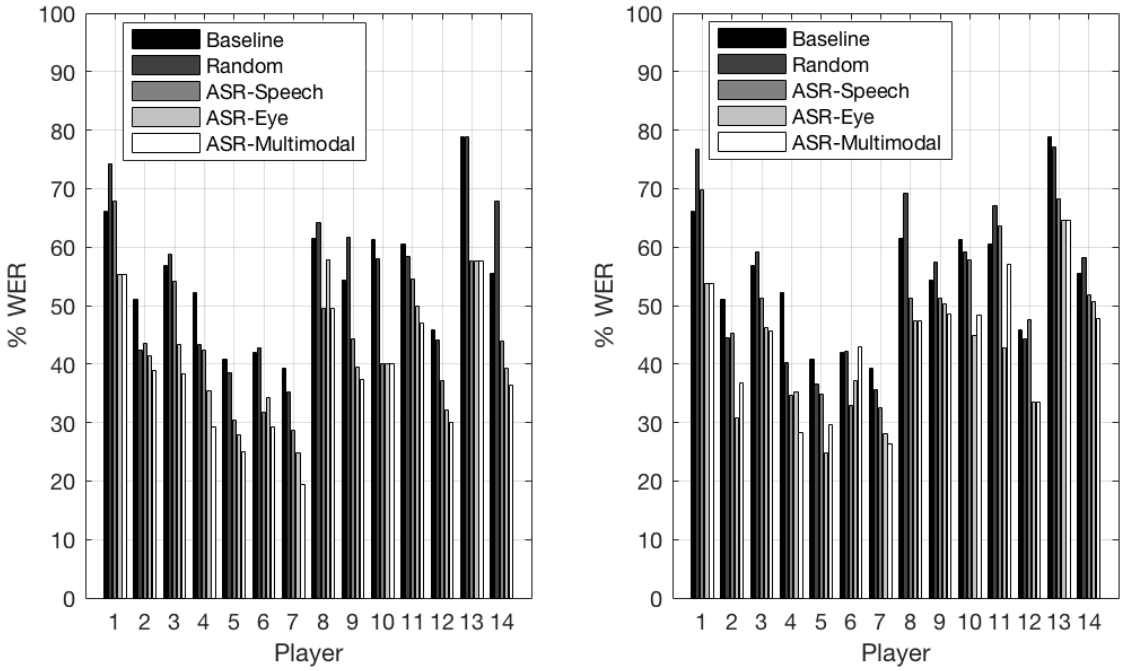


Figure 6.6: Performance of affect-robust ASR system on the arousal (left) and valence (right) dimensions for each player in the aGame corpus.

The results of integrating the ALMs into the affect-robust ASR system by late integration (lattice rescoring) are shown in Table 6.10. On the arousal dimension, the WER was observed to decrease when the ALMs were selected using any of the affect detection systems, with reductions ranging from 1.1% to 1.6%, compared with results before rescoring ($p < 0.05$ - paired sample t-test). Similar results were observed on the valence dimension ($p < 0.05$), with reductions ranging from 1.16% to 1.25%. The exception was when speech-based ADS was used ($p > 0.05$), as it was found to have a poor recognition performance compared to the other ADSs. Using the ALMs did not yield large performance improvements, as discussed in Section 6.6.3; however, it appears that the affect-robust ASR system might be further improved by selecting the ALMs based on the inferred affective state. Using another scheme of adaptation (e.g. a cache-based language model which

contains information about the current affective states) to dynamically rescore the ASR output might possibly lead to a larger performance improvement in terms of WER.

Table 6.10: Affect-robust ASR system WER before and after ALM integration based on ADS output.

	%WER	ADS-Speech	ADS-Eye	ADS-Multimodal
Arousal	Before ALM	44.7	40.6	38.1
	After ALM	43.1	39.3	37.0
Valence	Before ALM	49.5	42.2	43.6
	After ALM	49.5	40.9	42.5

6.7 Summary

This penultimate chapter has investigated the influence of affects on ASR system performance and the integration of affective information inferred from eye movement and speech signals either unimodally or multimodally, as described in Chapters 4 and 5, to improve the automatic recognition of affective speech.

Affective information was incorporated into the ASR system by acoustic model adaptation techniques and word lattice rescoring using the proposed affect-adaptive language models. For acoustic modelling, MAP, MLLR and the combination of both adaptation techniques were investigated to compensate for the impact of affects on acoustic modelling. Affect-adaptive language models were created, based on the observation that the linguistic content of the aGame corpus conveys affective information (Chapter 5). Following this observation, a vocabulary was created for each affective state, based on PPMI. It was used, in this chapter, in language modelling to incorporate the affective information conveyed by the vocabulary. The baseline language model was modified to create affect-adaptive language models that provided higher probabilities for the corresponding vocabulary.

A number of experiments were conducted to evaluate the proposed integration approach on the aGame corpus using the ground truth affective information. A baseline ASR system was developed for this purpose. The results show that the ASR system performed less well when presented with affective speech and that compensating for the influence of affects

using the proposed integration approach was beneficial. The best WER reduction was achieved by applying MLLR and MAP adaptation together. The results for the proposed affect-adaptive language model show significant reductions in perplexity over the baseline language model when it was used to estimate corresponding affective speech transcription. The recognition performance of affect-adaptive language models in the ASR system was measured by rescoring the generated word lattice, yielding a slight improvement in WER performance.

The affect-robust ASR system using the best acoustic model adaptation (MLLR followed by MAP) and the different ADSs were implemented and evaluated on all the players' affective speech. When system performance was compared in terms of WER, a greater improvement was observed using affective information from eye movement-based ADS compared to speech-based ADS. It was concluded that incorporating affective information from the multimodal-based ADS as prior knowledge to ASR system yielded better results than ADS based on either modality alone and that using affective information from eye movement-based ADS was better than speech-based ADS. The results demonstrate the promising benefit of integrating affective information from eye movements as an additional modality to enhance the automatic recognition of affective speech, in terms of WER reduction. This approach would be valuable in real-time systems, as information from eye movement is always available and can be integrated into ASR prior to the spoken utterance.

This work has discussed the integration of affective information that is detected over the whole duration of the game into an ASR system. Incorporating affective information continuously is valuable for real-time systems. Although it would be possible to repeat the experiments to provide affective information continuously, its continuous detection and integration would require further investigation which is beyond the scope of this research.

Chapter 7

Conclusion

Although considerable progress has been made towards the realisation of affect-intelligent machines, they are not yet a reality in everyday life. Accurate affect detection systems suitable for real-life HCI applications remain challenging for want of realistic data representative of individuals' felt affective experience. This thesis has reported research into affect-aware interfaces applicable for use in real-life applications to facilitate natural HCI through the development of affect detection systems utilising eye movement and speech data collected in an interactive virtual gaming environment. It is believed that in such an environment it is possible to elicit spontaneous affective responses close to those of real life, hence facilitating the evaluation of affect detection systems with generalised performance.

By analysing the use of eye movement and speech modalities in affect detection, the thesis concludes that spontaneous speech elicited while interacting with the virtual gaming environment via speaking aloud can provide information about the player's experienced affects and that the eye movement modality provides more reliable affective information across both the arousal and valence dimensions (research question 1 and 2 in Chapter 1, Section 1.2). Incorporating eye movement data with speech in multimodal fashion shows that the two modalities carry complementary affective information in the application of interactive virtual gaming environment and that both warrant incorporation where feasible (research question 3 in Chapter 1, Section 1.2). Based on the findings, this research has developed a framework for evaluating the use of affective information from eye movements as prior knowledge in ASR systems to enhance the recognition of affective speech and

achieve affect-robustness (research question 4 in Chapter 1, Section 1.2). Overall, the results show the value of considering eye movement as an additional modality to aid affect detection from speech, thus facilitating the development of an affect-aware speech-enabled interface without compromising the spontaneity of the interaction. This chapter enumerates the contributions made by the present research and offers recommendations for further research.

7.1 Contributions

The work described throughout this thesis makes original contributions to the general field of HCI and specifically to the area of affective computing. This section summarises these contributions.

7.1.1 The aGame corpus

This thesis has reported the development of a spontaneous multimodal affective corpus of eye movements and speech which was collected in an interactive virtual gaming environment designed to elicit different affective states according to Russell's circumplex model of affect. The participants, each having more than two years of gaming experience using a PC and/or gaming consoles, were invited to speak aloud while playing, thus eliciting natural affective speech responses. Eye movement data was captured using the EyeLink II sensor, a head-mounted eye tracker. Speech was simultaneously recorded using a studio quality microphone, then transcribed and reviewed manually by two independent individuals and time-aligned automatically. The illumination of the environment was recorded to compensate for PLR in the pupillary responses. On-screen activities were also recorded to allow visual inspection. All recorded data was subject to quality assessment. The software and hardware used for data collection allowed the modalities to be recorded synchronously.

Analysis of the players' self-reported affective experience confirmed that the interactive virtual gaming environment elicited the target affects as expected, according to the ground truth. This corpus is the first to contain recordings of affective eye movements and speech data elicited in a less constrained (spontaneous) environment, namely the interactive virtual gaming environment. It emphasises the need for the evaluation of multimodal affect

detection systems applicable to real-life natural HCI applications. The corpus can be used by the affective computing and speech recognition communities.

