Analysis of Gender Differences in Speech and Hand Gesture Coordination for the design of Multimodal Interface Systems

By

Jing Liu

A thesis submitted in fulfillment
of the requirements for the degree of
Doctor of Philosophy

Department of Computing
Faculty of Science
Macquarie University

Supervisor: A/Prof. Manolya Kavakli

June 2014

Statement of Candidate

I certify that the work in this thesis entitled Gender Differences in Speech and Hand Gestures for Their Integration in Multimodal Interface Systems has not previously been submitted for a degree nor has it been submitted as part of requirements for a degree to any other university or institution other than Macquarie University.

I also certify that the thesis is an original piece of research and it has been written by me. Any help and assistance that I have received in my research work and the preparation of the thesis itself have been appropriately acknowledged.

In addition, I certify that all information sources and literature used are indicated in the thesis.

Jing Liu

Acknowledgements

Foremost, I would like to express my sincere gratitude to my supervisor Associate professor Manolya Kavakli, for her support, supervision, encouragement and suggestions. Her wide knowledge and her passion in research and life have been of great value for me. Literally, without her continuous support and endless guidance, this work would not have been possible. It has been a great honour to be her PhD student.

In addition, my colleagues have helped me to develop this work. I wish to express my thanks to Dr. Iwan Kelaiah, Dr. Yifan Gao and John Porte for providing a friendly and enjoyable environment during these years.

I sincerely thank all the people in the Department of Computing, Faculty of Science, Macquarie University, for their warm support and help.

Last but not the least, I would like to dedicate this thesis to my parents, my husband Yong Yang and my daughter Iris, for their love, patience, and understanding.

List of Publications

- Jing Liu, Manolya Kavakli *Towards Accommodating Gender Differences in Speech-Gesture Based Multimodal Interface Design.* 8th International Conference on Information Technology and Applications (ICITA), July, 2013, Sydney, Australia. 2013

- Jing Liu, Manolya Kavakli *The Correlation of Speech and Hand Gestures for Multimodal Web Interaction.* International Conference on e-Learning, e-Business, EIS, and e-Government (EEE), July, 2013, Las Vegas, Nevada, USA.

- Jing Liu, Manolya Kavakli *Temporal Relation between Speech and Co-verbal Iconic Gestures in Multimodal Interface Design.* Gesture and Speech in Interaction (GESPIN), September, 2011, Bielefeld, Germany.

- Jing Liu, Manolya Kavakli *Hand gesture recognition based on segmented singular value decomposition.* Knowledge-Based and Intelligent Information and Engineering Systems, pages 214–223, July, 2010, Cardiff, UK.

- Jing Liu, Manolya Kavakli *A survey of speech-hand gesture recognition for the development of multimodal interfaces in computer games.* 2010 IEEE International Conference on Multimedia and Expo (ICME), pages 1564–1569, July, 2010, Singapore.

Abstract

The research studies about Multimodal Interface Systems (MMIS) involving speech and hand gestures have intensified in the past three decades. Understanding the correlations of speech and hand gestures has gained significance in MMIS design. Gesture is known to correlate with speech in a number of levels in general, but less is known about the gender differences in this kind of correlation.

When users interact multimodally with MMIS, we hypothesise that there are gender differences in the coordination of speech and hand gestures internally and externally. The investigation of such user related factors can benefit MMIS through accommodating gender adaptive processing strategies for different gender groups which can potentially improve the system performance. The main methodology used in this thesis is video annotation, including hand gesture annotation and speech annotation, to identify the gender differences in the descriptions of two objects using speech and hand gestures.

Our aim is to search for answers to the following questions:

Firstly, are there any gender differences in the coordination of speech and hand gestures? We found that females use more hand gestures than males for the same task. This may imply that females and males have different preferences in using speech and gestural modalities in MMIS. The temporal integration patterns are similar for males and females, but the temporal alignment intervals of gesture strokes and corresponding lexical affiliates are shorter for females than males.

Secondly, do males and females employ different cognitive processing models in

the coordination of speech and hand gestures? Our findings demonstrated that males and females have different distribution in cognitive actions. In general, males have more perceptual actions than functional actions, while females have more functional actions than perceptual actions. Gender differences in cognitive processing models might be the reason for the differences in the distribution of word types accompanying hand gestures. This implies that MMIS can potentially achieve better performance if information processing strategies are designed for different gender groups.

Thirdly, are there any differences in brain activities of males and females, when speech accompanies hand gestures? Our findings showed that the differences in lateralisation of brain activities associated with speech and hand gestures are quite minor in gender. However, we found that females show stronger beta spectral moment and more significant changes in spectral moment from alpha to beta band. This may explain the shorter temporal alignment of speech and hand gestures for females.

We demonstrated that gender differences in speech and hand gestures occur both internally (in cognitive processing and brain activities) and externally (in the presentation of speech and hand gestures). Based on the external differences, we developed models to predict the gender of users by evaluating their multimodal actions (using decision tree, neural network and logistic regression respectively). Our results show that a reasonable performance can be achieved by logistic regression model with an accuracy over 70%. Thus, we demonstrated that various gender prediction models can be successfully implemented using our findings and our results are promising for the design of gender adaptive MMIS.

Contents

List of Figures

List of Tables

1

Introduction

1.1 Multimodal Interface Systems (MMIS)

Mulitimodal Interface Systems (MMIS) are defined as systems that include different types of input methods beyond the traditional keyboard and mouse input/output, such as natural speech, facial expression, handwriting and manual gestures [2]. Even brain wave signals can be used in Human Computer Interaction (HCI) as an input mode [3]. Ideally, MMIS can process all combined user input modes to facilitate the overall HCI performance.

The possibility and ability to develop MMIS are upheld by the desire to simulate human cognition. Apparently human beings make full use of all their available modalities when they communicate with others. They can acquire and convey information

through different modes (such as speech, hand gestures and facial expressions). Physically, sensory information from different modalities has different nerve-pathways to the primary sensory area and can be parallel-processed [4]. However, interconnection also exist in some brain areas for multimodal integration and dispersion. Some researchers [5, 6] provided evidence that structures for multimodal perception and cognition in human physiology appear to be collaborative and to accommodate multimodal information interaction.

The main aim of MMIS is to make HCI more similar to human-human communication in future. One promising aspect of MMIS is their flexibility in providing users with a choice of input. They offer greater accessibility to a wide range of users with better performance than a single-modality system. They can also accommodate adaptability in switching modes as necessary. The synchronous input possibilities provided by MMIS allow for more flexible and efficient input. MMIS can also take advantage of mutual disambiguation to improve error correcting capability of the whole system Nowadays, research about MMIS has been more centralised on the integration of various user input modalities in a natural way.

Besides conventional direct-manipulation devices such as keyboard, mouse, and touch screens, technologies used in MMIS input modes also benefit from more advanced recognition technologies such as speech recognition, gesture recognition, lip movement, gaze tracking, face tracking and even the detection of brain waves. The development of MMIS dates back to Bolt's [7] original "Put That There" demonstration system which processed simple speech and hand pointing commands. The most mature research in the field of MMIS to date combines speech with hand gestures in pointing, handwriting, or lip movement tracking. With the development of technologies to track hand gestures (e.g. Data Gloves, magnetic trackers and vision-based approaches), the focus of MMIS research has become the integration of speech and hand gestures (rather than only pointing gestures).

Many efforts have been made to improve the performance of unimodal interpreters using speech and hand gesture recognition in past decades. For example, speech recognition accuracy has achieved a score of over 95% for small vocabularies [8]. The

accuracy for discrete hand gesture recognition has also reached a score of over 90% [9]. The vision-based hand gesture recognition has also achieved a classification accuracy of over 95% [10]. There are also attempts to improve the performance of bi-modal interpreters integrating speech and hand gestures [11, 12].

However, in order to build MMIS, as shown in Fig. 1.1, each level of the structure needs to be investigated and aligned properly to achieve better performance. At the top of this structure is the user. Users' gender, cultural background and age range may have impact on MMIS design. There might be users with various disabilities. Investigation of the user-related factors in MMIS input modes can provide useful insight for the design and implementation of adaptive processing strategies for MMIS design. However, user-related factors such as expertise, age, cultural or ethnic background and gender of the user have received far less attention in MMIS design.

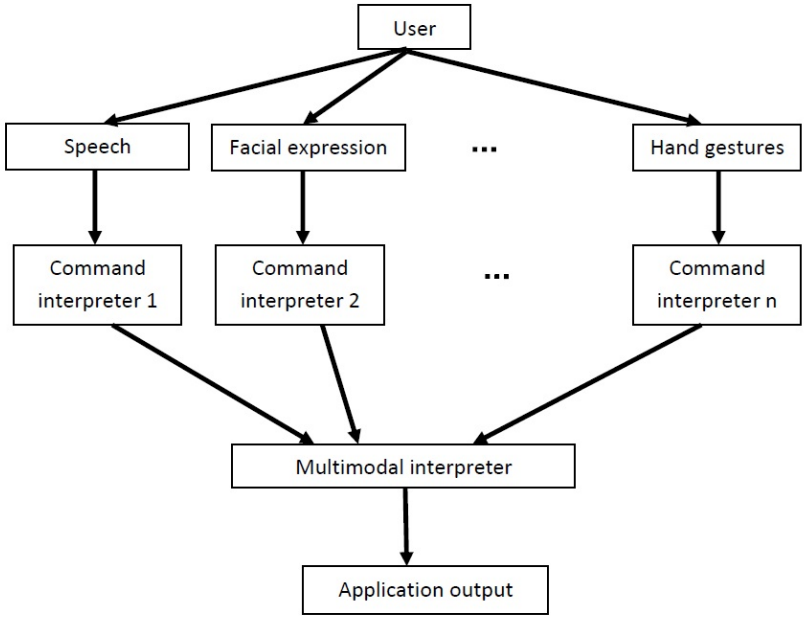


Figure 1.1: A typical MMIS struture

How user-related factors affect the design of MMIS is still veiled but has already started to attract research attention. For example, the research on multimodal integration of pen and voice input has found two distinctive types of users: ones who present

speech and pen commands in an overlapped or simultaneous manner and others who deliver signals sequentially with speech input lagging pen input [13]. These studies also found that everyone has a dominant integration pattern that is 95-96% consistent and kept stable over time [14]. In some earlier work, cultural differences have also been reported between users in the integration of speech and lip movement modalities [15, 16]. Cultural differences also have an impact on gesture-based interfaces [17]. These findings support that future MMIS could achieve greater performance through accommodating adaptive strategies for different user groups.

Gender as a user-related factor in the design of MMIS has not been extensively studied, even though it has been addressed in some other areas. In this thesis, our aim is to investigate gender differences in:

- the presentation of speech and hand gestures;
- cognitive processing in speech and hand gestures;
- brain activities when speech and hand gestures are used together.

We hope to shed light on user-adaptive systems with our findings regarding gender differences in speech and hand gestures. To this end, we present several studies that fit into an overall framework as in Fig. 1.2. Gender differences in speech and hand gestures will be investigated internally (such as differences in cognitive processing and brain activities associated with speech and hand gestures) and externally (the presentation of speech and hand gestures). The internal differences might be the reasons for the external differences. Gender prediction models will be built based on the external differences which can potentially benefit the design of MMIS.

1.2 Research Problems and Hypotheses on Gender Differences in MMIS

Gender differences actually have been broadly investigated in sociology and psychology. Two views are formed in regard to gender differences. One view asserts that gender

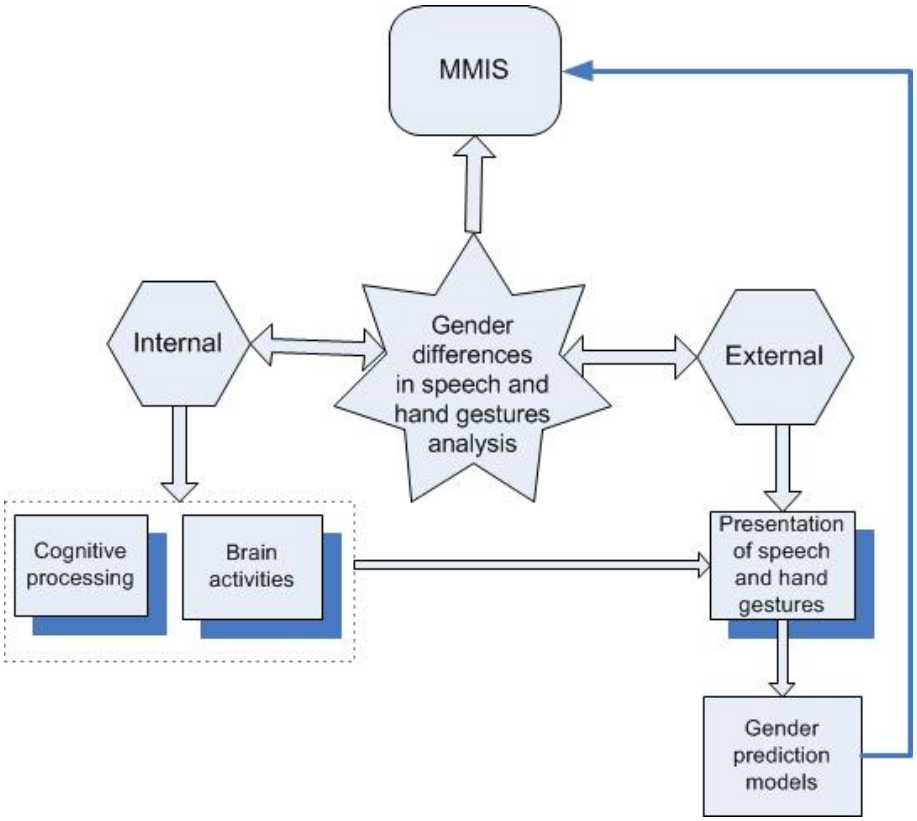


Figure 1.2: Thesis framework

differences are defined by our genetic makeup and exist from birth. Another view supports that gender differences are shaped by society and culture and formed after birth. For example, there are gender differences in the use of words [18, 19, 20, 21, 22], but apart from the language, are there any gender differences in the use of gestures?

In this thesis, our focus is the gender differences in the use of speech and gestures for MMIS design. It has been proven that speech and gestures share the same communication system [23]. As indicated in [13] that 'multimodal language does not differ linguistically from unimodal language', the general hypothesis in this thesis is that speech and hand gestures are integrated systems, but there are gender differences in processing of these two types of input. We investigate whether or not there are gender differences in the lexical affiliates of hand gestures. We assume that when speech is coordinated with hand gestures, males and females may use different keywords (lexical affiliates) to accompany gestures in particular tasks. We also investigate if males or

females have different preferences in using a specific vocabulary of speech and hand gestures.

As the structure in Fig. 1.1 shows, the performance of each input mode contributes to the performance of multimodal interpreter. Their integration, then, further affects the performance of the whole system. Users' gender has also been reported to have an impact on the performance of speech recognition systems [24, 25, 26, 27, 28]. Abdulla and Kasabov demonstrated that significant improvement could be achieved in the accuracy of word recognition in a gender-dependent database over a gender-independent approach [29]. Some studies have already been able to identify gender from speech due to the differences in their speech signals [30, 31, 32, 33, 34, 35, 36]. These studies demonstrate that there are gender differences in the speech characteristics.

Even though hand gestures have been broadly investigated in the past decade, users' gender has been mostly ignored in these studies. There are some general conclusions about gender differences in gestures, stating that females are more likely to use facial expressions and hand gestures to express their thoughts than males [37], and emphasising that men rely on more obvious gestures, while women use more subtle gestures [38]. As demonstrated in theses studies, only a small group of researchers to date focused on gender differences in the characteristics of hand gestures and gender prediction using hand gestures.

It has been presented that speech and gestures correlate with each other and gestures normally precede or synchronise with their corresponding keywords [39, 40, 41, 42, 43, 44, 45, 23, 46]. However, none of these investigations has considered in gender differences in studying the correlations or alignment of speech and hand gestures. In this thesis, we are interested in the coordination of speech and hand gestures.

The first hypothesis in this thesis is as follows:

H1: There are gender differences in the coordination of speech hand gestures as well as their temporal alignment in mutimodal information processing.

If so, it must be possible to make use of these different types of input to predict the users' gender through evaluating the multimodal interactions. This can be beneficial

for providing guidelines for the integration of multimodal input in MMIS.

As discussed in detail in Chapter 2, Cognitive scientists have explored multimodal cognitive processing at different levels [4]. According to Baddeley's Working Memory Theory, working memory comprises independent model processors (namely, Central Executive, Phonological Loop, Visual Spatial Sketchpad and Episodic Buffer) which are used to deal with specific productions (e.g. speech and hand gestures). These processors work together coordinately and synchronously [47]. A number of researchers have also demonstrated that working memory can be used more effectively to expand processing capabilities by presenting information in a dual-mode rather than in a single one [48]. According to the Multimodal Resource Theory [49], multiple but limited resources are available for information processing in human cognition. Various allocation strategies for limited resources in multimodal processing can be considered as possible reasons for high performance. The Cognitive Load Theory also builds on the assumption that working memory has limited capacity and duration [50]. According to this theory, subjects' preference and also superior performance in multiple modalities can be explained by ascribing the advantages of multimodal processing to effectively expanding of working memory [51, 52]. These theories support the fact that there are advantages to using multimodal processing in system design.

There are also studies demonstrating that gesturing is helpful in lightening a speaker's cognitive load [53] and also a user-centered interface design can free up mental resources and further improve user performance [54, 55]. Some studies show that individuals working memory load can affect the interaction between co-verbal hand gestures and working memory [56]. Based on these studies gender differences in cognitive processing is a significant research problem for the design of MMIS.

The second hypothesis is as follows:

H2: Males and females employ different cognitive processing models in the coordination of speech and hand gestures. Gender differences in cognitive processing might be a reason for the differences in the presentation of speech and hand gestures.

In this thesis we also investigate brain activities in speech and co-occurring hand

gestures of males and females. There are research studies demonstrating gender differences in the location of brain activities and information processing [57]. Males and females show differences in the strength of the brain activation linked to word generation [58]. These findings about gender differences in brain activities also support the claim that male brains are more lateralised than female brains with functions spread over both hemispheres of female brains. However, there are some studies arguing that gender differences in language lateralisation may be only observed in special cases and are absent at the population level [59].

The third hypothesis in this thesis is

H3: There are gender differences in brain activities in the coordination of speech and hand gestures.

We investigate if the lateralisation of the brain activities can still be observed in speech and hand gestures coordination for males.

In summary, we propose in this thesis that speech and hand gestures are integrated systems, but there are gender differences in their coordination internally and externally. The research questions corresponding to each hypotheses in this thesis are as follows:

RQ1: Are there any gender differences in using speech and hand gestures? What are the similarities and differences between males and females in the coordination of speech and hand gestures given the same tasks? Are there any gender differences in the integration or temporal alignment patterns of speech and hand gestures?

RQ2: Do males and females employ different cognitive processing models in the coordination of speech and hand gestures? Are the differences in the presentation of speech and coordinated hand gestures driven by the differences in cognitive processing?

RQ3: Are there any gender differences in brain activities in speech and coordinated hand gestures? Is the male brain more lateralising in the coordination of speech and hand gestures?

Based on these research questions, the goals of this thesis are introduced in the next section.

1.3 Goals of the Thesis

Based on our hypotheses and research questions, the main research objectives we are targeting are to:

G1: Discover gender differences in using speech and hand gestures.

G2: Investigate the differences in the temporal alignment of speech and hand gestures of males and females.

G3: Build models to recognise gender from speech and hand gestures.

G4: Study gender differences in the coordination of speech and hand gestures.

G5: Examine the gender differences in brain activities associated with the coordination of speech and hand gestures.

1.4 Methodology

We conducted two sets of experiments in this research to investigate gender differences in speech and hand gestures. In the first experiment, participants were required to describe two objects using speech and hand gestures naturally. Hand gestures were transcribed according to McNeill's definition of gesture categories from the collected video clips [40]. The lexical affiliates of hand gestures were also extracted from the corresponding audio clips. We investigated the gender differences in the annotations of hand gestures and their corresponding lexical affiliates. We analysed cognitive processes of males and females in the video/audio clips collected in the first experiment using protocol analysis. We coded the cognitive actions of participants using a coding scheme developed by Suwa et al. [1, 60].

In the second experiment, we investigated gender-based brain activities using electroencephalogram (EEG) signals collected from a practical EEG device - Emotive Neuroheadset. The brain wave signals were collected when participants use speech and hand gestures together. We used spectral moment analysis to analyse EEG signals to investigate the differences in the brain activities of males and females.

Gender differences in cognitive processing and brain activities in speech and hand

gestures can be viewed as internal differences, while differences in the usage and presentation of speech and gestures can be regarded as external differences which can be detected and predicted. We developed statistical models to predict gender through evaluating the speech and hand gestures. This might be beneficial for the design of MMIS accommodating the adaptive processing strategies for different gender groups to improve system performance.

1.5 Contributions of the Work

The work in this thesis contributes in four main aspects to the study of gender differences in speech and hand gesture coordination for MMIS through two sets of experimental studies:

- To the best of our knowledge, this work is the first research effort studying gender difference in speech and hand gestures coordination. We study the gender differences in the time intervals between the onset of gesture stroke phases and the corresponding lexical affiliates. Our findings support the claim that gestures share the same communication system with speech and gestures precede the related lexical affiliates in general. However, we found gender differences in the length of the time intervals, which have not been studied before.

- We use protocol analysis to study the cognitive processing in speech and hand gestures. Some researchers studied individual differences in cognitive processing regarding verbal ability or visual spatial ability. However, controversial views exist. To the best of our knowledge, this work is also the first attempt to study gender differences in cognitive processing in speech and hand gestures together. There may be a relation between gender differences in the cognitive processing and the presentation of speech and hand gestures. This work provides a broader view to this area.

- We design an experiment to collect EEG signals through Emotiv Neuroheadset when the participant only use speech and hand gestures. The EEG signals are

used to study the brain activation regarding the process of speech and hand gestures. Gender differences in brain activities have been reported in different tasks, but there are only a few about the tasks involving speech and hand gestures, which is our focus in this work. Our experimental results do not provide any evidence for gender differences regarding the brain lateralisation, but our findings show some gender differences in beta spectral moment, which may be the reason for gender differences we found in cognitive processing and presentation of speech and hand gestures.

- We investigate three statistical models (decision tree, neural network and logistic regression) for the prediction of user gender based on our analysis results. Compared to other methods using only speech or applying complicated algorithms, the models explored in this work can achieve a reasonable performance with a simple but effective approach.

1.6 Thesis Organisation

The remainder of this thesis is organised as follows:

- Chapter 2 reviews the previous related research efforts on speech and gesture based MMIS. It introduces a number of theories in human cognition that may support the development of more effective MMIS, as well as gender studies.

- Chapter 3 describes the first experiment in which the video clips were collected. It also introduces the methodology for the analysis of gender differences in speech and co-occurring hand gestures. It presents the evaluation of experimental results at the end of this chapter.

- Chapter 4 explains the second experiment conducted to study gender differences in brain activities. It includes the methods used to analyse EEG signals and the discussion of gender differences in the brain activities in speech and hand gestures.

- Chapter 5 explores the possibility to predict gender based on the gender differences found in Chapter 3 and Chapter 4. It also provides a critical evaluation on the performance of decision tree, neural network and logistical models for gender prediction using hand gestures and accompanied lexical affilates.

- Chapter 6 presents conclusion and suggestions for future work.

2

Literature Review

2.1 Introduction to Multimodal Interfaces

The increasing interest in MMIS design is inspired largely by shortening the distance between human-computer interaction (HCI) and face-to-face communication. This chapter will first focus on the general characteristics of MMIS.

As defined in [61] *MMIS process two or more combined user input modes (such as speech, pen, touch, manual gesture, gaze, and head and body movements) in a coordinated manner with multimedia system output. They are a new class of interfaces that aim to recognise naturally occurring forms of human language and behavior, and which incorporate one or more recognition-based technologies (e.g. speech, pen, vision).* To date, some of the human communication modalities (e.g. speech) have been extensively investigated. In controlled situations speech recognition rate has reached a high level

of performance. Commercially, successful products for speech recognition (e.g. IBM's Voice Type Application Factory) make it possible to use human voice in MMIS. At the same time, the use of human movements, especially hand gestures, has become a popular part of MMIS in recent years, which inspires a motivation for modeling, analysing and recognising hand gestures. The advances in gesture-based interfaces allow for many practical applications [62]. However, in interacting with computer systems, people prefer a combination of speech and gestures over either speech or gestures alone [63]. Different input modalities can complement each other, allowing greater expressiveness than each modality on its own. Table 2.1 displays the basic differences between traditional Graphic User Interfaces (GUI) and MMIS [4].

TABLE 2.1: Differences between GUI and MMIS

GUI	MMIS
single input stream	multiple input streams
atomic, deterministic	continuous, probabilistic
sequential processing	parallel processing
centralised architectures	distributed and time sensitive architectures

MMIS have been shown to improve error handling and reliability. There are studies claiming that users make 10% faster task completion and 36% less task-critical content errors while using MMIS compared to a unimodal interface [63]. The modalities can also enhance each other when similar concepts are expressed in many different ways, increasing reliability and decreasing mutual ambiguity in MMIS [63]. MMIS also provide greater expressive power, naturalness, flexibility and portability [13]. For example, in a text-editing session a user may delete a paragraph simply by circling the text and saying "delete"at the same time [64]. In a noisy environment noise may hamper the recognition of a spoken "delete" command, but the system can recover its meaning if it realises that the user has also drawn a circle on top of some text to emphasise the "delete" concept.

2.1.1 Modality

Modality is probably the most intuitive factor for MMIS. Some researchers state that human minds work in a modality-specific manner [65]. The selection of modalities for MMIS can affect the performance of the system to a great extent. The definition of modality varies from one field to another.

In computer science, modality can simply be defined as the form in which information is presented or exchanged (such as text, graph, touch, speech, and gestures) [66]. Each specific form of information is transferred to computer systems by users through specific media. For example, we normally use keyboard for text input. Speech is captured by a microphone. Gestures can be recorded by a camera. Different modalities have different properties and representation types. As such, a specific modality is more suitable for presenting certain types of information than others. In cognitive science, modality is commonly interpreted as the types of human sensation, namely vision, hearing, touch, smell and even taste [66]. Some computer input modes normally correspond to related human senses: cameras (vision), microphone (hearing), sensors (touch), olfactory (smell). Some other input modes, however, do not map directly to human senses, for example, keyboard, mouse or using tablet with a pen. Among different categories of modalities, visual and auditory modalities apparently dominate both in theoretical research area and in the practical application studies.

Input modalities are commonly distinguished from output modalities in MMIS studies [67]. For MMIS systems, input modalities normally carry information sent from users to the system, while output modalities deliver information generated by the system to users. Sometimes the same modality may rely on different media when it serves as input or as output. For example, hand gesture as an input modality may be carried out via a camera (users stand in front of a camera) or by a glove with colorful markers (users wear special gloves), while hand gestures as an output modality is realised via a display (the computer system displays hand gestures on the display). This thesis only addresses input modalities which are directly linked to users in MMIS.

2.1.2 User Input Modes

MMIS respond to more than one different user input mode such as speech, pen, touch, manual gestures, facial expression, gaze, and body movements in a coordinated way with MMIS output. MMIS become a new direction for the next generation computing, since they enable the paradigm to shift away from conventional GUI systems.

The earliest developed multimodal systems were probably ones that departed least from GUIs by including keyboard and mouse inputs. As speech recognition technology matured in the late 1980s and 1990s, these systems added speech input along with standard keyboard and mouse interfaces. The initial attempts used richer natural speech processing to support greater expressive power for the user [68, 69, 70, 71].

Bolt's "Put That There" demonstration system can be viewed as the original application of MMIS, which processes speech in parallel with manual pointing during object manipulation (see Fig. 2.1, a screenshot from [7]). In that system, the user communicated with a MMIS in a media room with a large screen display. The information from the hand was essentially transformed to a point on the screen by processing the x, y coordinates indicated by the Polhemus tracker. Semantic meaning from speech and pointing hand gestures (two modalities) was integrated to instruct interactions with MMIS. In the example command "Put That There", two deictic (pointing) hand gestures (one is referring to an object and one is for the intended location of the object) are integrated with spoken words to deliver information to MMIS about the mentioned object and location. In this system, voice is augmented with simultaneous pointing. The integration of two modalities disambiguate the command by matching information conveyed by speech and hand gestures.

Since Bolt's early concept, MMIS have emerged quickly in the past three decades. The developments of MMIS have arisen in diverse sub-fields. The development of hardware and software techniques is of great importance in supporting key components incorporated within MMIS, like integrating parallel input streams. MMIS have also diversified to combine new modality inputs (for instance, speech and pen input [72, 73, 74, 75, 76, 77, 78], speech and lip movements [79, 80], speech and hand or body

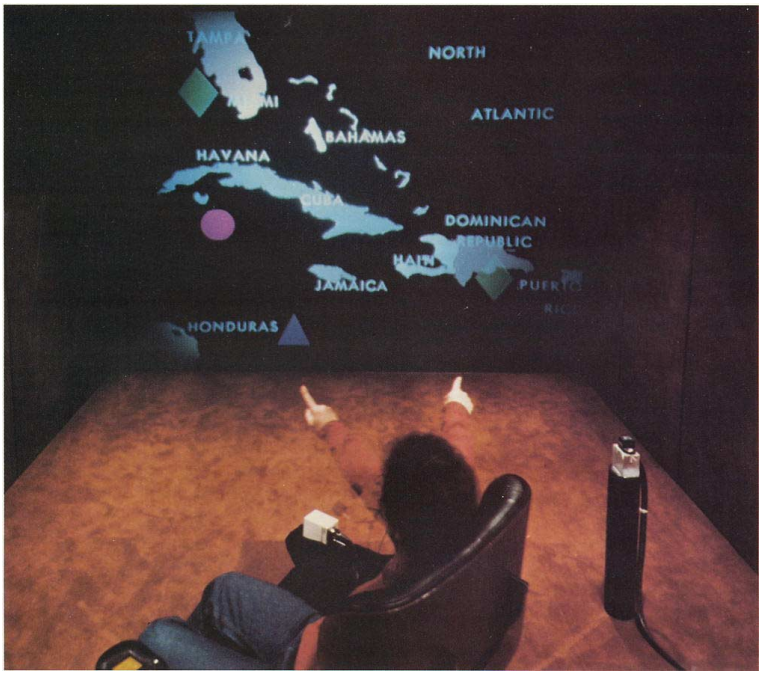


Figure 2.1: Bolt's "Put That There" system.

gestures [81, 82]). Modality choice is an important issue in the design of MMIS.