7.1.2 Eye movements-based affect detection

This work has demonstrated the feasibility of taking a less intrusive approach-based on eye-movement to affect detection, using data collected in interactive virtual gaming environments. Various features were extracted from pupillary response with luminosity compensation, fixation, saccade and blinking, and new pupillary response features based on the Hilbert transform were proposed for affect detection. Features of the elicited eye movements were analysed to determine how they varied across different affective states on the arousal and valence dimensions. The results confirm that playing in the virtual interactive gaming environment successfully elicited the expected affective states. An SVM classifier was employed to detect these affects from the eye movement features and their performance was compared in terms of detection accuracy. It was found that pupillary response with Hilbert transform features yielded a better performance on the arousal dimension, whereas using saccade and fixation features together gave a superior performance on the valence dimension. These results highlight the potential for the real-life application of affect-aware interfaces based on sensing eye movements.

7.1.3 Speech-based affect detection and integration with eye movements

An affect detection system was developed, using acoustic signals and the lexical content of speech collected by prompting players to speak aloud while playing in the interactive virtual gaming environment. Various acoustic signals were extracted for affect detection, including prosodic, voice quality and spectral features. Affect elicitation analysis shows that affective responses were successfully elicited from the players' speech. Affect detection was performed using an SVM classifier on acoustic features. The results demonstrate that it was easier to detect arousal than valence states from the acoustic speech signal. As to the lexical content, an affective lexical dictionary was developed by measuring the association of the spoken word with the elicited affects on the arousal and valence dimensions, using the 'affective salience' measure based on pairwise mutual information. Various lexical features

of the affective lexical dictionary were extracted from transcribed speech. The results show that lexical features yielded a better performance when used to detect players' affective states on the valence dimension. Results derived from both acoustic signals and the lexical content of the aGame corpus reveal that affect is detected more effectively from eye movement data than speech data. Combining affective information from eye movements with acoustic speech information in a multimodal approach improved the detection performance, indicating that the two modalities carry complementary information and that both warrant incorporation where feasible.

7.1.4 A multimodal affect-robust ASR system

Following the development of ADSs based on eye movement and speech data, unimodally and multimodally, the study investigated their application to an ASR system in order to enhance its capabilities during affective speech. An affect-robust ASR system was thus constructed by incorporating affective information from these two modalities using acoustic and language model adaptations. Affect-adaptive language models were developed to model the affective information of spoken words in the affective lexical dictionary. A research-level baseline ASR system was developed using the Kaldi toolkit to investigate the influence of affects on the system's performance in term of WER and to evaluate the use of the ADSs developed earlier in producing an affect-robust ASR. MLLR and MAP acoustic model adaptations were used to demonstrate the effectiveness of incorporating affective information from the ADSs at the optimal settings. The performance of the affect-adaptive language models was evaluated in terms of perplexity and WER when used in the ASR, in the latter case via word lattice rescoring. The complete affect-robust ASR system was evaluated on all of the players' affective speech by the integration of the ADSs unimodally and multimodally, and performances were reported in terms of WER. The results demonstrate the benefit of using eye movements as an additional modality to enhance the affect-robustness of an ASR, thus demonstrating the feasibility of affect-aware speech-enabled interfaces for natural HCI.

7.2 Recommendations for Future Work

The work presented in this thesis contributes towards facilitating affect-aware interfaces in a natural HCI by the use of eye movement as an additional modality to speech to enhance the affect detection process and demonstrates the value of its application to ASR systems for affect-robustness. There is nonetheless a need for further research before affect-awareness become a reality in natural HCI. This section offers some suggestion for future work.

- The development and evaluation of affect detection systems reported in this thesis were based on players' affective experience over the whole duration of each game, assuming that the same affective state persists throughout a game session. However, affective experience can change over the course of a game in response to different game events. The development of continuous affect detection is thus of interest for adaptive HCI and this requires the establishment of a method for continuous determination of ground truth to be investigated.
- It would be beneficial to develop the multimodal integration of eye movement and speech modalities based on their mutual relationship with the elicited affects. For example, by applying the information-theoretic concept of mutual information one can investigate the amount of information they convey about the elicited affective states. It would also be of interest to investigate their shared representation upon affects using deep learning techniques.
- This work employed a wearable eye movement tracking sensor, EyeLink II, for its ability to track pupillary responses with high precision. However, it can be invasive and uncomfortable for some players. Advances in the development of remote eye movement tracking sensors offer the potential for players to interact more freely within the virtual gaming environment without being restricted and without compromising naturalness of interaction. It would be interesting to evaluate the proposed approach using a remote eye movement tracking sensor.
- This work investigated the affect-robustness of a GMM-based ASR system, but state-of-the-art DNN-based ASR systems have been shown to provide superior recognition performance and can better account for variability in speech. It would therefore be of interest to investigate to what degree DNN-based ASR systems can

recognise affective speech and whether there is any benefit in using the proposed approach for affect-robustness.

- This study has investigated the benefits of using eye movements as an additional modality to speech in developing an affect-robust ASR system. Since information from the eye movement modality is always available, considering its integration into ASR prior to speech is of interest for real-time interaction.
- During the course of the work reported in this thesis, the deep learning approach to affect detection has emerged with promising performance. Investigating this approach to affect detection using data from the aGame corpus would be of great interest. However, the corpus is relatively small because it was difficult to recruit participants. One could consider collecting a larger corpus by means of crowdsourcing, in an online interactive virtual gaming context, for example.
- The possible areas in which this research could be applied are:
    - Speech-enabled interfaces.
    - Adaptive experiences in virtual reality and gaming.
    - Affect-aware multimodal human-computer interfaces.
    - Natural human-robot interaction.
    - YouTube transcription service for gamers.

7.3 Summary

This thesis has presented our work towards developing an affect detection system utilising spontaneous affective responses from eye movements and speech modalities elicited in an interactive virtual gaming environment. A framework is proposed to address the issue of affects in ASR systems using the ADSs developed here, highlighting the importance of using a more accurate modality for affect-aware speech-enabled interfaces which does not compromise natural interaction.

Appendix A

Overview of available affect corpora

Table A.1 provides a list of affective corpora that contain speech or eye movements data found in the literature. Early examples such as SUSAS [232] comprise a wide variety of affective and stressed speech collected in both acted and spontaneous real-life situations, such as roller coaster rides. Affective speech corpora may contain material in English or other languages, such as German in the case of the FAU Aibo [69] corpus and the Berlin affective speech dataset (EmoDB) [7]. The FAU Aibo corpus contains speech recordings of children interacting with a robot dog by giving him commands to perform some tasks, a Woz scenario, while. EmoDB contains acted affective speech data portrayed by actors. Affective speech data are mainly obtained from actors, from real-life situations, or by using Woz scenarios. In the last two cases the data are typically annotated by external observers.

Early corpora contained affective data from a single modality, mainly speech. In recent years, affective computing researchers have shown increasing interest in more complex multimodal data to improve recognition performance. The majority of available multimodal affective corpora are captured from audio and visual modalities, an example being the IEMOCAP corpus [138], which was developed to further the understanding of human multimodal affective behaviour. The visual data comprises facial expressions and/or bodily gestures. In addition to audio-visual corpora, a few affective multimodal physiological corpora are also available, including data such as EEG and eye movement data (e.g. [130]). The MAHNOBA-HCI corpus [129] captures both audio-visual and physiological data, including eye movements, EEG and peripheral signals[51] in response to affective movie clips.

[51] Peripheral signals include: GSR, ECG, breathing rate, and skin temperature.

However, the audio data consists only of laughter and the amount is insufficient for affect analysis. For a comprehensive review of available affective corpora, the reader should refer to [4].

Table A.1: A list of available affective corpora.

Corpus	Type	Method	Affects	Modaility	Brief description
SUSAS [232]	Spontaneous/acted Anotated	Real life and acted situations	Stress, fear, anxiety, anger	Speech	Speech under simulated and actual stress.
EMO-DB [7]	Acted	Reading sentences	Neutral, anger, fear, joy, sadness, disgust and boredom.	Speech	Actors simulating affects while reading sentences.
Lee and Narayanan [8]	spontaneous Anotated	Call center	Negative and non-negative affects.	Speech	Speech obtained from customers engaging in a converstation with an agent over telephone, frustrated customer was expected to show negative affects.
FAU Aibo [69]	Spontaneous Anotated	Robot interaction Woz	Neutral, joyful, surprised, emphatic angry, motherese and bored.	Speech	Children interacting with a robot dog.
VAM [59]	spontaneous Anotated	TV talk show	Arousal, valence, dominance	Speech/Visual	German show “Vera am Mittag”.
Belfast SAL [67]	Spontaneous Anotated	Woz	Arousal and valence	Speech/Visual	Data obtained from indiviudals engaging in conversations with Sensitive Artificial Listener (SAL) interface designed to evoke specific affective states.
SmartKom (SmK) [68]	Spontaneous Anotated	Woz	Neutral, joy, anger, helplessness, pondering and surprise	Speech/Visual	Collected mainly for training artificial agent as a tourist information guide and also used for affect detection.
IEMOCAP [138]	Acted Anotated	Dyadic interaction (scripted scenarios)	Affective keywords: anger, happiness, sadness, and neutrality, Dimensional affects: arousal valence and dominance	Speech/Visual	Developed to aid understanding of human multimodal affective behaviour.
RECOLA [117]	Spontaneous Anotated	Dyadic interaction (Remote Collaborative task)	Arousal, and valence	Speech/Visual ECG and electro-dermal activity (EDA)	Developed to aid understanding of human multimodal social affective behaviour in the context of computer collaborative task.
MAHNOB-HCI [129]	Spontaneous	Video	Arousal, valence, and dominance Affective keywords: neutral disgust, amusement, joy, fear, and sadness.	Speech/visual, EEG, eye movements and Peripheral signals	Subjects watch movie clips and self-report their affective expreince.
Lu et al. [130]	Spontaneous	Video	Valence	EEG and eye movements	Subjects watch movie clips and self-report their affective expreince.