In more recent MMIS, input modalities have been expanded from simple mouse or touch-pad pointing to two parallel input streams like speech and pen input [72, 73, 74, 75, 76, 77, 78] or speech and lip movement [79, 80]. These MMIS aim to recognise two natural forms of input modalities which are capable of conveying rich information. In these systems, the traditional input (keyboard and mouse) are discarded. Among pen/voice MMIS, some still limit pen input to pointing [83, 84]. Some can process speech along with more complex symbolic pen-based hand gestural input [73].

The developments of multimodal speech and lip movement originate from cognitive science research about audio-visual perception and speech input with coordinated lip and facial movements [85, 86, 87, 80]. The classic work on speech and lip movement combination was conducted by Brooke and Petajan [85]. Benoit, et al. [88] illustrate some examples of systems and applications which are related to speech and lip movements. Other recent systems using speech and lip movements can be seen in [89, 90, 91, 92, 93, 94].

Beside the maturity of MMIS including speech and pen input or speech and lip movements, recognition of other input modes is also emerging and beginning to be applied to new kinds of MMIS. In particular, there is growing interest in combining speech and free manual gestures [81, 82]. These systems are different from Bolt's original "Put That There" system. In Bolt's system, the hand was essentially transformed to a point on the screen. The actual hand posture did not matter, even if it was not in a pointing shape. After the emergence of "Put That There" system, speak-and-point has been regarded as the prototype of multimodal design. Unfortunately, this type of MMIS similar to speak-and-point just implements the function for selection of objects as the mouse does. Actually linguistic analysis about spontaneous manual gestures during multimodal communication indicates that simple pointing gestures only account for less than 20% of all gestures [40]. Studies about users' integrated pen/voice input also suggest that speak-and-point patterns only comprise 14% of all spontaneous multimodal utterances [63]. In contrast, modern MMIS that transmit gesturing are capable of generating much more expressive information than simple pointing and selection.

Nowadays, with the development of new hardware and software technologies, people may expect MMIS to accommodate more input modes like manual hand gestures and even body gestures. People gesture when they talk. This can be observed from people of all ages, cultures and backgrounds. Gesture has been regarded as a cognitive aid in the realisation of thinking, and also a carrier of different semantic content than speech [95]:

"*speech and gestures are different material carriers ... they are not redundant but are related, and so the necessary tension can exist between them to propel thought forward... to make the gesture is to bring the new thought into being on a concrete plane.*"

Integrating hand gestures to MMIS will make human computer interaction as natural as human-to-human interaction. MMIS that incorporate hand gestures presently adopts different acquisition technologies to track hand shape as well as hand position to recognise hand gestures. The most common tracking technologies for free hand gestures include Data Gloves (or cybergloves), magnetic trackers and vision-based approaches.

Acquisition using glove-based tracking technique (including gloves and magnetic trackers) can track movements of each finger independently which is efficient and accurate. Some data gloves using haptic devices can allow the user not only to feed information to the computer but can receive information from the computer in the form of a felt sensation on some part of the body. But this kind of tracking technique suffers from the need to wear restrictive and cumbersome devices. Fig. 2.2 displays examples of 5DT data glove and cyberglove. Research studies about glove-based gesture recognition are mature with high recognition rates [96, 97, 98, 99, 9]. A comprehensive survey about glove-based systems and their applications have been detailed elsewhere [100].

Figure 2.2: 5DT Data Glove and Cyberglove

An alternative to data gloves is to use a camera and computer vision to track the 3D pose and trajectory of the hand, but at the cost of tactile feedback. Vision-based approach is to some extent restricted with in precision compared to glove-based acquisition technology [101, 102, 103, 104]. But the ultimate goal of MMIS is to make human computer interaction as natural as the human-to-human interactions which use verbal and non-verbal modes in parallel. In order to achieve it, devices and sensors used for MMIS should be transparent and passive as much as possible, and machines should capture relevant human communication modalities. Vision-based approach for gesture tracking could be a better choice for this purpose.

However, compared with other input modalities applied in MMIS, speech and natural hand gesture integration is less mature up to now. In the following sections, we will

introduce some issues related to the design of MMIS using speech and hand gestures.

2.1.3 Input Information Fusion

In human communication, the use of speech, lip movement or hand gestures is completely coordinated. Unfortunately, the devices used to interact with computers in MMIS have not been designed at all to cooperate with each other. For example, the differences between time responses of different devices can be very large. Unlike human beings who are naturally able to fuse multimodal signals and to interpret them to work out the information conveyed, computer systems with multimodal interfaces present challenges for the integration of complementary modalities to form a highly collaborative blend.

Basically, there are three levels for multimodal architectures designed to handle joint processing of input signals [105]:

Intuitively, information processing in MMIS always starts with signal or data acquisition from different channels. The captured signals or information can be called raw data. The first method is to integrate signals at the signal level (also called the data level). At this level, two or more raw signals captured from input devices, are combined together directly. At the data level, the possibility of information fusion can only be performed well for highly synchronous signals with the same traits (e.g. two webcams collecting the same scene from different angles). Otherwise, it is hard to obtain satisfactory performance.

After the completion of data acquisition, useful features or characteristics of data will be extracted for subsequent analysis. The second method is to fuse signals at the feature level, also called early fusion. Generally in a feature-fusion architecture, closely coupled and synchronised signals such as speech and lip movements are integrated. High synchronization is a typical issue associated with information fusion at this level. At this level a greater volume of data for training to get reliable features is required which results in an increasing computational intensity.

The last step for information processing is to interpret it to get the semantic meaning of the input information. So the third method is to combine information at the

semantic level or decision level, which is also called late fusion (contrasting to early fusion). Instead of directly mixing raw signals or features of signals together, this architecture extracts semantic information from individual modes respectively and integrates the sequential recognised results from each mode. Fusion input modalities at this level has the ability to manage loosely-coupled modalities (e.g. speech and hand gestures). It guarantees multimodal fusion advantages of steering clear of the requirement of synchronization issues. Generally, these individual modes in semantic-fusion MMIS can be trained and maintained separately by using unimodal data and can be changed according to system requirements without retraining. Table 2.2 summarises the characteristics of the three levels of fusion methods [4].

Table 2.2: Summary of fusion levels

	Signal-level fusion	**Features-level fusion**	**Semantic-level fusion**
Input type	Raw signal of same type	Closely synchronised	Loosely coupled
Level of information	Highest level of information detail	Moderate level of information detail	Mutual disambiguation by combining data from modes
Noise/failures sensitivity	Highly susceptible to noise or failures	Less sensitive to noise or failures	Highly resistant to noise or failures
Usage	Rarely used for combining multi-modalities	Used for fusion of particular modes	Most widely used type of fusion

Combining speech and hand gestures is a challenging task. First, they have totally different characteristics: speech consists of audio signals and gestures normally consist

of video signals or electrical signals according to the sensing techniques for hand gesture acquisition. Second, speech and gestural input channels may provide asynchronous but complementary information with different characteristics (e.g. time scales). Last but not the least, corpora for speech and gesture multimodal training are hard to obtain in the current state. Therefore, we can see from Table 2.2 that integrating speech and hand gestures at signal or feature level is subject to failure. Fusing information at sematic level is more suitable in this situation.

2.1.4 Frameworks for Input Information Fusion

An appropriate framework is one of the most important requirements for the design of multimodal systems. Basically, a multimodal system framework is required to fuse inputs from subsystems to handle message exchange between users and application systems. At least, the framework should support time stamping of the beginning and end of individual input. Speech and gesture streams are supposed to be delivered either sequentially or simultaneously [106]. Time is an essential factor in MMIS which integrate multimodal input modes. It is necessary to assign time stamps to all messages produced by the user.

Fig. 2.3 illustrates a typical framework using semantic-level fusion for speech and hand gestural input. User-related issues are the focus of the thesis, so we do not display the complete information processing procedures for a full multimodal dialogue system. Instead, we include here only a basic architecture for multimodal dialogue management.

This framework includes four main components: Input Analyser, Multimodal Manager, Output Designer, and Application Information Database [107]. In such a framework, speech and gestures are recognised in parallel, and each is processed by an input analyser. The results are semantic representations that are fused by output designer. Multimodal manager exchanges information between output designer and application information database to implement real-time control.

However, the use of homogeneous programming language in every part of MMIS

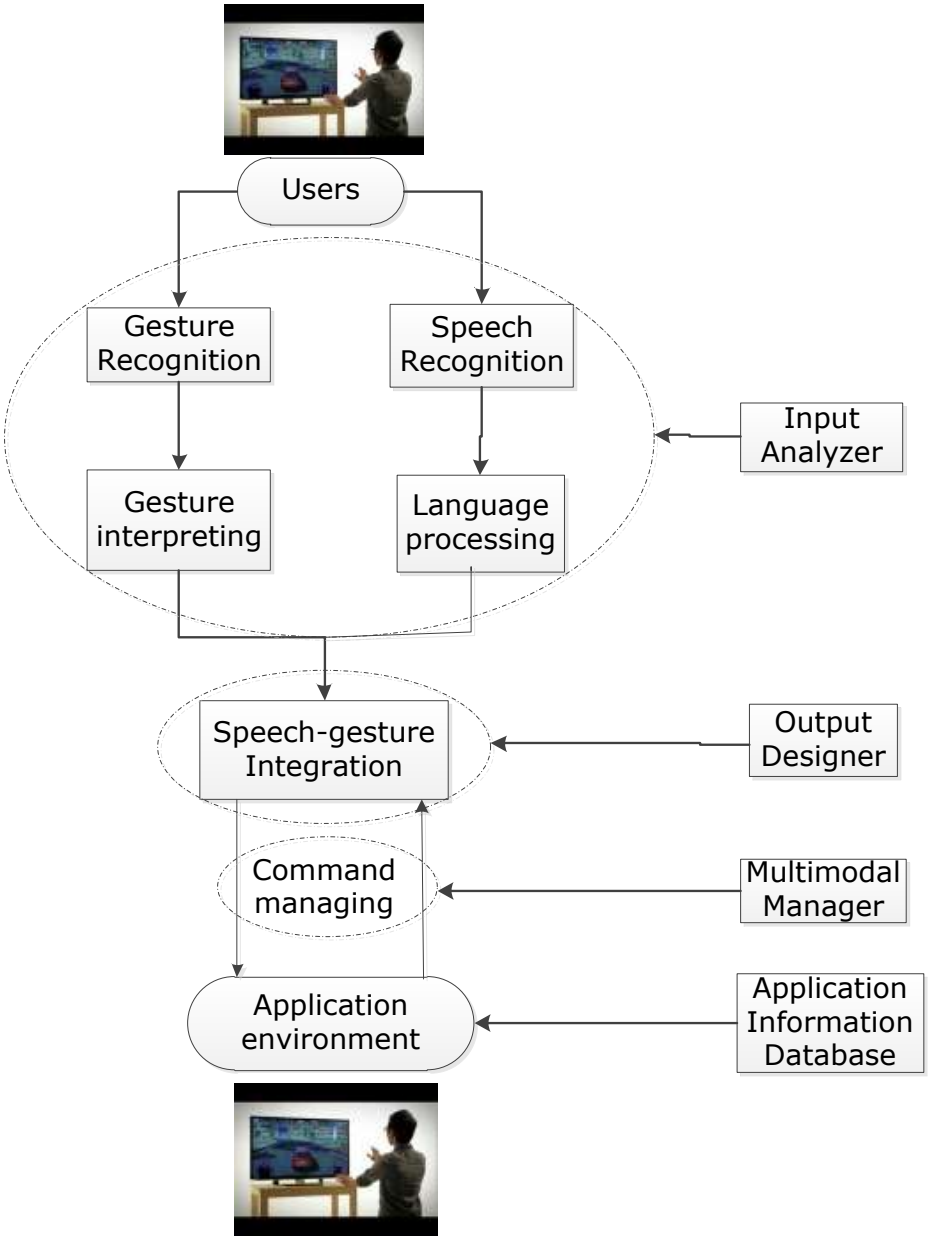


Figure 2.3: One framework for speech and gestures integration

is vital for the framework displayed in Fig. 2.3. This might be difficult in some cases. In order to overcome it, some researchers suggest multiagent architecture [105]. Wachsmuth [108] presents a method which conceptualises a multimodal user interface

on the basis of timed agent systems. They use multiple agents for the purpose of polling pre-semantic information from different sensory channels and integrating them to multimodal data structures. The data structures can be processed by an application system based on agent architectures. This kind of architecture provides an agent or a central facilitator which enables each component to communicate via a standard language over TCP/IP. Fig. 2.4 illustrates the basic framework based on the facilitator for MMIS.

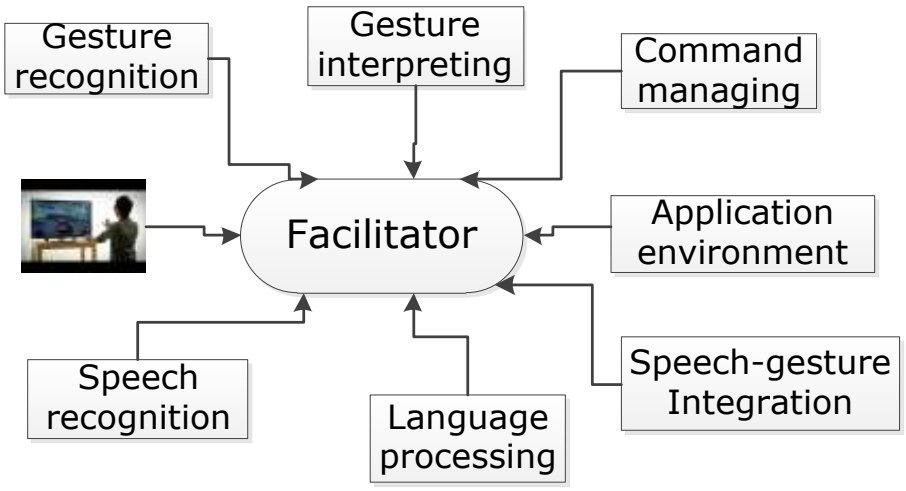


FIGURE 2.4: Facilitator for speech and gesture multimodal interface

As indicated in [109], one of the unique characteristics of MMIS is its time-sensitivity. Temporal constraints need to be explored on multimodal information processing. All the modalities used in MMIS must be properly time-stamped and integrated synchronously. In the time-sensitive architecture of MMIS, temporal thresholds for time-stamping of start and end of each input signals have to be established, in order to identify two command sequences. Apparently, when two commands are performed together in parallel, it is important to figure out in which order the commands occur and need to be integrated since the interpretation will vary accordingly. There was an

example to explain the different interpreting strategies for a MMIS in which voice and gestures are used simultaneously to control a music player [4]:

- ⟨pointing⟩ "Play next track": will result in playing the track following the one selected with a gesture;
- "Play" ⟨pointing⟩ "next track": will result in first playing the manually selected track and then passing to the following at the time "next" is pronounced;
- "Play next track" ⟨pointing⟩: In this case, the system should interpret the commands as being redundant.

Recent research has actually found that some users integrate multi-commands simultaneously and some do it sequentially [14]. The two types of users are also found to keep their habitual integration pattern across the whole session. The special integration pattern can be detected almost immediately on the very first multimodal interaction. The different user integration patterns as well as preferences in using different types of input modes have been reported in different task domains [110, 111]. In short, empirical studies suggest that individual differences exist in the mulitmodal integration patterns.

This means that user-related factors are essential factors in determining the temporal thresholds in the design of MMIS. Ideally, if adaptive temporal thresholds can be applied for different users in MMIS, it could result in significant improvements in system processing speed and the accuracy of command integration. For example, it has been demonstrated systems that adjust the time window dynamically according to the user have superior performance to a system with fixed-duration time window [112]. Gender, age, culture and any other individual all could be influential factors in MMIS. For instance, children, adults and seniors have demonstrably different intermodal lags during sequential integration [110]. It means different time thresholds exist in different age groups. Gender as a user related factor is rarely taken into account when users interact with MMIS. Our aim in this thesis is to address basic gender differences in the use of speech and hand gestures. We hope to shed light on gender differences in temporal thresholds for speech and hand gesture interaction in MMIS.

2.2 Human Cognition and MMIS

The development of MMIS is supported by cognitive scientists at a number of levels [4]. In this section, we will introduce the main cognitive science findings that are relevant to MMIS.

2.2.1 Integrated Systems Hypothesis

As we all know, humans can acquire and produce information through different modalities (e.g. speech, facial expression and hand gestures). As explained in [113], the human communication channel consists of sensory organs, the central nervous system, various parts of the brain and muscles or glands. Sensory information from each individual modality has its own specific pathway to the primary sensory cortex and can be processed in parallel.

In the brain some association areas that exist for specific multimodal integration (input) and diffusion (output) are highly interconnected. The evidence has been provided for the claim that neural processing in language comprehension involves the simultaneous incorporation of information coming from a broader domain of cognition than only verbal semantics [5].

The neural evidence for similar integration of information from speech and gesture emphasises the tight interconnection between speech and gestures that are used together with speech (co-speech gestures). Using neuroimaging technology like functional Magnetic Resonance Imaging (fMRI), Dick et al. examined how gestures influence neural activity in brain regions associated with processing semantic information [6]. They showed that perceiving hand movements during speech modulates the distributed pattern of neural activation involved in both biological motion perception and discourse comprehension. The results from [114] confirmed the **integrated-systems hypothesis** about speech and gestures and demonstrated that gesture and speech form an integrated system in language comprehension. Multimodal perception and cognition structures in the human brain appear to have evolved to be collaborative and to produce multimodal information.

2.2.2 Tripartite Working Memory Model for MMIS

Baddeley and his co-workers proposed their **tripartite working memory** model as shown in Fig.2.5 [47]. This modal has become the dominant view in the field of working memory, which suggests that working memory consists of independent model processors that work together in a coordinated and synchronous manner. The four main components of this modal include:

1. The Central Executive: A supervisory system that controls and regulates cognitive processes. It also intervenes when interactions between the modal slave systems go astray. These different input modes are to be processed concurrently or in a coordinated way.

2. The Phonological Loop: As the first slave system, it is also called the articulatory loop that copes with verbal, auditory and linguistic tasks. It consists of two subcomponents: a short-term phonological store with auditory memory traces that can hold speech or acoustic information for 1-2 seconds; and an articulatory rehearsal component that is responsible for reviving the memory traces.

3. The Visual Spatial Sketchpad: As the second of three slave systems, it is assumed to exclusively process information about what we see. It is activated in the processing of imagery and spatial tasks. There are at least two types of functionality within this slave system: imagery-based functionality which allows the remembering and recall of colours, forms, shapes, textures etc; and spatial functionality that is involved in tasks such as navigation, map reading and route descriptions.

4. The Episodic Buffer: As the last slave system, it is used to link information across domains to integrate units of visual, spatial, and verbal information with time sequencing. It also temporally stores schema retrieved from long term memory or the central executive modal or either of the two other slave systems.

Baddeley's model gains support from psychological studies and neurological pathologies as well as empirical findings. It integrates a large number of findings from research

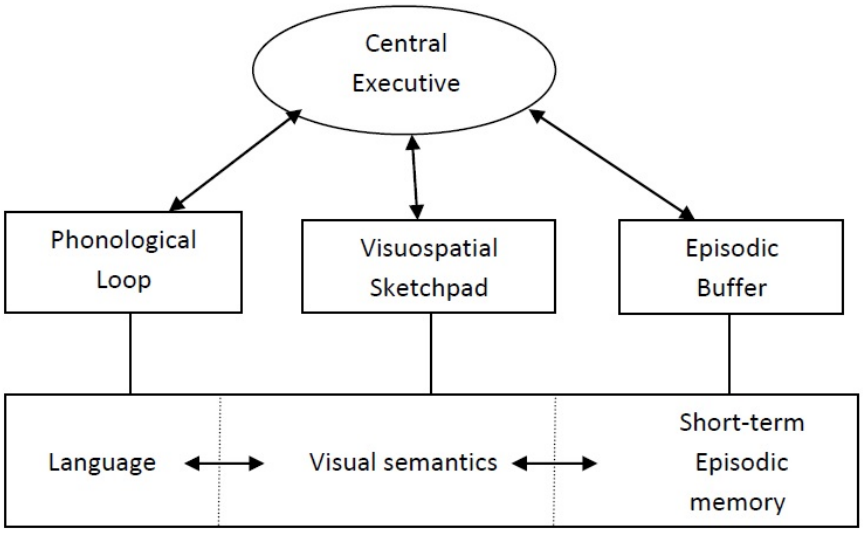


FIGURE 2.5: Schematic of Baddeley's working memory model

studies on short-term and working memory. The model can give a general view of multimodal information processing in human cognition. It suggests that different types of resources are probably used to deal with specific productions. For instance, speech information is likely produced by the phonological component while gestural input is likely to be managed by the visual/spatial component. Mousavi et al. [115] also suggest working memory might be partially independent processors for managing visual and auditory information. Tindall-Ford et al. [48] confirmed that by presenting information in a dual-mode form rather than a purely visual one, effective working memory can be increased and then further expand processing capabilities.

However, Baddeley's model emphasises the distinction between visual/spatial and auditory verbal processing, which makes it hard to explain integrated or interleaved cognitive processes that may involve both components, such as those that occur in multimodal production and many activities that involve both symbolic and linguistic information such as using speech and gestures to present images.

2.2.3 Multiple Resource Theory

The **Multiple Resource Theory** is quite different from the working memory model, and explains the various cognitive effects disclosed by empirical evidence. The motivation behind Multiple Resource Theory is to predict the level of performance and productivity of a human operator who carries out multiple real-world tasks [49] and the compatibility between the modalities used for input and output processing. Rather than modelling the structure of working memory, the theory is focused on the idea that we have multiple and limited modal resources available to process information in realtime. The models primary advantage is that it accounts for performance effects in highly taxing applied task combinations, such as driving and speaking on the phone, or monitoring aircraft visually and attending to incoming auditory messages.

Multiple Resource Theory (see Fig. 2.6) describes a set of modally-organised central resources that are taxed when a user completes a task. These resources have limited capacity and can be shared by multiple tasks being completed at once. Task interference occurs when two concurrent tasks requiring the same resource compete or interfere with one another, causing performance degradation on both tasks. The notion of task-interference is able to account for behaviour exhibited in situations of task overload, especially in tasks that require the same resource simultaneously, at some stage of processing. The corollary of this is that tasks that do not require the same resource at any stage of processing will not significantly interfere with one another and will therefore allow the subject to maintain a high level of performance [116]. This accounts for the higher level of performance in cross modal time-sharing tasks. In this account of modal cognitive resources, working memory itself is seen as a central processor that sits outside the modally structured resources.

One of the latest instantiations of Multiple Resource Theory is a four-dimensional Multiple Resource Model (see Fig. 2.7) that depicts the processes and resource-types described by Wickens. This model is used to analyse the modal and processing resources necessary for multiple task completion. Interference is predicted if the same

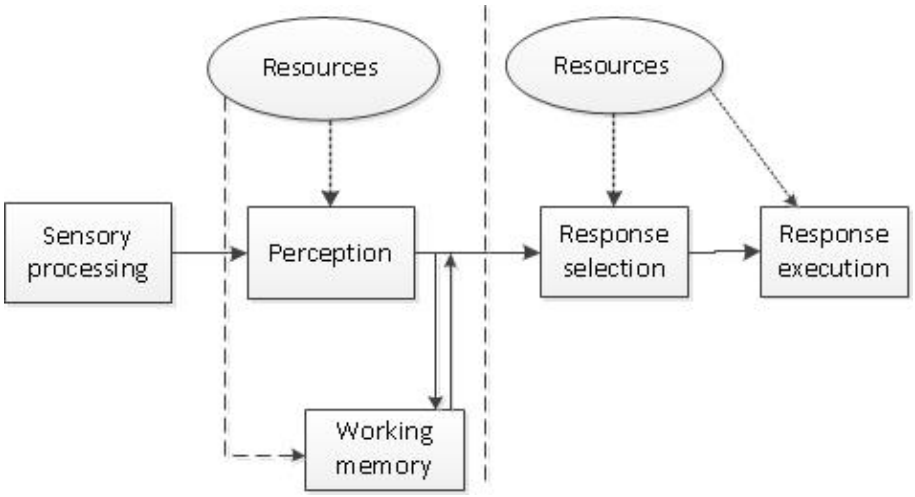


FIGURE 2.6: Wicken's model of working memory in context

modal resources are required by more than one task at any stage of the cognitive process. Conflicts do not exist within a single task, as processing is assumed to occur sequentially within a single task [49].

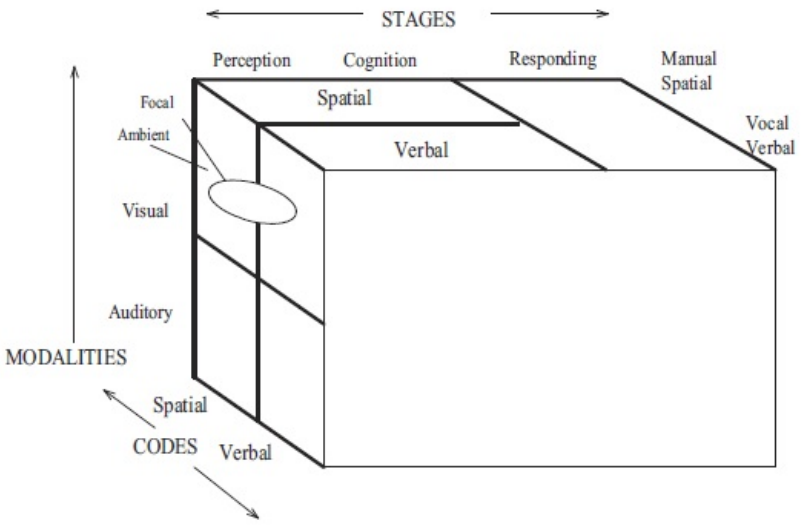


FIGURE 2.7: Wicken's 4D model

While the Multimodal Resource Theory can be useful in predicting whether modal interference will occur in a multiple-task situation and ultimately, an estimate of the

performance of the operator, there is less information articulated in the theory, or the model, about the processes involved in sharing these resources, e.g., modal sharing within a single task, where all processes are semantically dependent on each other. Also, the definition of what qualifies as a single task and how it is modally categorised into its perceptual, cognitive and response counterparts can be complicated and difficult to apply in individual situations, as shown by the multi-step process required for analysis in the computational model [49]. Interference and competition for modal resources across tasks can only tell half the story; the advantages of cross-modal task performance may be due to more than just separate physical perception channels. Conscious and automated time or other sharing of modal resources by various allocation strategies also need to be considered as possible alternative reasons for high performance in some types of tasks with multimodal processing.

Along the stages axis in this model, as can be seen in Fig. 2.7, the processing model is similar to the human processor model, even though it is not stated by Wickens. The human processor model [117] uses the cognitive, perceptual, and motor processors along with the visual image, working memory, and long term memory storages. It separates information processing into three subsystems vertically: perceptual subsystem, cognitive subsystem and motor subsystem, which correspond to perception, cognition and responding in Fig. 2.7. Each subsystem consists of different components horizontally to support its functionality. The main advantage of the human processor model is to calculate the cycle and decay time for each processor. The estimated cycle times are 50-200ms (Mean=100ms) for perceptual processor, 25-170ms (Mean=70ms) for cognitive processor and 30-100ms (Mean=70ms) for motor processor. The typical values of decay time for different components in subsystems are also estimated [117]. These values for human processor model have been estimated to be different for younger and older adult groups [118]. The value of the human processor model allows a system designer to predict the performance with respect to task time by users.

2.2.4 Cognitive Load Theory for MMIS

Cognitive load theory is another method used by researchers to identify multimodal processing for improving performance in tasks that induce high mental demand. Cognitive load theory attempts to interpret the experience of mental demand, adding an interesting dimension to performance assessment. The theory builds on the assumption that human beings' working memory has limited capacity and duration [50]. One can view working memory as the random access memory of a computer. It is capable of remembering information over a very brief interval (a few seconds), and it is what allows us to keep our mind on a task. Effective use of working memory processing is important for achieving high performance. Subjects presenting similar levels of performance may also differ in their individual cognitive load experience.

It was found that compared to single modality, subjects demonstrated not only a performance superiority but also preference for multiple modalities, if different modalities were used in tasks [51]. Cognitive load theory interprets this phenomenon by ascribing the advantages of multimodal processing to effective expanding of working memory by a set of modality-specific working memory resources when multi-modality is used [52]. Furthermore, evidence shows subjects seem to change and adjust their multimodal behaviour in complex, high-load tasks. For instance, when tasks become more difficult, users prefer to interact in multi-modalities rather than uni-modally across varieties of application domains. It is believed that the use of multi-modalities facilitates more effective use of modality-based working memory resources and gives users a hand in self-managing cognitive load [52, 119, 120]. Oviatt et al. [54, 55] investigate user-centered design principles and strategies for MMIS in educational applications and show that user-interface design that minimises cognitive load can free up mental resources and improve student performance.

Morsella and Krauss [121] found that participants use more gestures when describing visual objects from memory and when describing objects that were difficult to remember and express verbally. Participants especially use gestures when describing an object accessed visually. If gestures are restricted, they are likely to produce non-fluent speech even when spatial memory is untaxed. This may suggest that gestures

can directly affect both working memory and lexical retrieval. Speakers gesture when they talk and those gestures can have a positive effect on communication. Gesturing has been proven to be helpful in lightening a speaker's cognitive load in numerous experimental conditions [53].

Cook et al. [122] concluded that speakers often gesture with their hands while talking, even when they do not appear to have any difficulties with lexical access or fluency. Their findings suggest that gestural movements can function to lighten one's working memory load. Importantly, they also suggest that moving the hands in rhythmic synchrony with speech (like beats gesture) does not lighten the load on working memory. Speakers need to move their hands in meaningful ways in order to reduce the working memory load.