Appendix B

CONSENT FORM

Dear participant,

You are invited to take part in a research that aims to improve multimodal interfaces by utilising eye movement data as additional input modality to speech, in order to better interpret user's behavior. Before you agree to take part in the research experiment you must read the information sheet as it contains information about the experiment and will help you to decide whether or not to participate.

Your participation in this research is voluntary; however, upon completion of the experiment you will receive a £10 iTunes voucher as appreciation for giving us your time. You may choose to refuse to participate or withdraw from the experiment at any time. If you do so, there will be no influence on your relationship with the EESE school and all your data will be destroyed and not be used in any part of the study.

By signing below, I declare that I have been given information about the experiment. I discussed the research with Ashwaq who is conducting this research as part of a PhD degree supervised by Dr Neil Cooke in the EESE school at the University of Birmingham. I am indicating my consent to participate in the research. I understand that the data collected from my participation will be dealt with confidentially and used primarily for the PhD thesis, and will also be used for conference and journal publications, and I consent for it to be used in that manner.

Signature: .....

Date: ...../...../.....

Given DI: .....

Appendix C

Questionnaire

Figure C.1 depicts the questionnaire form that appears on completion of each game. The questionnaire consists of two sections, the first (Figure C.1(a)) evaluates the player affective experience in terms of valence and arousal on a seven-point scale (range [-3, 3]). Positive values indicate more aroused/pleasure affective experience, and negative values indicate more sleepy (low aroused)/displeasure affective experience. The second lists eight discrete affects (relaxed, content, happy, excited, angry/annoyed, afraid, sad and bored), as shown in (Figure C.1(b)), and asks the player to choose the affective label closest to his/her experience during the game.



(a) (b)

Figure C.1: A screenshot of the questionnaire, from the interactive virtual gaming environment developed by [10].

Appendix D

Equipment

Figures D.1 to D.5 show the equipment used in collecting the recordings for the aGame corpus.



Figure D.1: Eyelink II eye movement tracking sensor, adapted from [27].



Figure D.2: Shure WH20 microphone.



Figure D.3: Edirol UA-25 speech recorder.



Figure D.4: Logitech Wingman 3D force feedback joystick.



Figure D.5: Extech HD45 Luxmeter.

Appendix E

The Hilbert transform

The Hilbert transform (HT) is a signal processing tool for nonlinear and non-stationary signals. It acts as a filter that shifts the phases of a given signal in all frequency components by $\pm\pi/2$ radians and returns a complex representation, as given by Equation E.1, called an analytic signal [176].

$$z(t) = x(t) + j\tilde{x}(t) \tag{E.1}$$

where $x(t)$ is the real part and $j\tilde{x}(t)$ is the imaginary part, corresponding to the input signal and the Hilbert transform, respectively.

Appendix F

Overview of GMM-HMM based acoustic modelling

A standard ASR system employs a triphone context-dependent HMM. In this appendix, the GMM-HMM acoustic model is briefly described.

F.1 Hidden Markov model

The hidden Markov model (HMM) is a probabilistic model that can model observed time series data in an unknown state [202], such that at any given time t a sequence of observations $Y = \{y_1, y_2, \ldots . y_n\}$ is produced by a sequence of hidden states $Q = \{q_1, q_2, \ldots . q_n\}$. The hidden states are assumed to be conditionally independent from the other states except the current one at time t. In HMM acoustic models, words are modelled as sequences of word units (phones), such that each word unit is represented by a sequence of hidden states and the probability distribution associated with it [202]. Figure F.1, illustrates a five-state HMM word unit model. Each state s_i has transition probabilities a_{ij} and output probabilities b_i for the observed acoustic features y_i. The start and end states are not associated with output probabilities; hence they are called 'non-emitting states. The state output probabilities are modelled as a mixture of Gaussians, in order to model the asymmetric and multimodal densities of the observed acoustic features [202].

The common technique for training an HMM model is Maximum Likelihood (ML) estimation, where the aim is to maximise the model parameters given the observed data. The standard algorithm for ML estimation in ASR is the Baum-Welch algorithm, where the posterior probability for state s_i associated with observation y_i in a whole sequence is calculated using forward and backward probabilities. The forward probability refers to the

probability of all the states in the sequence up to state s_i at time t. The backward probability refers to the probability that the model produces the remaining observed sequence given that the state s_i at time t was occupied. Alternatively, Viterbi decoding can be used; however, it only applies a forward probability pass. This method of HMM training is implemented in the Kaldi toolkit for its efficiency and low computational cost [233].

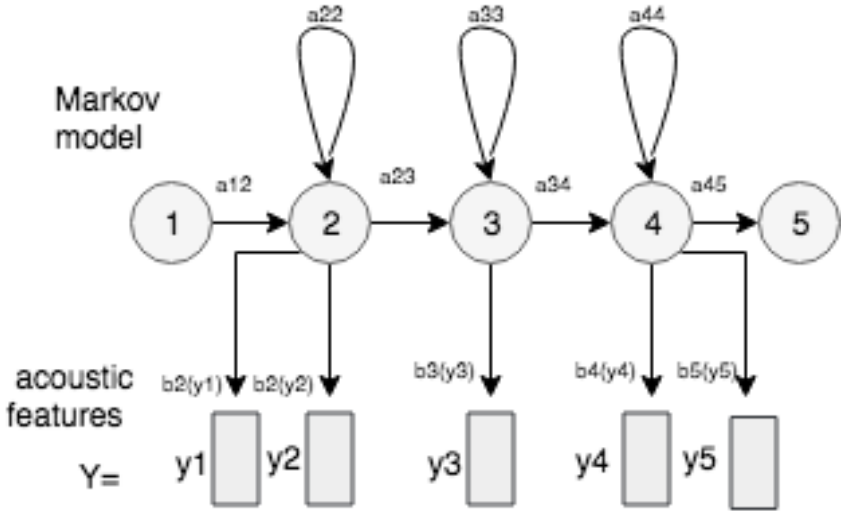


Figure F.1: A left-to-right five-state HMM, adapted from [202].

Typically, HMM acoustic model parameters are initially trained using a simple phone model called ‘monophones’. The model sets all state transition probabilities with uniform distribution. The state’s output probability distributions are initialised with a global mean and covariance matrix estimated from the observed data. In this model, the phones are modelled context-independently from the other phones. In the context-dependent model, so called ‘triphones’, a unique phone model is created by considering the left and right context of each observed phone.

The number of HMM model parameters tends to increase, as the Gaussian mixture models (GMMs) are estimated for each HMM state separately, yielding an increase in the model complexity. Hence, more training data are required in order to create accurate HMM models. To reduce the HMM model’s complexity, HMMs with similar parameters can be combined to model data with acoustic similarities. Similarity between models is evaluated

by measuring the distance between their parameters, and models that fall within a certain distance are clustered together. The combined model is referred to as a 'tied state model'. For further details of HMM-based ASR, the reader can refer to [202].

F.2 Gaussian mixture model

The GMM is a weighted sum of *K* Gaussian component densities with means and diagonal covariance matrix [234]. The mixture weights $\emptyset$ of the model are commonly estimated using the Expectation Maximization (EM) estimation technique [235], such that $\sum_1^k \emptyset_k = 1$. The GMM probability density function (PDF) is defined by Equation F.1 [234].

$$p(x) = \sum_1^k \emptyset_{\,k}\, p(x|\mu_k,,\Sigma_k) \tag{F.1}$$

where x is the input feature vector of the *D*-dimension. $p(x|\mu_k,,\Sigma_k)$ is the *k*th Gaussian component density, consisting of two parameters, mean μ_k and covariance matrix Σ_k, and defined by Equation F.2 [234].

$$p(x|\mu_k,,\Sigma_k) = \frac{1}{(2\pi)^{D/2}|\Sigma_k|^{1/2}} \exp(-1/2(x-\mu_k)^T\ {\Sigma_k}^{-1}(x-\mu_k)) \tag{F.2}$$

References

[1] J. Tao and T. Tan, *Affective information processing*. Springer-Verlag London, 2009.

[2] R. W. Picard, *Affective computing*. MIT Press, 2000.

[3] S. Jerritta, M. Murugappan, R. Nagarajan, and K. Wan, "Physiological signals based human emotion recognition: a review," in *7th International Colloquium on Signal Processing and its Applications*, 2011, pp. 410-415.

[4] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: from unimodal analysis to multimodal fusion," *Information Fusion*, vol. 37, pp. 98-125, 2017.