As a summary, our brains work multimodally to acquire and process information and our working memory also deals with multimodal input in a coordinated manner. In this way, human communication can be seen to exploit our natural ability to easily process and produce multimodal information. These theories can be directly applied to the design and implementation of multimodal interfaces by allowing users to make flexible use of the entire gamut of modal productions (e.g. gaze, gesture and speech). However, a more recent study [56] demonstrated that the interaction between co-speech gestures and working memory is affected by an individual's working memory load.

2.2.5 Gender Studies in Human Cognition

Regarding the individual differences in cognitive processing, controversial views exist. Some studies state that there are gender differences in verbal ability, quantitative ability and visual spatial ability in human cognition [123]. A general view is that men outperform women on visuospatial tasks and women outperformed men on tests of verbal fluency. Males are also demonstrated to show significantly higher mean scores on the arithmetical computations, arithmetical reasoning, and spatial cognition measures [124].

However, gender differences in human cognition are still controversial. For example, studies conducted by Hyde and her colleagues show that gender differences in cognitive

abilities (verbal, quantitative, visuospatial) are quite small and therefore, insignificant [125, 126].

Gender differences have been studied in many other aspects. A review of gender difference studies relevant to MMIS design will be detailed in Section 2.4.

2.3 Correlation of Speech and Hand Gestures

As presented in the previous section, the design of MMIS including speech and gestures input modes gains immense support from cognitive studies. In this section we will introduce the correlation of speech and hand gestures that is important for the design of MMIS.

2.3.1 Gesture Types

Before we review the correlation of speech and hand gestures, we will first present McNeill's taxonomy of gestures, and then discuss his view on the relationships of gesture and speech, as we will use his gesture classification in this work. McNeill's definition of gesture categories includes four types [40]:

Deictic gestures are the familiar pointing motions that identify an entity under discussion. A variant, abstract deictic gestures, are used to specialise and locate in physical space where entities under discussion have been placed.

Iconic gestures represent a concrete idea. They mostly convey information about the outline of a picture of shape or object in space or the hands represent the shape or the object itself. These gestures are imagistically representational. For example, a speaker describing a chair which has a square seat, may use two hands to draw a square in the air to represent the square seat as in Fig. 2.8.

Metaphoric gestures are also representational, but they are more associated with abstract ideas related to subjective notions, rather than the object itself. They can be viewed as de facto iconic gestures. A simple and frequently used metaphoric gestures might be a thumb up hand gesture in Fig. 2.9. When you praise someone for the good

Figure 2.8: Iconic hand gestures

job he has done, you may say "Excellent!" with a thumb up hand gesture. The thumb up hand gesture represents your notion to praise someone rather than any real object.

Figure 2.9: Metaphoric hand gestures

Beats, as the fourth type of gestures of McNeill's classification, is named batons by others. Beat gestures are small baton-like hand movements that serve to mark the speech pace normally. They are timed with the "rhythm" of speech. These gestures are not considered to convey any semantic information. Beats vary in size, and can be large, noticeable movements. Often, however, they are small, barely perceptible flicks of the wrist or finger.

2.3.2 Relationship between Gesture and Speech

Currently there are three different views about the relationship between speech and gestures followed by researchers. The first one points out that speech and gesture are separately communicated [43, 127, 128, 129]. According to this view, the primary role of gestures is to compensate for speech when verbal communication is temporarily unavailable (e.g. coughing or hard to express by words). They argued that the process

of gesture production has no effect on the process of speech production or the cognitive process related to speech.

The second point of view proposed initially by Krauss and fellows [130, 131] is that speech and gestures are linked reciprocally at a specific point during speech production. They pointed out that the production of gestures is activated when speakers come across difficulty in lexical retrieval. The activation of gestures in turn activates the lexical affiliate of that concept in mind, which results in articulating of the word successfully. According to this view, gesture is linked with speech only to the extent that it stimulates the activation of word retrieval in speech at a moment.

The third one articulated by McNeill [40] argues that speech and gestures form an integrated system of communication. The links between speech and gesture are presented at the different levels of speech production (e.g. discourse, syntax, semantics and prosody). From this standpoint, speech and gesture co-occur with one another during the same underlying thought process, even though the two modalities may capture and reflect different aspects of the common underlying cognitive process. The processes of the production of gesture and of speech should therefore influence each other at any disrupted point. He claimed in [132] that "gestures and speech are parts of the same psychological structure and share a computational stage." He gives five arguments as evidence, which are based on the very close temporal, semantic, pragmatic, pathological, and developmental parallels between speech and referential and discourse-oriented gestures.

First, gestures occur only during speech. Gestures as he defined above normally occur overwhelmingly in speech situations. Gestures by listeners may also occur, but they are extremely rare. More specifically, the majority of gestures (90% based on McNeill's count) occur during the speaker's actual articulation, and not, for example, during pauses.

Second, gestures have semantic and pragmatic functions that parallel those of speech. Gestures are symbols equivalent to various linguistic units in meaning and function as well. The form of the gesture is to some extent determined by the content being conveyed. All four types of gestures present content (or perform discourse

functions) related to their lexical affiliates. Often the content conveyed by gesture complements that of speech. For example, gesture may provide additional information with speech, such as a manner of action, or the physical relationship of two entities.

Third, speech and gestures are temporally synchronised within linguistic units. Gestures happen at the same time with their lexical counterparts as semantically and pragmatically parallel linguistic units. In fact, speakers seem to regulate their gestural time, by performing a hold before or after the stroke, to ensure this synchronization. Moreover, gestures "almost never cross clause boundaries" which ensure that they stay within their lexical counterparts propositional phrase.

Fourth, gestures and speech are affected in parallel ways by the neurological damage that produced aphasia. Broca's aphasics can speak "telegraphically" with command of content words. But they cannot relate these words into fluent sentences. They produce numerous iconic gestures (parallel to the lexical affiliates) but only few beats (parallel to the lack of higher-level discourse ability). In contrast, Wernicke's aphasics speak "vacuously" with fluent sentences but, however, little concrete semantics. They produce beat gestures and metaphoric gestures occasionally, but they produce few or even no iconic gestures.

Finally, gestures develop in parallel with speech in children. Speech abilities of children progress nearly from deictics and concrete words to discourse coding grammatically. Their gesture abilities also progress roughly from deictics and iconics to metaphorics and beats.

In McNeill's argument, gesture is communicative and provides meaning apart from that of speech, stems from a common source with speech, and is produced in an interactive , parallel fashion. This view laid the foundation for applying speech and gestures in MMIS. As speech and gestures are temporally synchronised, they can be integrated together sequentially. As gestures convey related content for speech either complementarily or redundantly, speech and gestures might compensate each other or debug each other in MMIS.

McNeill's view is more prevalent and widely accepted nowadays. There is actually already some neuropsychological and neurophysiologic evidence supporting the idea

that speech and gesture share the same communicating system [23]. We will not explore the deep or original relation between speech and gestures. We accept that speech and gestures are synchronised with each other. In this thesis, our study is grounded in McNeill's theoretical viewpoint that speech and gesture spring from a common origin. But one of the goals of this study is to indeed "throw useful light on McNeill's view" and provide quantitative evidence for the correlation of speech and gestures from cognitive and temporal aspects. Our main intention is to identify, if any, gender differences existing in using speech and gestures which can benefit the design of MMIS.

2.3.3 Temporal Synchrony of Speech and Gesture

Most findings regarding the temporal synchrony of speech and gesture suggest that gestures normally precede or fully synchronise with their lexical affiliates [39, 40, 41, 42, 43, 44, 45, 23, 46]. That is, to our best knowledge, no study has provided evidence that gestures are happening after their lexical affiliates. While measuring the temporal synchrony of speech and co-occurring gestures, different measurement points have been used for calculating the time interval between the manual and speech movements. For example, some measured gestural onset (the start point of a gesture) to speech onset (the start point of the lexical affiliates) [39, 23], while others have measured the interval between the apex of gestural stoke and the stressed point in related keywords [133].

The conclusion from [40] is that both gesture strokes and spoken utterances are performed together at more or less regular intervals. These intervals turned out to be between 1 and 2 seconds. Morrel-Samuels and Krauss [39] examined 60 carefully selected gestures and found that the time interval between the onset of gestures and the onset of their lexical affiliates is to range from 0 to 3.75s, with a mean of 0.99s and a median of 0.75s. None of the sixty gestures was initiated after articulation of the lexical affiliate. A more recent study [23] examined the temporal synchronization of meaningful speech and gestures versus pseudo-words and gestures in more controlled paradigm with Italian speakers. One of their findings is that gestural onset always preceded the lexical affiliate even for the pseudo-word and pseudo-gesture. The interval

between gesture onset and speech onset significantly differed from the same interval for pseudo-words and meaningful gestures (693ms), and for words and meaningless gestures (375ms). One multimodal system, called Quickset, combined speech and gesture inputs which are overlapped or fall within a certain lag relation temporally. They found that it is proper to integrate a gesture with speech which follows within 4s interval [134]. Intuitively, we believe that 4-s intervals are longer than normal time intervals in natural communication. Different measurement methods might be one explanation for the different conclusions when quantifying the time interval between speech and related gestures. Other potential factors that can affect the results could be different gesture types used in experiment, different tasks selected for experiment and user-related factors (e.g. user gender, age, cultural background and other individual differences).

However, the answer to which way is best to measure the time synchrony of speech and gesture is unclear given the available literature. In this thesis, our focus is not to examine if gestures precede or synchronise with the lexical affiliate in general, but to determine if there are any gender differences in the synchronization between gestures and speech. Nevertheless, the experimental methodologies in this thesis will lend insight into whether gestures always occur before or simultaneously with the lexical affiliate. The onset of gesture stroke will be temporally measured with the onset of lexical affiliate.

2.4 Gender Differences

There are two views regarding gender: the essentialist and the social constructionist [135]. Based on the essentialist view, gender is part of our genetic make-up and we were born with it. Men and women are therefore distinct identities and their behaviors are shaped accordingly. From a social constructionist point of view, gender is shaped by society, culture and time. They advocate the idea that psychological conditions in early life leads to who we are.

It is in vain to say if gender differences are definitely born with genes or only

formed in society. Some researchers studied the basic biological differences between the male and female brain and they demonstrated that the brain differences make it impossible for the sexes to present equal emotional or intellectual characteristics [136]. Others have also given evidence that sociological, cultural and religious factors can affect gender differences in communication [137]. The focus of this dissertation is on the accommodation of gender differences in MMIS design. This review on the gender differences relevant to MMIS design including speech and hand gestures first presents a survey on gender differences in word use and gestures, and then discusses "Gender HCI" which was established in 2004.

2.4.1 Gender Differences in Word Use

An empirical study by Mulac and Lundell [18] revealed that male speakers used more *impersonals, fillers, elliptical sentences, units, justifiers, geographical references, and spatial references* in a task of describing landscape photographs orally, while female speakers used more *intensive avderbs, personal pronouns, negations, verbs of cognition, dependent clauses with subordinating conjunctions understood, oppositions and pauses.* They applied these language variables to predict the gender of speakers with 87.5% accuracy. Some other studies have also reported that, for example, women have been found to use more conjunctions such as "but", and more modal auxiliary verbs for particular cases [19, 20, 21, 22]. Men have been found to use more longer words and use more references to location [138, 21].

Another finding reported in [21] about gender differences in word use is that females are more likely to use first-person singular than males. This finding was verified by Cohen [139]. In Cohen's study, on average, men tend to have shorter intervals between their use of pronouns, while women have longer narratives between two subsequent utterances of pronouns in personal narratives.

There are no studies addressing the particular gender differences in word use specifically in MMIS. However, if gender can be predicted by words they use [18] and they present significant differences in word use, we can assume that males and females will have their own preferences to use some words when they interact with MMIS including

speech and gestures. This implies that the design of MMIS will potentially achieve greater performance by accommodating gender differences.

2.4.2 Gender Differences in Gestures

Based on the available literature, there are no studies addressing gender differences on gestures. It was suggested in [37] that women more often use face expression and hand gestures to express their thoughts than men. Regarding nonverbal communication there are differences between females and males. Women use more expressions and nonverbal behaviors than men. Women are more skilled at sending and receiving nonverbal messages [140]. Men are louder and more interruptive and display more nervous, defluent behaviors. These studies refer to nonverbal communication in general which includes facial expressions, eye movements, head movements etc. They are not restricted to hand gestures. One report from Evergreen Valley College reveal that the differences in the mean use of hand gestures used by men and women was statistically significant in a social bar setting [141].

Freeman states that men are likely to use their hands to express themselves and they rely on more obvious gestures. Women, on the other hand, present more subtle gestures and they restrain and exhibit deferential gestures [38]. However, studies listed here about gender differences in gestures are set up in a social environment. To our best knowledge, no studies particularly illuminate whether any gender differences present in gesture use while people communicate with computer systems.

A recent finding [142] reveals that the number of gestures made with the right hand during speech is significantly higher for males, while during listening the number of gestures made with left hand is significantly higher. We have no results regarding the females left or right handed gestures. However some other studies state that their results did not reveal any difference in the degree of hand preference between pointing gestures produced along with speech and gestures produced on their own [143].

2.4.3 Gender HCI

The aim of MMIS is to allow user with strong preference to adopt new technologies for human computer interaction because of usability opportunities and flexibilities they provide. Our motivation to examine the gender differences for MMIS design is grounded on the gender differences found in HCI.

Findings from fields such as psychology, computer science, marketing, neuroscience, education, and economics strongly suggest that males and females solve problems, communicate, and process information differently. The term "Gender HCI" was coined in 2004 by Laura Beckwith, a PhD candidate at Oregon State University, and her supervisor Margaret Burnett [144]. Gender HCI is a subfield of HCI that focuses on the design and evaluation of interactive systems for humans, with an emphasis on differences in how males and females interact with computers. Even though this subfield was named in 2004, gender differences in HCI were found earlier than that.

A study in investigating gender differences regarding computer attitudes and perceived self-efficacy in the use of computers stated that there are gender differences in perceived self-efficacy regarding completion of complex tasks in both word processing and spreadsheet software [145]. Another study about computer displays argued that large displays helped to reduce the gap in gender differences in navigating virtual environments. With larger displays, females' performance with computers improved while males' was not significantly affected [146, 147]. It is a well known fact that boys and girls show different preferences in computer video games [148]. Investigation of tangible and proximity based HCI suggested that it is important to be cognisant of gender with respect to the interactions they facilitate [149].

2.4.4 Gender Prediction

As presented in previous sections, gender differences exist in many aspects of HCI, but gender recognition actually has not been broadly studied. This section introduces the works related to gender recognition by speech or hand gestures or both of them together.

Gender recognition from speech started from around the 1990s. Wu & Childers demonstrated that fundamental frequency and formant characteristics are reliable indicators for gender discrimination in early studies [30, 31]. But it seems gender prediction from speech did not boom after that preliminary study. One study suggested that different speech characteristics differentiate gender in different age ranges [150]. From this perspective, some researchers paid attention to age and gender recognition from speech patterns [34, 35, 151, 152, 153]. But the accuracies of these approaches varied in different data sets and were actually not very high. The modeling methods used in these articles are also complex, combining more than three different approaches.

Gender prediction based on hand gestures seems to be an unexplored area of research, since few studies have addressed the differences of hand gestures in gender. As for paired speech and hand gestures, to the best of our knowledge, the work in this thesis is novel in studying the gender differences in the alignment of speech and hand gestures. If gender can be predicted from speech and hand gestures used, adaptive processing strategies can be explored for the integration of multimodal modalities with better performance in the design of MMIS or gender HCI in future.

2.4.5 Gender Differences in Brain Activities

As mentioned previously, there are two views regarding gender differences: the essentialist and the social constructionist. Gender differences are related to biological factors. However this old theory is viewed as sexist by some since it has been used to subjugate women. The biological basis of gender differences in the brain has been recently investigated again in an increasing number of studies [154, 155]. These studies try to explain gender differences with differences in brain structure, chemistry and function of gender. These variations occur in different parts of the brain associated with language, memory, emotion, vision, hearing and navigation. Moir and Jessel [57] discuss the differences in male and female brain structure in relation to information processing.

A number of findings indicate that gender differences in language processing are related to functional asymmetry of the hemispheric brain. For example, McGlone [156]

states that the male brain may be more asymmetrically organised than the female brain, both for verbal and nonverbal functions. These differences are often significant in the mature organism. In children, they are more or less the same. A study [157], using echo-planar functional magnetic resonance imaging (fMRI) to investigate brain activation during phonological tasks, reports that brain activation in males is more lateralised to the left inferior frontal gyrus regions. In females, the patterns of activation are quite different and engage more diffused neural systems that involve both the left and right inferior frontal gyrus.

Fig. 2.10 (screenshot from [158]) displays images of the distribution of activated posterior language areas for three males and females [158]. This study states that females use the posterior temporal lobes more bilaterally during linguistic processing of global structures in a narrative than males do. An fMRI study [159] of gender differences in regional activation reveal that the activation is more left lateralised for the verbal and more right for the spatial tasks. While men show left activation for the spatial task, for women this is not the same. They also suggest that with the task difficulty increasing, more distributed activation is produced for the verbal tasks. And more circumscribed activation is produced for the spatial task. A more recent study in France found that males and females show different brain activation strength linked to word generation [58]. This study claim that there is a gender effect on cerebral activation. Men and women also show significant differences in mental rotation tasks, as reported in [160] that men show significantly stronger parietal activation, while women showed significantly greater right frontal activation.

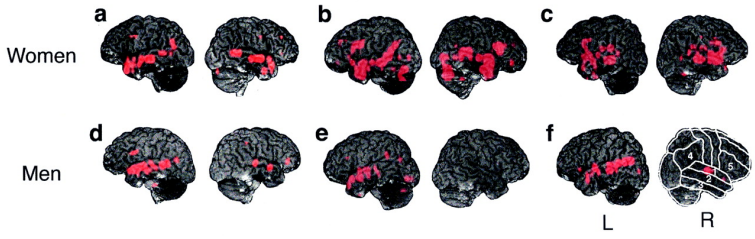


FIGURE 2.10: Images of the distribution of activated areas in posterior language areas

As a summary from these studies, the female brain is less lateralised with functions spread over both hemispheres of their brains and left-hemispheric dominance of language functions is greater in males than it is in females. However, there are also some controversial views about this conclusion. For example, Frost et al. states that no significant differences were found between the sexes in lateralization of activity in any region of interest or in intrahemispheric cortical activation patterns [161]. Weiss et al. demonstrate that men and women who do not differ significantly in verbal fluency task performance and show a very similar pattern of brain activation [162]. Sommer et al. state that the putative gender difference in language lateralization may be absent at the population level, or may be observed only with some, as yet not defined, language tasks [59].

Even though findings about the activated brain areas associated with language processing are controversial, it is still a great breakthrough in the study of gender differences in brain activities. However, it is still unknown if there are any gender differences in brain activities regarding non-verbal tasks specifically related with hand gestures. Some researchers have proven that speech and gesture share the same communication system [23]. We may assume that brain activities may also present different lateralization for males and females in speech and hand gesture coordination.

Most of the related studies listed above involve fMRI of the brain activities in various problem solving tasks. The experimental equipment of fMRI is expensive and requires a spacious professional environment. It is not easy for researchers to repeat these studies and to conduct their own experiments. In this thesis, we will use a commercially available and also affordable equipment-Emotiv Neuroheadset to study brain activities of males and females.

2.5 Conclusion

In this chapter, we reviewed the structures and the input information fusion methods for MMIS. The user input modes of MMIS evolved from conventional input interfaces (e.g. mouse and keyboard) to the user-related input (e.g., speech, pen, touch and

manual gestures), which provide more immersive user experience.

We also reviewed the cognitive processing theories related to multimodal processing. These theories reveal that human brain processes information multimodally and working memory deals with multimodal modes in a coordinated manner. User-centered interface design can free up mental resources and further improve user performance [54, 55]. The development of MMIS is upheld by cognitive theories at a number of levels. However, the gender differences in the processing of different types of input, particularly about speech and hand gestures, have not gained much attention.

To our best knowledge, there is few literature addressing gender differences in speech and hand gestures neither internally (the cognitive processing of these two modes) nor externally (the integration or presentation). After a review of gender differences in HCI, we believe that there are potential benefits to accommodate gender difference in the design of MMIS if gender can be predicted. We explored the internal and external gender differences in use of speech and hand gestures in the remaining chapters and introduced an attempt to build models to predict gender.

3

Experiment 1 and Analysis

3.1 Introduction

The purpose of this thesis is to explore gender differences in 1) speech and gestures used in same tasks, 2) cognitive processing, 3) brain activities while using speech and hand gestures.

For the first purpose, we will study speakers' preferences in using speech or gestures and check whether there are any gender differences across the presentation of speech and gestures, specifically looking at temporal synchronization of gestures and their lexical affiliates. For the second purpose, we will study the gender differences in cognitive processing while using speech and gestures. We will analyse participants' cognitive actions from video/audio clips and use the cognitive coding scheme developed by Suwa et al. [1, 60] for cognitive analysis. We conducted an experiment to film participants

when they described two objects using speech and hand gestures together. The video clips were annotated and coded for speech and hand gestures analysis. In this chapter, we will introduce the analysis methods we used in the first experiment, the experimental procedures and the analysis results as well.

For the third purpose, we conducted another experiment to collect EEG signals by Emotiv Neuroheadset from participants when they used speech and hand gestures only. EEG is the recording of electrical activity along the scalp and can track the state of the brain. The methodology and experimental procedures of the second experiment will be introduced in the next chapter.

3.2 Experiment 1

3.2.1 Task and Data Collection

The aim of the first experiment is to study gender differences in using speech and hand gestures. In this experiment, the participants were required to describe two types of chairs (Fig. 3.1 and Fig. 3.2) before a camera (Task 1 and Task 2 will be used in the following chapters to represent them). The process describing the characteristics of an object and therefore is similar to a design session. They were asked to describe the depiction of the objects in detail as if they drew the objects on paper, but they were required to use their speech and hand gestures instead of pen and paper.

A camera with an embedded microphone was placed in front of the participants to record their speech and hand gestures for later analysis. The 3D images of the objects were placed on the desk in the scope of speaker's view. The camera was placed in such a way that the upper body of participants was clearly recorded and the gestural space was included in the scope of the camera to capture their gestures clearly. They were encouraged to use as many gestures as possible, as well as to describe the objects as naturally as possible, to serve our ultimate goal which is to make human computer interaction as natural as human to human communications.

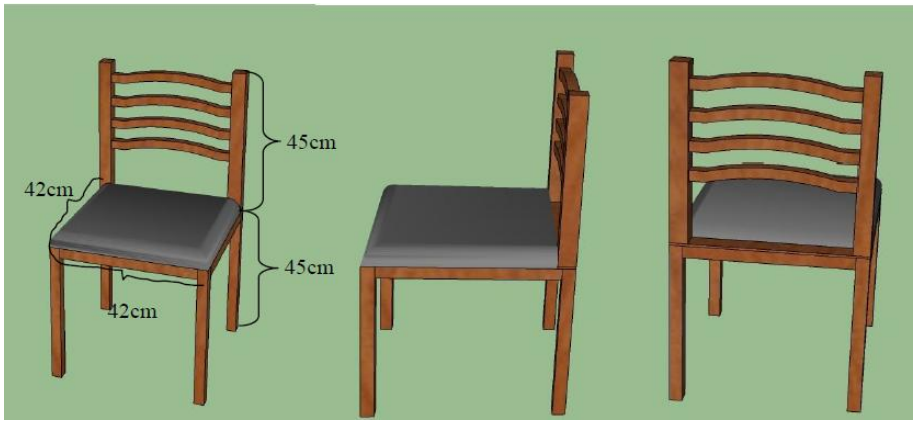


Figure 3.1: Task 1: simple chair description

Figure 3.2: Task 2: abstract chair descripton

3.2.2 Participants for Experiment 1

We obtained Human Ethic approval from Faculty Ethics Reviews Committees in Macquarie University and recruited participants according to it. English speakers without any history of disorders in language, speech, hearing or development were recruited in this study. They were not necessarily all native speakers, but spoke English fluently and had at least 6 months experience of living in Australia. Participants were recruited voluntarily via advertisements distributed at Macquarie University. They were not paid for their time.

Eighteen participants (9 males and 9 females) were filmed. Their ages varied from 20 to 50. None of them knew the exact task before they arrived in the experimental site

(VR lab at the Department of Computing, Macquarie Univerity). Participants were only told the topic of the study, the procedure involved and also the approximate time required for them to spend in the experimental room. If any individual showed any indication of unwillingness to join the experiment, then he or she was not scheduled for the completion of the experiment.

The data collection was not finished in one day. Filming was performed according to the availability of suitable participants over a month. Prior to the commencement of the experiment, each participant was asked to sign a consent form (see Appendix A). The chief experimental conductor went through each page of the information consent form with participants to make sure they totally understood any potential risk that might arise during the process. They were also asked to chat with the experimenter for a while to ensure they could speak English fluently and can be understood easily, as this would be important for the annotation process later.

3.3 Speech and Gesture Annotation

Before we study speech and hand gesture characteristics of males and females, they need to be extracted from video/audio clips. There are many speech and hand gesture recognition technologies nowadays. The aims of speech and gesture recognition are to identify what is said in speech and the movements of body by means of an algorithm implemented as a computer program.

Basically there exist three approaches to speech recognition: acoustic phonetic approach, pattern recognition approach and artificial intelligence approach [163]. Speech vocabularies are always needed for any speech recognition method. Computer programs can only recognise speech pre-defined in vocabularies. The accuracy rate of speech recognition can be over 90% for some vocabularies [164].

Gesture recognition can be conducted with techniques from computer vision and image processing. Some literature classify two different approaches in gesture recognition: a 3D model based and an appearance-based [165]. The 3D model approach uses volumetric information or skeletal representation, or a combination of the two to model

the movements of the body. The appearance-based approach derives the parameters directly from the images or videos using a template database rather than use a spatial representation of the body. Gesture recognition is more challenging than speech recognition regarding accuracy and complexity. At the moment hand gestures recognition are mostly implemented on only small vocabularies [104].

Even though speech and hand gestures can be recognised by various algorithms, our focus in this thesis is not to use these approaches to identify speech and hand gestures from video/audio clips. Our aim is to extract hand gestures and their corresponding keywords precisely for the post analysis to study their characteristics. We therefore use manual annotation in this thesis.

After McNeill's publication of his book "Hand and Mind: What Gestures Reveal about Thought" [40], the field of gesture studies has been broadly recognised by researchers. The definition of gesture classification and segmentation in his book became the foundation of subsequent studies. Many researchers in different area have adopted his theory to study gesture related fields, including multimodal interaction [166], gesture and speech relationship [133, 133], gesture modeling [167, 168, 169], gesture and cognition [170] etc. In these studies, manual annotation of gestures and speech are applied to exactly get the relevant information. These foregoing research has proven that manual annotation is a reliable method which we will adopt in this thesis. The pre-defined gesture coding schemes in a popular annotation software Anvil are also based on McNeill's gesture classification [171]. Anvil as one of the main annotation tools in the thesis will be introduced in the next section.

3.3.1 Annotation Tools

In order to explore the differences in speech and hand gestures used by male and female speakers, audio and video clips are annotated to extract hand gestures and their related lexical affiliates. During our experiments, speech of participants was recorded by the camera embedded microphone. We extracted speech from video clips for

each participant by a free software AVS Video Converter [1] before starting annotation. Gestures and speech are annotated in Anvil [172] and Praat [173] respectively. The speech annotations were imported into Anvil to synchronise gestures with their lexical affiliates.

Anvil is a free but powerful video annotation tool. It was originally developed for gesture research in 2000[171]. It offers multi-layered annotation based on a user-defined coding scheme. Anvil has been used by many researchers for video annotation in different areas [172, 174, 175, 176].

Praat is a very flexible and also free tool to do speech analysis [173]. The Praat program is maintained by Paul Boersma and David Weenink of the Institute of Phonetics Sciences of the University of Amsterdam. It offers a wide range of multiple operations (e.g. acoustic editing, acoustic measurements and creating pictures etc) and is also embedded with a scripting function. The most important function is that the Praat annotation file can be imported into Anvil panel, which allow us to do the post analysis of speech and hand gestures together.

3.3.2 Hand Gesture Annotation

Gesture annotation was done through Anvil. Fig. 3.3 displays a screenshot of gesture annotation in Anvil. All user related information is hidden in all figures from Praat and Anvil to comply with the privacy requirements of human ethics committee. The following two paragraphs introducing the Anvil user interface are quoted from [171]. Anvil user interface has four components: the Main Window, the Video Window, the Element Window, and the Annotation Board.

The Main window is located on the upper left which is also the first window seen after starting Anvil. It holds the main menu bar and, underneath, a tool bar that provides short-cuts to Anvil's most vital functions. In the middle part of this window, user actions and video information are listed. Near the bottom, the specification file used for the currently annotated video is displayed. Located at the very bottom are the

[1] http://www.avs4you.com/AVS-Video-Converter.aspx

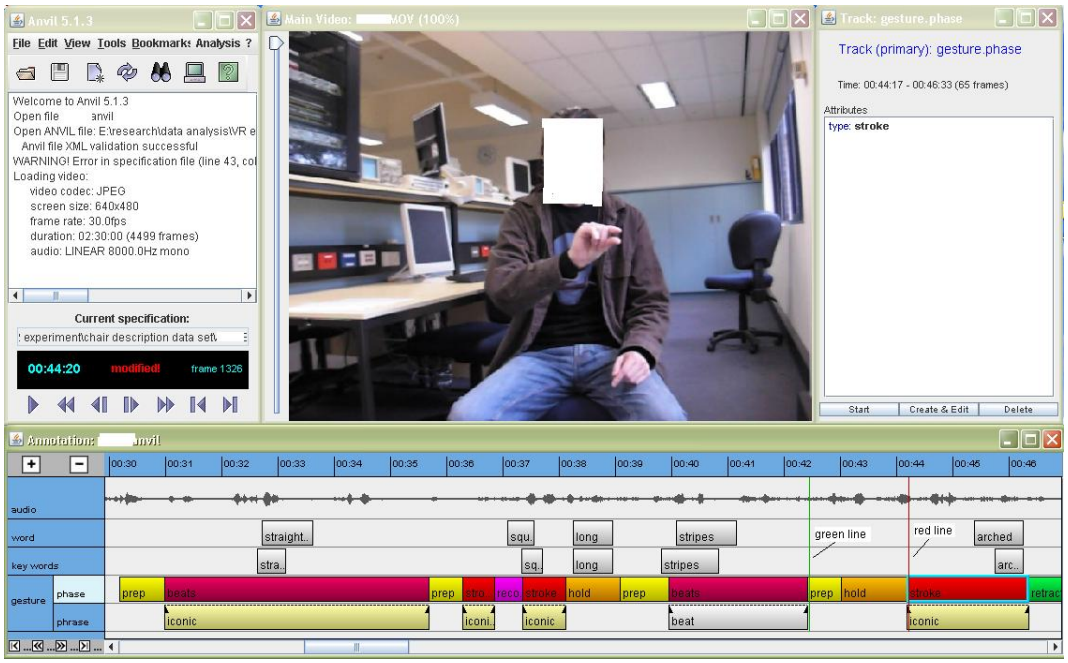


Figure 3.3: Sample view of gesture annotation in Anvil

video controls with variable playback speed and single-frame movement. The Video Window in the middle of the upper part displays the video loaded currently so that the user can watch the video and do annotations at the same time. The Element Window is on the left of upper part which gives information on the active tract and its currently selected element.