[5] J. A. Russell, "A circumplex model of affect," *Journal of Personality and Social Psychology*, vol. 39, no. 6, pp. 1161-1178, 1980.

[6] S. K. D'mello and J. Kory, "A review and meta-analysis of multimodal affect detection systems," *ACM Computing Surveys*, vol. 47, no. 3, pp. 1-36, 2015.

[7] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A database of german emotional speech," in *Proceedings of the 9th European conference on speech communication and technology*, 2005, pp. 1517–1520.

[8] C. M. Lee and S. S. Narayanan, "Toward detecting emotions in spoken dialogs," *IEEE transactions on speech and audio processing*, vol. 13, no. 2, pp. 293-303, 2005.

[9] M. Soleymani, M. Pantic, and T. Pun, "Multimodal emotion recognition in response to videos," *IEEE Transactions on Affective Computing*, vol. 3, no. 2, pp. 211-223, 2012.

[10] M. Moghimi, R. Stone, P. Rotshtein, and N. Cooke, "Influencing Human Affective Responses to Dynamic Virtual Environments," *Presence: Teleoperators and Virtual Environments*, vol. 25, no. 2, pp. 81-107, 2016.

[11] L. Ten Bosch, "Emotions, speech and the ASR framework," *Speech Communication*, vol. 40, no. 1-2, pp. 213-225, 2003.

[12] Z. H. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: audio, visual, and spontaneous expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 1, pp. 39-58, 2009.

[13] K. P. Truong, D. A. Van Leeuwen, and F. M. De Jong, "Speech-based recognition of self-reported and observed emotion in a dimensional space," *Speech communication*, vol. 54, no. 9, pp. 1049-1063, 2012.

[14] W.-L. Zheng, B.-N. Dong, and B.-L. Lu, "Multimodal emotion recognition using EEG and eye tracking data," in *36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2014, pp. 5040-5043.

[15] S. T. Iqbal, X. S. Zheng, and B. P. Bailey, "Task-evoked pupillary response to mental workload in human-computer interaction," in *Proceedings of the ACM Conference on Human Factors in Computing Systems*, Vienna, Austria, 2004: ACM Press, pp. 1477-1480.

[16] I. Kotsia, S. Zafeiriou, and S. Fotopoulos, "Affective gaming: A comprehensive survey," in *IEEE Computer Vision and Pattern Recognition Workshops*, 2013, pp. 663-670.

[17] S. I. Hjelm, "Research + design: the making of Brainball," *Interactions*, vol. 10, no. 1, pp. 26-34, 2003.

[18] M. Porta, S. Ricotti, and C. J. Perez, "Emotional e-learning through eye tracking," in *IEEE Global Engineering Education Conference (EDUCON)*, 2012, pp. 1-6.

[19] K. Holmqvist, M. Nyström, R. Andersson, R. Dewhurst, H. Jarodzka, and J. Van de Weijer, *Eye tracking: A comprehensive guide to methods and measures*. OUP Oxford, 2011.

[20] D. F. Fisher, R. A. Monty, J. W. Senders, and (eds), *Eye movements: cognition and visual perception*. Hillsdale, NJ: Lawrence Erlbaum, 1981.

[21] M. Wedel and R. Pieters, "A review of eye-tracking research in marketing," in *Review of marketing research*, N. K. Malhotra Ed., 2008, pp. 123-147.

[22] R. J. Jacob and K. S. Karn, "Eye tracking in human-computer interaction and usability research: ready to deliver the promises," in *The mind's eye: Cognitive and Applied Aspects of Eye Movement Research*, R. Radach, J. Hyona, and H. Deubel Eds. Amsterdam, The Netherlands: Elsevier, 2003, pp. 573-605.

[23] R. Ruddarraju *et al.*, "Perceptual user interfaces using vision-based eye tracking," in *Proceedings of the 5th international conference on Multimodal interfaces*, Vancouver, British Columbia, Canada, 2003, pp. 227-233.

[24] G. Chanel, M. Bétrancourt, T. Pun, D. Cereghetti, and G. Molinari, "Assessment of computer-supported collaborative processes using interpersonal physiological and eye-movement coupling," in *Humaine Association Conference on Affective Computing and Intelligent Interaction*, Geneva, Switzerland, 2013, pp. 116-122.

[25] F. Karray, M. Alemzadeh, J. A. Saleh, and M. N. Arab, "Human-computer interaction: overview on state of the art," *International Journal on Smart Sensing and Intelligent Systems*, vol. 1, no. 1, pp. 137-159, 2008.

[26] C. H. Morimoto and M. R. Mimica, "Eye gaze tracking techniques for interactive applications," *Computer vision and image understanding*, vol. 98, no. 1, pp. 4-24, 2005.

[27] SRResearch, *Eyelink II users manual,version 1.05*. Mississauga, Canada.: SR Research Ltd., 2002.

[28] L. E. Nacke and R. L. Mandryk, "Designing affective games with physiological input," in *Workshop on multiuser and social biosignal adaptive games and playful applications in fun and games conference (BioS-Play)*, 2010.

[29] B.-H. Juang and L. R. Rabiner, "Automatic speech recognition — a brief history of the technology development," in *Elsevier Encyclopedia of Language and Linguistics*, 2 ed.: Elsevier, 2005.

[30] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257-286, 1989.

[31] A. L. Gorin, G. Riccardi, and J. H. Wright, "How may I help you?," *Speech Commun.*, vol. 23, no. 1-2, pp. 113-127, 1997.

[32] U. Shrawankar and V. Thakare, "Adverse conditions and ASR techniques for robust speech user interface," *International Journal of Computer Science (ICJSI)*, vol. 8, no. 5, 2011.

[33] B. Schuller, J. Stadermann, and G. Rigoll, "Affect-robust speech recognition by dynamic emotional adaptation," in *Proceedings of ISCA Speech Prosody 2006*, Dresden, Germany, 2006.

[34] V. V. R. Vegesna, K. Gurugubelli, and A. K. Vuppala, "Application of emotion recognition and modification for emotional telugu speech recognition," *Mobile Networks and Applications*, vol. 24, no. 1, pp. 193-201, 2019.

[35] S. A. Karkhedkar, "Using emotion as inferred from prosody in language modeling," PhD thesis, The University of Texas at El Paso, 2013.

[36] K. Iwano, T. Yoshinaga, S. Tamura, and S. Furui, "Audio-visual speech recognition using lip information extracted from side-face images," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2007, no. 1, pp. 4-4, 2007.

[37] N. J. Cooke and M. Russell, "Gaze-contingent automatic speech recognition," *IET signal processing*, vol. 2, no. 4, pp. 369-380, 2008.

[38] A. Shen, N. Cooke, and M. J. Russell, "Selective use of gaze information to improve ASR performance in noisy environments by cache-based class language model adaptation," in *INTERSPEECH*, Lyon, France, 2013, pp. 1844-1848.

[39] A. Alhargan, N. Cooke, and T. Binjammaz, "Affect recognition in an interactive gaming environment using eye tracking," in *17th International Conference on Affective Computing and Intelligent Interaction (ACII)*, 2017.

[40] A. Alhargan, N. Cooke, and T. Binjammaz, "Multimodal affect recognition in an interactive gaming environment using eye tracking and speech signals," in *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, 2017: ACM Press, pp. 479-486.

[41] K. R. Scherer, "What are emotions? and how can they be measured?," *Social Science Information*, vol. 44, pp. 695-729, 2005.

[42] P. Kleinginna, Jr. and A. Kleinginna, "A categorized list of emotion definitions, with suggestions for a consensual definition," *Motivation and Emotion,* vol. 5, no. 4, pp. 345-379, 1981.

[43] C. Darwin, P. Ekman, Ed. *The expression of the emotions in man and animals*, 3 ed. Oxford University Press, USA, 1998.

[44] W. James, "What is an emotion?," *Mind,* vol. 9, pp. 188-205, 1884.

[45] P. Ekman *et al.*, "Universals and cultural differences in the judgments of facial expressions of emotion," *Journal of Personality and Social Psychology,* vol. 53, no. 4, pp. 712-717, 1987.

[46] J. A. Russell, "Culture and the categorization of emotions," *Psychological bulletin,* vol. 110, no. 3, pp. 426–450, 1991.

[47] A. Mehrabia, "A semantic space for nonverbal behavior," *Journal of Consulting and Clinical Psychology,* vol. 35, no. 2, pp. 248-257, 1970.

[48] D. Grandjean, D. Sander, and K. R. Scherer, "Conscious emotional experience emerges as a function of multilevel, appraisal-driven response synchronization," *Consciousness and Cognition,* vol. 17, no. 2, pp. 484-495, 2008.

[49] C. Aracena, S. Basterrech, V. Snáel, and J. Velásquez, "Neural networks for emotion recognition based on eye tracking data," in *IEEE International Conference on Systems, Man, and Cybernetics*, 2015, pp. 2632-2637.

[50] D. Wu, T. D. Parsons, and S. S. Narayanan, "Acoustic feature analysis in speech emotion primitives estimation," in *INTERSPEECH*, Makuhari, Chiba, Japan, 2010, pp. 785–788.

[51] L. F. Barrett, M. Lewis, and J. M. Haviland-Jones, Eds. *Handbook of emotions*, 4 ed. New York, NY, US: Guilford Press, 2016.

[52] B. Reeves and C. I. Nass, *The media equation: How people treat computers, television, and new media like real people and places*. Cambridge university press, 1996.

[53] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge," *Speech Communication,* vol. 53, no. 9-10, pp. 1062-1087, 2011.

[54] T. Friedrichs, C. Zschippig, M. Herrlich, B. Walther-Franks, R. Malaka, and K. Schill, "Simple games – complex emotions: automated affect detection using physiological signals," in *Proceedings of the 14th International Conference on Entertainment Computing - ICEC 2015,* Trondheim, Norway, 2015, pp. 375-382.

[55] M. Moghimi, R. Stone, and P. Rotshtein, "A psychophysiological evaluation of an effective recognition technique using interactive dynamic virtual environments," *World Academy of Science, Engineering and Technology, International Journal of Computer, Electrical, Automation, Control and Information Engineering,* vol. 11, no. 5, pp. 559-567, 2017.

[56] C. Peter, E. Ebert, and H. Beikirch, "Physiological sensing for affective computing," in *Affective Information Processing*, J. Tao and T. Tan Eds.: Springer-Verlag London, 2009, pp. 293-310.

[57] J. Kory and S. D'Mello, "Affect elicitation for affective computing," in *The Oxford Handbook of Affective Computing*, R. Calvo, S. D'Mello, J. Gratch, and A. Kappas Eds.: Oxford University Press, 2014.

[58] S. Shahid, E. Krahmer, and M. Swerts, "Real vs. acted emotional speech: comparing south-Asian and caucasian speakers and observers," in *Proceedings of the 4th International Conference on Speech Prosody*, Campinas, Brazil, 2008, pp. 669-772.

[59] M. Grimm, K. Kroschel, and S. Narayanan, "The vera am mittag german audio-visual emotional speech database," in *IEEE International Conference on Multimedia and Expo*, 2008, pp. 865-868.

[60] M. M. Bradley and P. J. Lang, "Measuring emotion: the self-assessment manikin and the semantic differential," *Journal of behavior therapy and experimental psychiatry*, vol. 25, no. 1, pp. 49-59, 1994.

[61] M. M. Bradley, L. Miccoli, M. A. Escrig, and P. J. Lang, "The pupil as a measure of emotional arousal and autonomic activation," *Psychophysiology*, vol. 45, no. 4, pp. 602-607, 2008.

[62] P. J. Lang, M. M. Bradley, and B. N. Cuthbert, "International affective picture system (IAPS): Affective ratings of pictures and instruction manual," Gainesville, FL, Tech. Rep. A-8, 2008.

[63] T. Partala and V. Surakka, "Pupil size variation as an indication of affective processing," *International Journal of Human-Computer Studies*, vol. 59, no. 1-2, pp. 185-198, 2003.

[64] M. M. Bradley and P. J. Lang, "International affective digitized sounds (IADS) stimuli instruction manual and affective ratings," Gainesville, FL, Tech. Rep. B-2, 1999.

[65] G. Riva *et al.*, "Affective interactions using virtual reality: The link between presence and emotions," *Cyberpsychology & Behavior*, vol. 10, no. 1, pp. 45-56, 2007.

[66] N. Dahlbäck, A. Jönsson, and L. Ahrenberg, "Wizard of Oz studies—why and how," *Knowledge-based systems*, vol. 6, no. 4, pp. 258-266, 1993.

[67] E. Douglas-Cowie *et al.*, "The HUMAINE database: addressing the collection and annotation of naturalistic and induced emotional data," in *International conference on affective computing and intelligent interaction*, Lisbon, Portugal, 2007, pp. 488-500.

[68] S. Steininger, F. Schiel, O. Dioubina, and S. Raubold, "Development of user-state conventions for the multimodal corpus in smartkom," in *Proc. of the Workshop on Multimodal Resources and Multimodal Systems Evaluation*, 2002, pp. 33-37.

[69] S. Steidl, "Automatic classification of emotion-related user states in spontaneous children's speech," PhD thesis, Universität Erlangen-Nürnberg, Germany, 2009.

[70] A. Rodriguez, B. Rey, M. Clemente, M. Wrzesien, and M. Alcaniz, "Assessing brain activations associated with emotional regulation during virtual reality mood induction procedures," *Expert Systems with Applications,* vol. 42, no. 3, pp. 1699-1709, 2015.

[71] G. N. Yannakakis and A. Paiva, "Emotion in games," in *The Oxford Handbook of Affective Computing*, R. Calvo, S. D'Mello, J. Gratch, and A. Kappas Eds.: Oxford University Press, 2014, pp. 459-471.

[72] B. Schuller *et al.*, "Paralinguistics in speech and language State-of-the-art and the challenge," *Computer Speech and Language,* vol. 27, no. 1, pp. 4-39, 2013.

[73] D. Ververidis and C. Kotropoulos, "Emotional speech recognition: resources, features, and methods," *Speech Communication,* vol. 48, no. 9, pp. 1162-1181, 2006.

[74] S. Narayanan and P. G. Georgiou, "Behavioral signal processing: deriving human behavioral informatics from speech and language," *Proceedings of the IEEE,* vol. 101, no. 5, pp. 1203-1233, 2013.

[75] P. Taylor, "Phonetics and phonology," in *Text-to-speech synthesis*: Cambridge university press, 2009.

[76] K. R. Scherer, "Vocal communication of emotion: A review of research paradigms," *Speech communication,* vol. 40, no. 1-2, pp. 227-256, 2003.

[77] T. Johnstone and K. Scherer, "Vocal communication of emotion," in *The Handbook of Emotions*, M. Lewis and J. Haviland- Jones Eds., 2 ed. New York: Guilford Press, 2000, ch. 14.

[78] C. E. Williams and K. N. Stevens, "Vocal correlates of emotional states," *Speech Evaluation in Psychiatry,* pp. 189–220, 1981.

[79] K. R. Scherer, "Vocal affect expression: A review and a model for future research," *Psychological bulletin,* vol. 99, no. 2, p. 143, 1986.

[80] M. Belyk and S. Brown, "The acoustic correlates of valence depend on emotion family," *Journal of Voice,* vol. 28, no. 4, 2014.

[81] I. R. Murray and J. L. Arnott, "Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion," *The Journal of the Acoustical Society of America,* vol. 93, no. 2, pp. 1097-1108, 1993.

[82] C. Gobl and A. N. Chasaide, "The role of voice quality in communicating emotion, mood and attitude," *Speech communication,* vol. 40, no. 1-2, pp. 189-212, 2003.

[83] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: features, classification schemes, and databases," *Pattern Recognition,* vol. 44, no. 3, pp. 572-587, 2011.

[84] R. Cowie *et al.*, "Emotion recognition in human-computer interaction," *IEEE Signal Processing Magazine,* vol. 18, no. 1, pp. 32-80, 2001.

[85] S. A. Ali, A. Khan, and N. Bashir, "Analyzing the impact of prosodic feature (pitch) on learning classifiers for speech emotion corpus," *International Journal of Information Technology and Computer Science (IJITCS),* vol. 7, no. 2, 2015.

[86] S. Q. Wu, T. H. Falk, and W. Y. Chan, "Automatic speech emotion recognition using modulation spectral features," *Speech Communication,* vol. 53, no. 5, pp. 768-785, 2011.

[87] B. Schuller, S. Steidl, and A. Batliner, "The INTERSPEECH 2009 Emotion Challenge," in *Proc. Interspeech*, Brighton, UK, 2009, pp. 312–315.

[88] B. Vlasenko and A. Wendemuth, "Processing affected speech within human machine interaction," in *Proc. Interspeech*, Brighton, UK, 2009, pp. 2039–2042.

[89] S. Bedoya-Jaramillo, E. Belalcazar-Bolanos, T. Villa-Canas, J. R. Orozco-Arroyave, J. D. Arias-Londono, and J. F. Vargas-Bonilla, "Automatic emotion detection in speech using mel frequency cesptral coefficients," in *Proceedings of the Symposium of Image, Signal Processing, and Artificial Vision (STSIVA)*, 2012, pp. 62-65.

[90] B. Schuller, "Recognizing affect from linguistic information in 3d continuous space," *IEEE Transactions on Affective Computing,* vol. 2, no. 4, pp. 192-205, 2011.

[91] F. Eyben, A. Batliner, B. Schuller, D. Seppi, and S. Steidl, "Cross-corpus classification of realistic emotions–some pilot experiments," in *Proc. 3rd Internat. Workshop on EMOTION (Satellite of LREC): Corpora for Research on Emotion and Affect*, Valletta, Malta, 2010, pp. 77–82.

[92] H. P. Espinosa, C. A. R. García, and L. V. Pineda, "Features selection for primitives estimation on emotional speech," in *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, 2010, pp. 5138-5141.

[93] M. Tahon, G. Degottex, and L. Devillers, "Usual voice quality features and glottal features for emotional valence detection," in *Speech Prosody*, 2012, pp. 693–696.

[94] B. Schuller *et al.*, "The relevance of feature type for the automatic classification of emotional user states: low level descriptors and functionals," in *Interspeech*, Antwerp, Belgium, 2007, pp. 2253-2256.