The most important window is the Annotation Board since all coding takes place here. It displays a time aligned view on all tracks and their contained elements which are user-defined annotation types in a specification file. The track hierarchy is seen on the left and the current track is highlighted. Time, represented by the horizontal x-axis, is marked in seconds on the top bar (small ticks represent video frames, 30 ticks between each full second). There is also a red vertical line, called the playback line, going across all tracks and marking the current frame in the video. The user can navigate through the video by dragging this line. Double-clicking a point in a track will bring a green line which can locate the starting frame of an interval of interest (e.g. stroke).

The gestures are annotated in two aspects: gesture type and gesture phase. We use McNeill's gesture type classification to categorise gestures into four types: deictic gesture, iconic gesture, metaphoric gesture and beat gesture. Gesture phases are segmented based on the Anvil built-in gesture phase description files. In the description file, a gesture is segmented into 7 phases which also postulated by McNeill [40]and then extended by Kita et al. [177]:

Prep preparation phase, bringing arm and hand into stroke position. This means the limb moves away from a rest position into the gesture space where it will begin the stroke.

Stroke the most energetic part of the gesture movement and also the requisite part of a gesture. A gesture is not said to happen with stroke phase absent. It is also the gesture phase with meaning and effort.

Beats a number of successive strokes (beats); all beats should be covered by this phase. This should be identified with the gesture type: beat gesture. If there is a hold in-between then continue with prep phase.

Hold a phase of stillness just before or just after the stroke, usually used to defer the stroke so that it coincides with a certain word. The hold can be a "post-stroke" hold or "pre-stroke" hold.

Recoil directly after the stroke the hand may spring back so as to emphasise the harshness of the stoke.

Retract movement back to rest position (not always the same position as at the start). In some situations, there may not be this phase if the speaker immediately moves into a new gesture.

Partial-retract retraction movement that is stopped midway to open another gesture phase.

For these seven phases, only the stroke phase is a compulsory part of a gesture while others are optional. Actually in most situations, it is rare to see all these seven

phases presented during a gestural process. The following figures display three phases for a gesture:

The speaker uses gestures to describe a "square" in this example. From the rest position in Fig. 3.4, the speaker moves her hands to the start position of the stroke phase in Fig. 3.5 and then gestures "square" in Fig. 3.6. At the last stage of this gesture, the speaker holds for almost 1 second (a screenshot is shown in Fig. 3.7) before she starts another gesture in Fig. 3.8 without a retract phase for "square". It is easy to identify a gesture if it starts from the rest position. However as can be seen in Fig. 3.7 and Fig. 3.8, some gestures just start from the end of the previous gesture which may not end with a retract phase. Actually during continuous speech, many gestures are finished without a preparation phase or any other phases except for a stroke phase. Fig. 3.9 illustrates the proportion for different gesture phases of one speaker. As seen in this example, 34 out of 55 gestures have prep phases and only 18 out of 55 gestures have retract phases. How to identify each phase of a gesture during continuous speech is a major problem.

Figure 3.4: Rest position of Prep phase

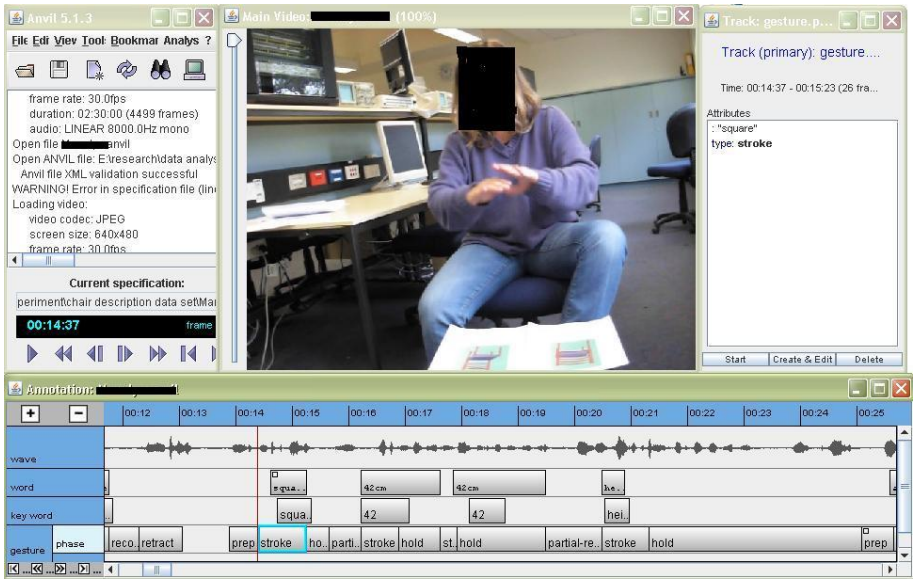


FIGURE 3.5: Start position of Stroke phase

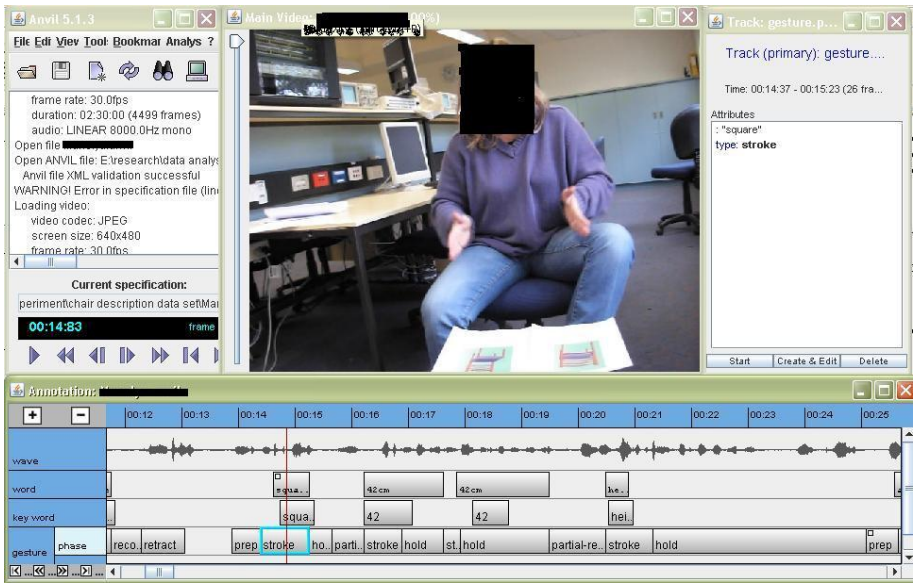


FIGURE 3.6: A point during Stroke phase

Kita et al. [177] give some instructions about how to identify phase types. As a *stroke* phase, it is exerted with more force than neighbouring phases. The acceleration

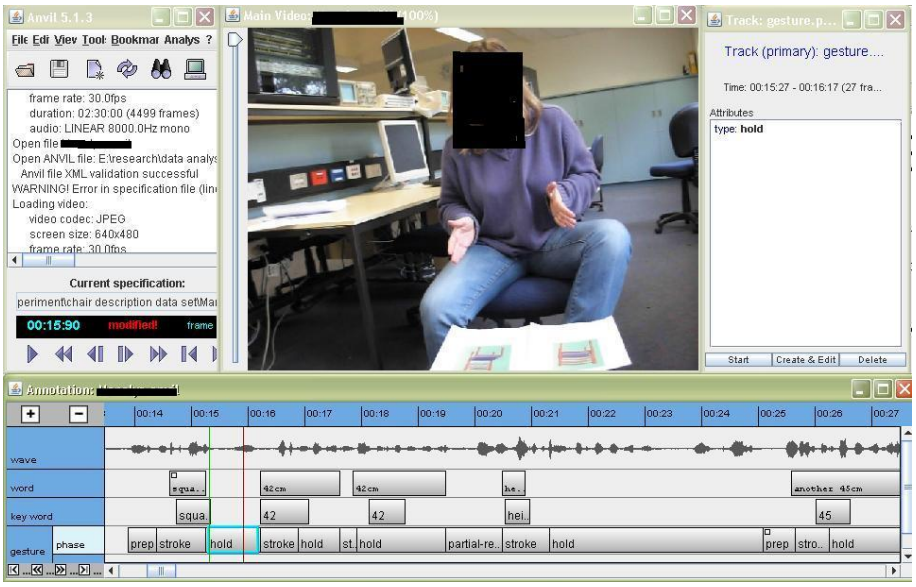


Figure 3.7: A point during Hold phase

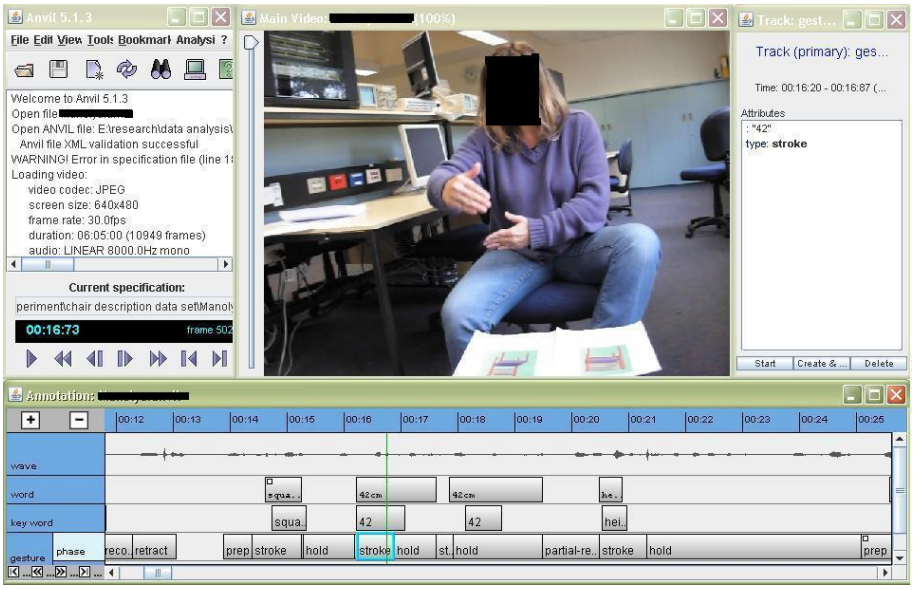


Figure 3.8: A point in Stroke phase for another gesture

and deceleration are good indicators of exerted force. The stroke phase is the most energetic part and requisite for a gesture movement. The movement for a gesture stroke is often apparent in the video frames as a blurring of the hands; the cessation of the blurring in one stroke movement can be taken as the end of a gesture stroke [178].

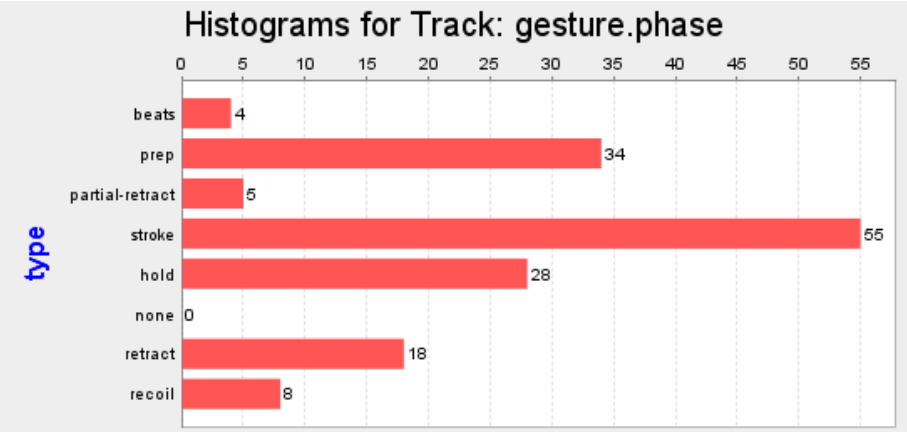


FIGURE 3.9: Example of proportions for different gesture phases for one participant

In a *hold* phase, the hand is motionless (or nearly motionless), since the hand is rarely perfectly motionless. The decision of hold phase is related to the neighbouring phases. A phase whose movement has no perceptible target direction is considered to be a hold phase. Sometimes a hold is performed with a distinctively 'active' hand shape at a position.

A *Preparation* phase is non-stroke movement that either departs from the resting position or moves a limb between two strokes. Beside the movement, the preparation phase also includes hand-interval preparation which can be the change of hand shape and the change of the orientation of palm and knuckles.

A non-stroke phase that arrives at the resting position is a *retraction*. Sometimes, the hand makes a non-stroke movement that goes toward a potential retraction phase, but shifts to a phase for another gesture before reaching the rest position. This movement is called a *partial retraction*.

With these instructions, the boundaries between each phase were discernible. There were inevitably occasional fuzzy boundaries, but these were not the norm and did not affect the post-annotation analysis.

While coding gestures, some studies did the annotation with sound muted [177, 179]. At the beginning, we hoped to code gestures without audio information to reduce subjective bias, but finally we found it is not possible to annotate a gesture without

sound in the McNeill style of annotation even it is possible to annotate speech without video, since the accompanying speech is crucial to interpreting and coding gestures [133].

3.3.3 Speech Annotation

It is possible to annotate speech without visual information. But our aim in speech annotation is to obtain lexical affiliates of gestures, so it is reasonable to code speech together with video tracks playing during the annotation. Speech annotation software is helpful for annotating speech accurately and saves countless hours for the coders.

Speech annotation in this thesis is actually completed in two steps.

Annotation in Anvil

In the first step, when we annotate gestures, we create a track named "word" under the "audio" wave track in Anvil to record the approximate location of gesture lexical affiliates. This means that the positions of a gesture related keyword (or keywords) in the audio stream are estimated in Anvil while coding gestures (see Fig. 3.3). Due to the limited resolution in Anvil, the exact position of a word can not be identified. But the word track in Anvil will provide a time window that include the keyword of a gesture. This time window will be used in speech annotation in Praat, which save plenty of coding time.