[95] Y. Sun, G. Wen, and J. Wang, "Weighted spectral features based on local Hu moments for speech emotion recognition," *Biomedical Signal Processing and Control,* vol. 18, pp. 80-90, 2015.

[96] B. Schuller *et al.*, "The INTERSPEECH 2010 Paralinguistic Challenge," in *Interspeech*, Makuhari, Chiba, Japan, 2010, pp. 2794- 2797.

[97] B. Schuller, S. Steidl, A. Batliner, F. Schiel, and J. Krajewski, "The INTERSPEECH 2011 Speaker State Challenge," in *Interspeech*, Florence, Italy, 2011, pp. 3201-3204.

[98] B. Schuller *et al.*, "The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," in *Interspeech*, Lyon, France, 2013, pp. 148–152.

[99] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *Proceedings of the 21st ACM international conference on Multimedia*, Barcelona, Spain, 2013, pp. 835-838.

[100] A. Jović, K. Brkić, and N. Bogunović, "A review of feature selection methods with applications," in *Proceeding of the 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, Opatija, Croatia, 2015, pp. 1200-1205.

[101] M. Grimm and K. Kroschel, "Emotion estimation in speech using a 3D emotion space concept," in *Proceeding of the IEEE Automatic Speech Recognition and Understanding Workshop*, 2005, pp. 381–385.

[102] C. N. Anagnostopoulos, T. Iliou, and I. Giannoukos, "Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011," *Artificial Intelligence Review*, vol. 43, no. 2, pp. 155-177, 2015.

[103] M. R. Amer, B. Siddiquie, C. Richey, and A. Divakaran, "Emotion detection in speech using deep networks," in *ICASSP*, 2014: Florence, Italy, pp. 3752–3756.

[104] K. P. Murphy, "Deep learning," in *Machine Learning A Probabilistic Perspective*: MIT Press, 2012, pp. 995-1007.

[105] R. Plutchik, *The psychology and biology of emotion*. New York, NY, US: HarperCollins College Publishers, 1994.

[106] C. M. Lee, S. S. Narayanan, and R. Pieraccini, "Combining acoustic and language information for emotion recognition," in *Interspeech*, Denver, USA, 2002, pp. 873-376.

[107] T. Polzehl, A. Schmitt, F. Metze, and M. Wagner, "Anger recognition in speech using acoustic and linguistic cues," *Speech Communication*, vol. 53, no. 9-10, pp. 1198-1209, 2011.

[108] Q. Jin, C. Li, S. Chen, and H. Wu, "Speech emotion recognition with acoustic and lexical features," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane, QLD, Australia, 2015, pp. 4749-4753.

[109] K. W. Gamage, V. Sethu, and E. Ambikairajah, "Salience based lexical features for emotion recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 5830-5834.

[110] A. Batliner, K. Fischer, R. Huber, J. Spilker, and E. Noth, "How to find trouble in communication," *Speech Communication*, vol. 40, no. 1-2, pp. 117-143, 2003.

[111] M. Bulut, S. Lee, S. Narayanan, and Isca, "Analysis of emotional speech prosody in terms of part of speech tags," in *Interspeech*, Antwerp, Belgium, 2007, pp. 626–629.

[112] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Emotion recognition from speech: putting ASR in the loop," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Taipei, Taiwan, 2009, pp. 4585-4588.

[113] T. Athanaselis, S. Bakamidis, I. Dologlou, R. Cowie, E. Douglas-Cowie, and C. Cox, "ASR for emotional speech: clarifying the issues and enhancing performance," *Neural Networks,* vol. 18, no. 4, pp. 437-444, 2005.

[114] F. Metze, A. Batliner, F. Eyben, T. Polzehl, B. Schuller, and S. Steidl, "Emotion recognition using imperfect speech recognition," in *Interspeech*, Makuhari, Chiba, Japan, 2010, pp. 478-481.

[115] L. Kessous, G. Castellano, and G. Caridakis, "Multimodal emotion recognition in speech-based interaction using facial expression, body gesture and acoustic analysis," *Journal on Multimodal User Interfaces,* vol. 3, no. 1-2, pp. 33-48, 2010.

[116] F. Ringeval *et al.*, "Av+ ec 2015: The first affect recognition challenge bridging across audio, video, and physiological data," in *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*, Brisbane, Australia, 2015, pp. 3-8.

[117] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions," in *10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, 2013, pp. 1-8.

[118] I. B. Mauss and M. D. Robinson, "Measures of emotion: A review," *Cognition and emotion,* vol. 23, no. 2, pp. 209-237, 2009.

[119] J. Beatty and B. Lucero-Wagoner, "The pupillary system," in *Handbook of psychophysiology*, J. T. Cacioppo, L. G. Tassinary, and G. G. Berntson Eds. New York, NY, US: Cambridge University Press, 2000, pp. 142-162.

[120] J. T.-Y. Wang, "Pupil dilation and eye-tracking," in *A Handbook of Process Tracing Methods for Decision Research: A Critical Review and User's Guide*, M. SchulteMecklenbeck, M. Kühberger, and R. Rayard Eds. USA: Psychology Press, 2011, pp. 185-204.

[121] D. Guitton, R. Simard, and F. Codère, "Upper eyelid movements measured with a search coil during blinks and vertical saccades," *Investigative Ophthalmology & Visual Science,* vol. 32, no. 13, pp. 3298-3305, 1991.

[122] S. Baltaci, D. Gokcay, and Acm, "Negative Sentiment in Scenarios Elicit Pupil Dilation Response: An Auditory Study," in *Proceedings of the ACM International Conference on Multimodal Interaction (ICMI)*, Santa Monica, California, USA, 2012, pp. 529-532.

[123] S. Kawai, H. Takano, and K. Nakamura, "Pupil diameter variation in positive and negative emotions with visual stimulus," in *International Conference on Systems, Man, and Cybernetics (SMC)*, Manchester, UK, 2013, pp. 4179-4183.

[124] A. Herbon, C. Peter, L. Markert, E. van der Meer, and J. Voskamp, "Emotion studies in HCI—a new approach," in *Proceedings of the 2005 HCI International Conference*, Las Vegas, 2005.

[125] J. A. Stern, L. C. Walrath, and R. Goldstein, "The endogenous eyeblink," *Psychophysiology,* vol. 21, no. 1, pp. 22-33, 1984.

[126] K. Kaneko and K. Sakamoto, "Spontaneous blinks as a criterion of visual fatigue during prolonged work on visual display terminals," *Perceptual and Motor Skills,* vol. 92, no. 1, pp. 234-250, 2001.

[127] Y. Tanaka, "Arousal level and blink activity," *Shinrigaku kenkyu: The Japanese journal of psychology,* vol. 70, no. 1, pp. 1-8, 1999.

[128] R. Subramanian, D. Shankar, N. Sebe, and D. Melcher, "Emotion modulates eye movement patterns and subsequent memory for the gist and details of movie scenes," *Journal of Vision,* vol. 14, no. 3, p. 18, 2014.

[129] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, "A multimodal database for affect recognition and implicit tagging," *IEEE Transactions on Affective Computing,* vol. 3, no. 1, pp. 42-55, 2012.

[130] Y. Lu, W.-L. Zheng, B. Li, and B.-L. Lu, "Combining eye movements and EEG to enhance emotion recognition," in *24th International Joint Conference on Artificial Intelligence (IJCAI'2015)*, 2015, pp. 1170–1176.

[131] V. Raudonis, G. Dervinis, A. Vilkauskas, A. Paulauskaite-Taraseviciene, and G. Kersulyte-Raudone, "Evaluation of human emotion from eye motions," *International Journal of Advanced Computer Science and Applications (IJACSA),* vol. 4, no. 8, 2013.

[132] S. Alghowinem, M. AlShehri, R. Goecke, and M. Wagner, "Exploring eye activity as an indication of emotional states using an eye-tracking sensor," in *Intelligent Systems for Science and Information*, L. Chen, S. Kapoor, and R. Bhatia Eds., (Studies in Computational Intelligence: Springer International Publishing, 2014, pp. 261-276.

[133] D. Novak, M. Mihelj, and M. Munih, "A survey of methods for data fusion and system adaptation using autonomic nervous system responses in physiological computing," *Interacting with computers,* vol. 24, no. 3, pp. 154-172, 2012.

[134] S. Steidl, A. Batliner, D. Seppi, and B. Schuller, "On the impact of children's emotional speech on acoustic and language models," *EURASIP Journal on Audio, Speech, and Music Processing,* vol. 2010, no. 1, 2010.

[135] B. Vlasenko, D. Prylipko, and A. Wendemuth, "Towards robust spontaneous speech recognition with emotional speech adapted acoustic models," in *Proceedings of 35th German Conference on Artificial Intelligence*, Saarbrücken, Germany, 2012, pp. 74-77.

[136] C. M. Whissell, "The dictionary of affect in language," in *The measurement of emotions*, 1989, pp. 113-131.

[137] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "Switchboard: telephone speech corpus for research and development," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, San Francisco, CA, USA, 1992, pp. 517-520.

[138] C. Busso *et al.*, "IEMOCAP: Interactive emotional dyadic motion capture database," *Journal of Language Resources and Evaluation,* vol. 42, no. 4, pp. 335-359, 2008.

[139] N. Cooke, "Gaze-contingent automatic speech recognition," PhD thesis, University of Birmingham, 2006.

[140] A. Shen, "The selective use of gaze in automatic speech recognition," PhD thesis, University of Birmingham, 2014.

[141] J. S. Robert and F. P. Hannigan, "Applications of virtual environments: an overview," in *Handbook of Virtual Environments:Design, Implementation and Applications*, K. S. Hale and K. M. Stanney Eds.: CRC Press, 2014, pp. 883-957.

[142] K. A. Ericsson and H. A. Simon, "Verbal reports as data," *Psychological review*, vol. 87, no. 3, pp. 215-251, 1980.

[143] J. Nielsen, T. Clemmensen, and C. Yssing, "Getting access to what goes on in people's heads?: reflections on the think-aloud technique," in *Proceedings of the second Nordic conference on Human-computer interaction*, 2002, pp. 101-110.

[144] S. Soleimani and E. L.-C. Law, "What can self-reports and acoustic data analyses on emotions tell us?," in *Proceedings of the 2017 Conference on Designing Interactive Systems*, 2017, pp. 489-501.

[145] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an ASR corpus based on public domain audio books," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206-5210.

[146] B. Gagl, S. Hawelka, and F. Hutzler, "Systematic influence of gaze position on pupil size measurement: analysis and correction," *Behavior Research Methods*, vol. 43, no. 4, pp. 1171-1181, 2011.

[147] J. Carletta, "Assessing agreement on classification tasks: the kappa statistic," *Computational linguistics*, vol. 22, no. 2, pp. 249-254, 1996.

[148] V. Pappu and P. M. Pardalos, "High-dimensional data classification," in *Clusters, Orders, and Trees: Methods and Applications*, vol. 92, F. Aleskerov, B. Goldengorin, and P. Pardalos Eds., 2014, pp. 119-150.

[149] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273-297, 1995.

[150] C. J. Burges, "A tutorial on support vector machines for pattern recognition," *Data mining and knowledge discovery*, vol. 2, no. 2, pp. 121-167, 1998.

[151] L. Bottou and C.-J. Lin, "Support vector machine solvers," in *Large scale kernel machines*, L. Bottou, O. Chapelle, D. DeCoste, and J. Weston Eds.: MIT Press, 2007.

[152] C.-W. Hsu, C.-C. Chang, and C.-J. Lin, "A practical guide to support vector classification," Department of Computer Science and Information Engineering, National Taiwan University, Taiwan., Tech. rep 2003.

[153] S. S. Keerthi and C.-J. Lin, "Asymptotic behaviors of support vector machines with Gaussian kernel," *Neural computation*, vol. 15, no. 7, pp. 1667-1689, 2003.

[154] H.-T. Lin and C.-J. Lin, "A study on sigmoid kernels for SVM and the training of non-PSD kernels by SMO-type methods," Technical report 2003.

[155] C.-W. Hsu and C.-J. Lin, "A comparison of methods for multiclass support vector machines," *IEEE transactions on Neural Networks,* vol. 13, no. 2, pp. 415-425, 2002.

[156] S. Knerr, L. Personnaz, and G. Dreyfus, "Single-layer learning revisited: a stepwise procedure for building and training a neural network," in *Neurocomputing*, vol. 68, S. F.F. and H. J. Eds.: Springer, 1990, pp. 41-50.

[157] J. Wagner, E. Andre, F. Lingenfelser, and J. Kim, "Exploring fusion methods for multimodal emotion recognition with missing data," *IEEE Transactions on Affective Computing,* vol. 2, no. 4, pp. 206-218, 2011, doi: 10.1109/T-AFFC.2011.12.

[158] R. E. Bellman, *Adaptive control processes: a guided tour*. Princeton university press, 1961.

[159] A. R. Webb, *Statistical pattern recognition*. John Wiley & Sons, 2003.

[160] G. Stemmler, M. Heldmann, C. A. Pauls, and T. Scherer, "Constraints for emotion specificity in fear and anger: The context counts," *Psychophysiology,* vol. 38, no. 2, pp. 275-291, 2001.

[161] Q. Gu, L. Zhu, and Z. Cai, "Evaluation measures of the classification performance of imbalanced data sets," in *International Symposium on Intelligence Computation and Applications*, 2009, pp. 461-471.

[162] C. Ellis, "The pupillary light reflex in normal subjects," *British Journal of Ophthalmology*, vol. 65, no. 11, pp. 754-759, 1981.

[163] M. Nakayama, "Influence of blink on pupillary indices," in *IEEE Biomedical Circuits and Systems Conference*, 2006, pp. 29-32.

[164] M. Nakayama and Y. Shimizu, "Frequency analysis of task evoked pupillary response and eye-movement," in *Proceedings of the 2004 symposium on eye tracking research and applications*, San Antonio, Texas, 2004, pp. 71-76.

[165] S. Mathôt, "A simple way to reconstruct pupil size during eye blinks," ed, 2013.

[166] F. Onorati, R. Barbieri, M. Mauri, V. Russo, and L. Mainardi, "Reconstruction and analysis of the pupil dilation signal: Application to a psychophysiological affective protocol," in *Proceeding of the 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2013, pp. 5-8.

[167] C. M. Privitera, L. W. Renninger, T. Carney, S. Klein, and M. Aguilar, "Pupil dilation during visual target detection," *Journal of Vision,* vol. 10, no. 10, p. 3, 2010.

[168] V. Peysakhovich, M. Causse, S. Scannella, and F. Dehais, "Frequency analysis of a task-evoked pupillary response: Luminance-independent measure of mental effort," *International Journal of Psychophysiology*, vol. 97, no. 1, pp. 30-37, 2015.

[169] M. Pomplun and S. Sunkara, "Pupil dilation as an indicator of cognitive workload in human-computer interaction," in *Human-Centred Computing: Cognitive, Social, and Ergonomic Aspects. Vol. 3 of the Proceedings of the 10th International Conference on Human-Computer Interaction*, Crete, Greece, D. Harris, V., Duffy, M. Smith, and C. Stephanidis, Eds., 2003.

[170] F. T. Oliveira, A. Aula, and D. M. Russell, "Discriminating the relevance of web search results with measures of pupil size," in *Proceedings of the 27th International Conference on Human Factors in Computing Systems*, Boston, MA, USA, 2009, pp. 2209-2212.

[171] V. F. Pamplona, M. M. Oliveira, and G. V. G. Baranoski, "Photorealistic models for pupil light reflex and iridal pattern deformation," *ACM Transactions on Graphics (TOG)*, vol. 28, no. 4, 2009.

[172] A. Longtin and J. G. Milton, "Modelling autonomous oscillations in the human pupil light reflex using non-linear delay-differential equations," *Bulletin of Mathematical Biology*, vol. 51, no. 5, pp. 605-624, 1989.

[173] P. Moon and D. E. Spencer, "On the stiles-crawford effect," *Journal of the Optical Society of America*, vol. 34, no. 6, pp. 319-329, 1944.

[174] C. M. Privitera and L. W. Stark, "A binocular pupil model for simulation of relative afferent pupil defects and the swinging flashlight test," *Biological Cybernetics*, vol. 94, no. 3, pp. 215-224, 2006.

[175] J. Allen, "Short term spectral analysis, synthesis, and modification by discrete fourier transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 25, no. 3, pp. 235-238, 1977.

[176] A. D. Poularikas, "The hilbert transform," in *The Handbook of Formulas and Tables for Signal Processing*, A. D. Poularikas Ed.: Boca Raton: CRC Press LLC, 1999, ch. 15.

[177] G. Hossain and M. Yeasin, "Understanding effects of cognitive load from pupillary responses using Hilbert analytic phase," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, Columbus, OH, USA, 2014, pp. 381-386.

[178] J. W. Freeman, "Hilbert transform for brain waves," in *Scholarpedia* vol. 2, ed, 2007.

[179] L. Festinger, *A theory of cognitive dissonance*. Stanford university press, 1962.

[180] A. Ojha, F. Ervas, and E. Gola, "Emotions as intrinsic cognitive load: an eye movement analysis of high and low intelligent individuals," in *IEEE International Conference on Cybernetics (CYBCONF)*, Exeter, UK, 2017, pp. 1-6.

[181] G. Chanel, C. Rebetez, M. Bétrancourt, and T. Pun, "Boredom, engagement and anxiety as indicators for adaptation to difficulty in games," in *Proceedings of the 12th international conference on Entertainment and media in the ubiquitous era*, Tampere, Finland, 2008: ACM, pp. 13-17.

[182] G. Chanel, C. Rebetez, M. Betrancourt, and T. Pun, "Emotion assessment from physiological signals for adaptation of game difficulty," *IEEE Transactions on Systems Man and Cybernetics Part a-Systems and Humans,* vol. 41, no. 6, pp. 1052-1063, 2011.

[183] B. Kerous, F. Skola, and F. Liarokapis, "EEG-based BCI and video games: a progress report," *Virtual Reality,* vol. 22, no. 2, pp. 119-135, 2018.

[184] M. Callejas-Cuervo, L. A. Martínez-Tejada, and A. C. Alarcón-Aldana, "Emotion recognition techniques using physiological signals and video games-systematic review," *Facultad de Ingeniería,* vol. 26, no. 46, pp. 19-28, 2017.

[185] P. Taylor, *Text-to-speech synthesis*. Cambridge university press, 2009.

[186] B. Nirmalkar and S. Kumar, "A review paper on voiced/unvoiced classification," *IJSRD - International Journal for Scientific Research & Development,* vol. 4, no. 04, pp. 1065-1070, 2016.

[187] T. Giannakopoulos, "A method for silence removal and segmentation of speech signals, implemented in Matlab," *University of Athens, Athens,* vol. 2, 2009.

[188] S. G. Koolagudi and K. S. Rao, "Emotion recognition from speech: a review," *International journal of speech technology,* vol. 15, no. 2, pp. 99-117, 2012.

[189] J. W. Picone, "Signal modeling techniques in speech recognition," *Proceedings of the IEEE,* vol. 81, no. 9, pp. 1215-1247, 1993.

[190] K. R. K. Anne, Swarna Vankayalapati, Hima Deepthi, "Emotion recognition using prosodic features," in *Acoustic modeling for emotion recognition*, 1 ed.: Springer International Publishing, 2015, ch. 2, pp. 7-15.

[191] P. Boersma, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound," in *Proc. Inst. Phonetic Sci*, 1993, pp. 97–110.

[192] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," in *Readings in speech recognition*, A. Waibel and K.-F. Lee Eds.: Elsevier, 1990, pp. 65-74.

[193] S. Young *et al.*, *The HTK book (version 3.2)*. Cambridge Universit, Engineering Department, 2002.

[194] J. F. Blinn, "What's that deal with the DCT?," *IEEE Computer Graphics and Applications,* vol. 13, no. 4, pp. 78-83, 1993.

[195] S. S. Stevens, J. Volkmann, and E. B. Newman, "A scale for the measurement of the psychological magnitude pitch," *The Journal of the Acoustical Society of America,* vol. 8, no. 3, pp. 185-190, 1937.

[196] R. C. Snell and F. Milinazzo, "Formant location from LPC analysis data," *IEEE transactions on Speech and Audio Processing,* vol. 1, no. 2, pp. 129-134, 1993.

[197] J. E. Markel and A. H. Gray, *Linear prediction of speech*. Springer-Verlag, 1982.

[198] T. M. Cover and J. A. Thomas, *Elements of information theory*. John Wiley & Sons, 2012.

[199] P. D. Turney and P. Pantel, "From frequency to meaning: vector space models of semantics," *Journal of Artificial Intelligence Research,* vol. 37, pp. 141-188, 2010.

[200] G. Salton, A. Wong, and C. S. Yang. (1975) A vector space model for automatic indexing. *Communications of the ACM*. 613-620.

[201] X. Huang and L. Deng, "An overview of modern speech recognition," in *Handbook of Natural Language Processing*, N. Indurkhya and F. J. Damerau Eds., 2 ed.: Chapman & Hall/CRC, 2010.

[202] M. Gales and S. Young, "The application of hidden Markov models in speech recognition," *Foundations and Trends in Signal Processing,* vol. 1, no. 3, pp. 195-304, 2007.

[203] D. Jurafsky and J. H. Martin, *Speech and Language Processing: An introduction to speech recognition, computational linguistics and natural language processing*, 3 ed. 2018.

[204] G. Hinton *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine,* vol. 29, no. 6, pp. 82-97, 2012.

[205] S. Katz, "Estimation of probabilities from sparse data for the language model component of a speech recognizer," *IEEE transactions on acoustics, speech, and signal processing,* vol. 35, no. 3, pp. 400-401, 1987.

[206] S. F. Chen and J. Goodman, "An empirical study of smoothing techniques for language modeling," *Computer Speech & Language,* vol. 13, no. 4, pp. 359-394, 1999.

[207] S. Furui, "Speaker-independent isolated word recognition using dynamic features of speech spectrum," *IEEE Transactions on Acoustics, Speech, and Signal Processing,* vol. 34, no. 1, pp. 52-59, 1986.

[208] B. S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *the Journal of the Acoustical Society of America,* vol. 55, no. 6, pp. 1304-1312, 1974.

[209] A. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," *IEEE transactions on Information Theory,* vol. 13, no. 2, pp. 260-269, 1967.

[210] A. Ljolje, F. Pereira, and M. Riley, "Efficient general lattice generation and rescoring," in *Sixth European Conference on Speech Communication and Technology (EUROSPEECH),* 1999, pp. 1251-1254.

[211] A. Ogawa, K. Takeda, and F. Itakura, "Balancing acoustic and linguistic probabilities," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Seattle, WA, USA, 1998, vol. 1: IEEE, pp. 181-184.

[212] J.-L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE transactions on speech and audio processing,* vol. 2, no. 2, pp. 291-298, 1994.

[213] P. C. Woodland, "Speaker adaptation for continuous density HMMs: A review," in *Proc. ISCA Workshop on Adaptation Methods for Speech Recognition,* Sophia Antipolis, France, 2001, pp. 11-19.

[214] R. Bippus, A. Fischer, and V. Stahl, "Domain adaptation for robust automatic speech recognition in car environments," in *Proceedings of the Sixth European Conference on Speech Communication and Technology (EUROSPEECH),* 1999, pp. 1943-1946.

[215] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer speech & language,* vol. 9, pp. 171-185, 1995.

[216] C. Leggetter and P. Woodland, "Flexible speaker adaptation using maximum likelihood linear regression," in *Proc. ARPA spoken language technology workshop*, 1995, pp. 110-115.

[217] J. R. Bellegarda, "Statistical language model adaptation: review and perspectives," *Speech communication,* vol. 42, no. 1, pp. 93-108, 2004.

[218] C.-H. Chueh and J.-T. Chien, "Topic cache language model for speech recognition," in *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP),* Dallas, TX, USA, 2010, pp. 5194-5197.

[219] N. J. Cooke and M. J. Russell, "Cache-based language model adaptation using visual attention for asr in meeting scenarios," in *Proceedings of the international conference on Multimodal interfaces,* Cambridge, Massachusetts, USA, 2009, pp. 87-90.

[220] S. Qu and J. Chai, "An exploration of eye gaze in spoken language processing for multimodal conversational interfaces," in *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL),* 2007, pp. 284-291.

[221] N. G. Ward, A. Vega, and T. Baumann, "Prosodic and temporal features for language modeling for dialog," *Speech Communication,* vol. 54, no. 2, pp. 161-174, 2012.

[222] N. G. Ward and A. Vega, "Towards empirical dialog-state modeling and its use in language modeling," in *Interspeech,* 2012, pp. 2314-2317.

[223] R. Kneser, J. Peters, and D. Klakow, "Language model adaptation using dynamic marginals," in *Proceeding of the Fifth European Conference on Speech Communication and Technology (EUROSPEECH),* Rhodes, Greece, 1997, pp. 1971–1974.

[224] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals," *Soviet Physics Doklady,* vol. 10, 1966.

[225] N. Cooke, A. Shen, and M. Russell, "Exploiting a 'gaze-lombard effect' to improve asr performance in acoustically noisy settings," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* Florence, Italy, 2014, pp. 1754-1758.

[226] S. F. Chen, D. Beeferman, and R. Rosenfeld, "Evaluation metrics for language models," in *Proceedings of DARPA Broadcast News Transcription and Understanding Workshop*, 1998, pp. 275-280.

[227] M. Bisani and H. Ney, "Joint-sequence models for grapheme-to-phoneme conversion," *Speech communication*, vol. 50, no. 5, pp. 434-451, 2008.

[228] M. Suzuki, R. Tachibana, S. Thomas, B. Ramabhadran, and G. Saon, "Domain adaptation of CNN based acoustic models under limited resource settings," in *Interspeech*, San Francisco, USA, 2016, pp. 1588-1592.

[229] D. Povey *et al.*, "The Kaldi speech recognition toolkit," in *Proceedings of the Automatic Speech Recognition & Understanding (ASRU) Workshop*, 2011.

[230] A. Stolcke, "SRILM-an extensible language modeling toolkit," in *Seventh international conference on spoken language processing*, 2002.

[231] S. Ghosh, M. Chollet, E. Laksana, L.-P. Morency, and S. Scherer, "Affect-lm: A neural language model for customizable affective text generation," in *Proc. ACL'17*, 2017, pp. 634–642.

[232] J. Hansen and S. Bou-Ghazale, "Getting started with SUSAS: a speech under simulated and actual stress database," in *Proceedings of the Eurospeech 1997*, Rhodes,Greece, 1997, vol. 5, pp. 2387–2390.

[233] L. J. Rodríguez and I. Torres, "Comparative study of the baum-welch and viterbi training algorithms applied to read and spontaneous speech recognition," in *Pattern Recognition and Image Analysis. IbPRIA 2003. Lecture Notes in Computer Science*, vol. 2652, F. J. Perales, A. J. C. Campilho, N. P. de la Blanca, and A. Sanfeliu Eds.: Springer, Berlin, Heidelberg, 2003, pp. 847-857.

[234] D. Reynolds, "Gaussian mixture models," *Encyclopedia of biometrics*, pp. 827-832, 2015.

[235] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the royal statistical society. Series B (methodological)*, vol. 39, pp. 1-38, 1977.