Annotation in Praat

In the second step, when we annotate speech in Praat, we have the video playing in Anvil to have references to gesture annotation.

In Praat, it is easy to identify word boundaries with the speech intensity contour displayed in annotation window. Fig. 3.10 displays a screenshot of speech annotation in Praat. In this figure the first tier on the top is the real speech signal waves. The second tier below the top one gives the speech intensity contour. The bottom tier is the annotation tier which displays the boundaries of extracted keywords. When doing

annotation in Praat, we get the time reference from Anvil annotation in the first step to find the approximate location of the keywords. This can be simply done by clicking anywhere in the first tier, the time will be displayed at the top this tier. The onset and offset of a word can be easily marked with the assistance of the intensity contour which always has a turn point at the boundary of two words. By double clicking the onset and offset points in the bottom tier, two lines will be generated to indicate the word boundaries. When the annotated section is selected, the time points of the boundaries and the duration of the keyword can be clearly seen on the top of the first tier as in Fig. 3.10.

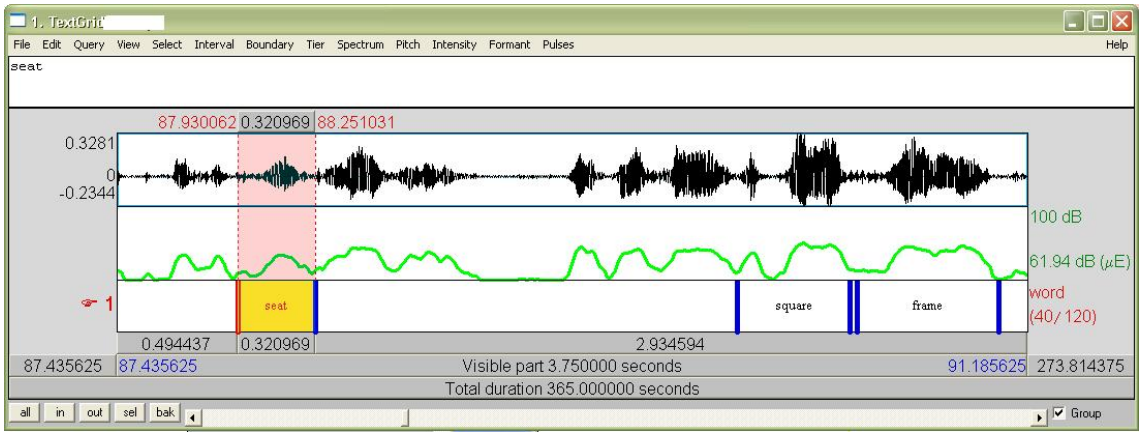


FIGURE 3.10: Sample view of speech annotation in Praat

In most situations during speech annotation, it is straightforward to determine which word corresponds to which gesture. However there were some fuzzy cases in our experiment. For example, the speaker made a gesture as illustrated in Fig. 3.6, while saying "the seat of the chair is square". In this situation, it was undoubted that the keyword "square" is affiliated with the gesture stroke. We, therefore, synchronised this gesture with its lexical affiliate "square". Another participant, when describing the same part of this chair, however, said, "the chair has a square seat" as shown in Fig. 3.11.

In this case, it was ambiguous to determine if this gesture was related to "square"

Figure 3.11: Another gesture stroke made for similar description with Fig. 3.6

or "seat". The best way to solve such ambiguity is to confirm with the participant who made the gestures, but it is impossible to check with every participant, since not all of them are available when we annotate their clips. In our case, we eventually checked with three participants who used this kind of description and found that they all agreed that the first word should be affiliated with the corresponding gesture in this situation. We then used this criterion to make decisions for all annotation involving this kind of ambiguity. Fortunately, this situation did not happen frequently.

After the completion of speech annotation for one participant, the annotation file from Praat was exported to an Anvil track named "keywords" under the "word" track and above the "gesture.phase" track (as shown in Fig. 3.3). In Fig. 3.12 we can see the obvious time bias of each keyword for the "word" and "keywords" tracks.

3.3.4 Post Annotation Analysis

In Anvil, each track can be exported to a table which can include start time, end time and duration for each annotated section in this track. For the gesture track,

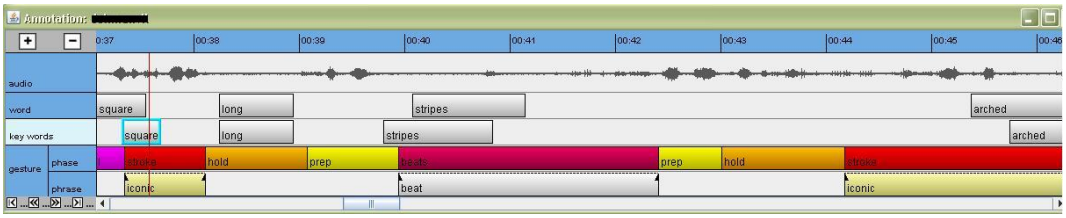


FIGURE 3.12: Match for keywords estimated in Anvil and keywords coded in Praat

the exported table has 3 columns corresponding to start time, end time and duration for each gesture. For the keywords track, the exported table also has the same 3 columns corresponding to start time, end time and duration for each keyword. For each participant, we imported the two tables to one Excel file which was used for post annotation processing.

Statistical analysis was applied in post annotation processing to explore if there is any significant differences for male and female speakers in the following aspects:

- Time interval between the onset of gestures and their corresponding lexical affiliates
- Time length of gesture strokes
- Time length of keywords
- Preferences in using speech and hand gestures

The analysis results will be discussed in the Section 3.5.

3.3.5 Pilot Annotation and Analysis

As mentioned before, the experiment was not completed in one day. We normally annotated the clips for one participant as soon as they finished the experiment. We annotated the clips for three participants at the initial stage to get familiar with the analysis software and also the annotation process. The pilot annotation is quite important for the later coding and analysis. The first three participants included two females and one male.

The first three participants produced about 15 minutes of videotapes in total. A total of 24986 frames (the frame rate is 30.0 frames per second (fps)) were analysed and 121 gestures were annotated eventually. For the initial analysis, we treated the start point of a gesture as the onset of the gesture no matter which phase the gesture starts with. As a whole, we noticed that more than 90% gestures started before the related words within 2 seconds in our experiments. These initial findings confirmed that gestures start before the onset of the related speech and are synchronised with speech as we reported before [180].

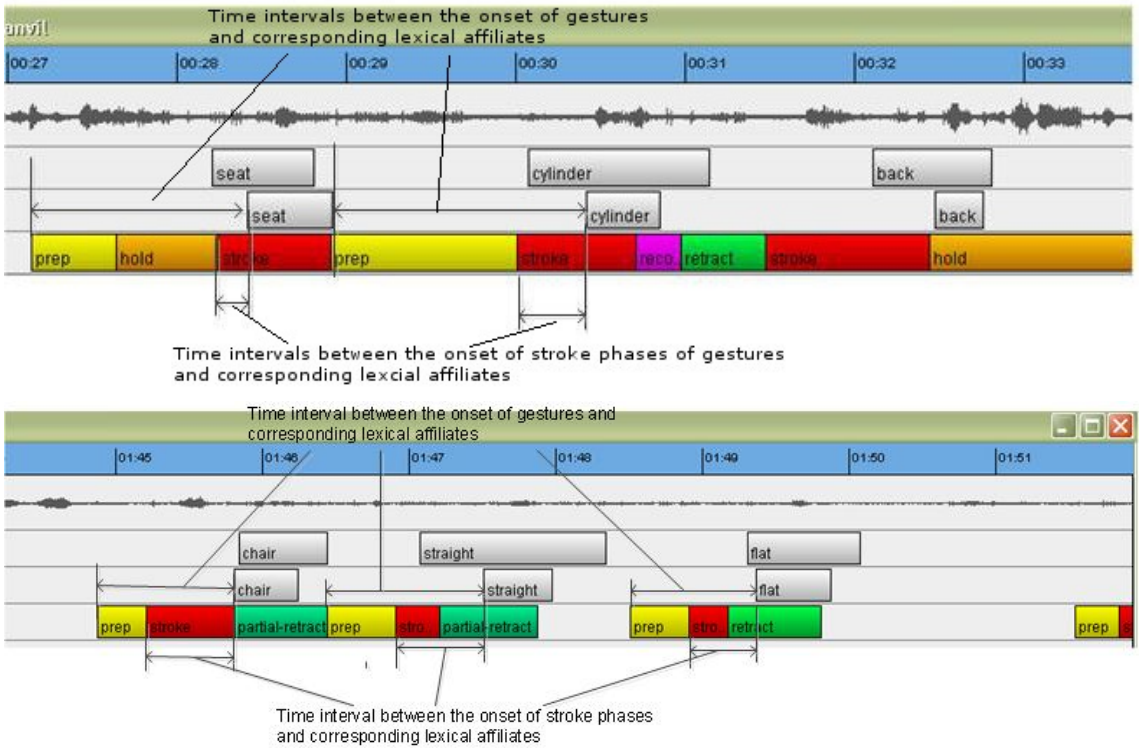


Figure 3.13: The onset time interval comparison

However, afterwards we found that the onsets of gesture lexical affiliates were closer to the onset of stroke phases than to the start points as explained in [180]. It can be clearly seen in Fig. 3.13 which are screenshots of two annotations for two different participants. The analysis of the time interval between stroke phases of gestures and their lexical affiliates based on the concrete numbers also verified our first observation.

In Table 3.1 and Table 3.2, '-1∼0s' indicates that the onset of speech precedes the co-verbal gestures (gesture strokes), while '0∼1s', '1∼2s' and '2∼3s' indicate that the onset of gestures (gesture strokes) precede the related words and the time intervals between them are within 1 seconds, 2 seconds and 3 seconds respectively. We can see from the two tables that, on average, the percentage of time intervals within 1∼2s and 2∼3s are less taking the start point of the stroke phases as the onset (11.02% and 2.20%) than taking the start point of whole gesture as the onset (20.66% and 4.96%). For each participant, the percentage also changes in the same way. There is a significant increase in the percentage of time intervals located between -1∼0s in Table 3.2 (from 2.48% to 11.79%).

TABLE 3.1: Time intervals between the onset of whole gesture and lexical affiliate

Participants	-1∼0s	0∼1s	1∼2s	2∼3s
P1	1.88%	64.15%	26.42%	7.55%
P2	4.54%	81.82%	13.64%	0
P3	0	70.83%	20.83%	8.33%
Average	2.48%	71.9%	20.66%	4.96%

TABLE 3.2: Time intervals between the onset of stroke phase and lexical affiliate

Participants	-1∼0s	0∼1s	1∼2s	2∼3s
P1	8.18%	76.66%	11.86%	3.30%
P2	16.91%	78.11%	4.08%	0
P3	10.38	68.69%	17.13%	3.8%
Average	11.79%	74.99%	11.02%	2.20%

Inspired by these initial results, we used the stroke phase of a gesture to represent

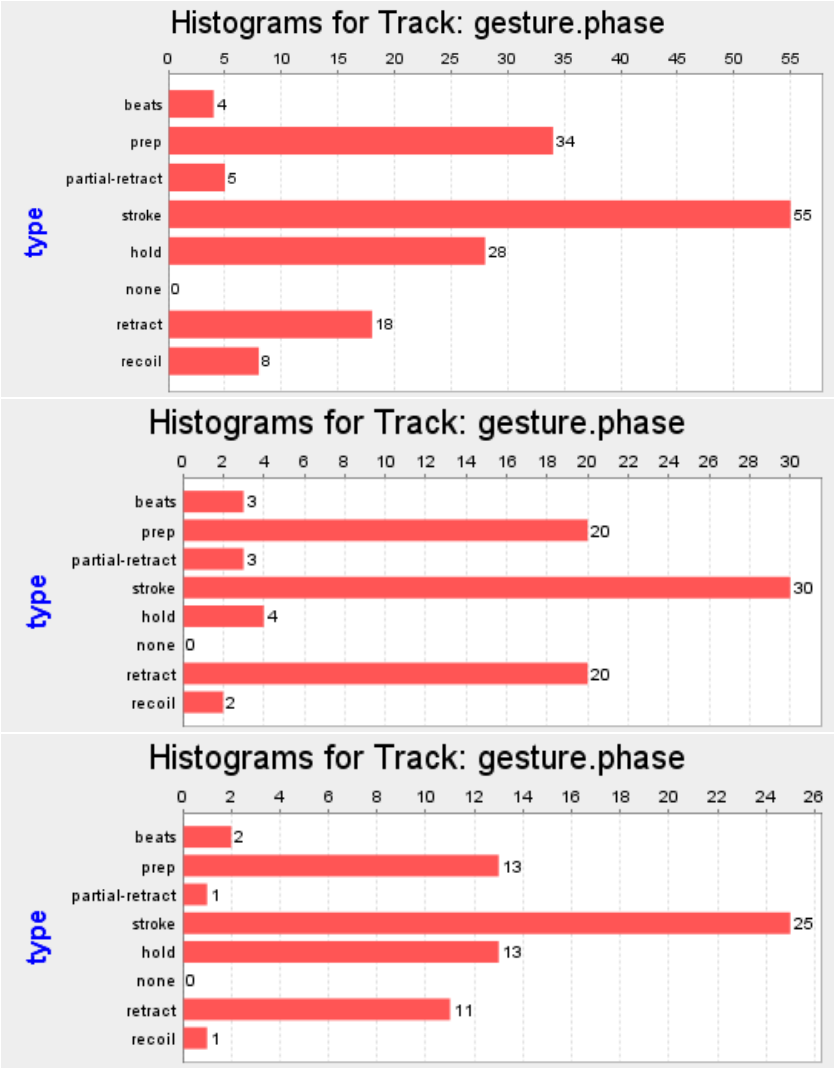


Figure 3.14: Histograms of gesture phases in pilot analysis for three participants

the whole gesture in the following annotation and analysis sections for other participants. Also as indicated in previous chapter, only the stroke phase is a compulsory phase of a gestural movement and many gestures are finished without preparation phase or some other phases. The histograms of different gesture phases for these participants are shown in Fig. 3.14. We can see in this figure that the majority of gestures produced by the three participants do not have other phases except stroke phases. This is another reason that we chose the stroke phase of a gesture to represent it.

This way, we achieved to save plenty of time, since the full annotation process was

extremely time-consuming. It also made the annotation more explicit, since the stroke phase of a gesture is the most energetic part of the gesture movement and also the compulsory part of a gesture. For the annotation, the movement for a gesture stroke was often apparent in the video frames as a blurring of the hands; the cessation of the blurring in one stroke movement was taken as the end of a gesture. Other phases were not recorded since the beginnings of other phases for each gesture were subject to greater subjectivity and difficulty in identification.

Inter-coder Agreement

Before we went to the analysis step, one important thing for any annotation project is to validate the performed annotations by measuring the agreement between different coders, since manual annotation mainly relies on the human coder's comprehension of segmenting and classifying the data. It is essential to evaluate how objective the annotations are.

The validation normally can be done by measuring inter-coder (multiple coders annotate the same media) or intra-coder (the same coder annotates the same media after some time has passed) agreement. In both cases, the degree of correspondence between two annotation files has to be measured. Since the annotation process is significantly time-consuming, it was difficult to ask someone to do the whole annotation as we did for all recorded clips. However another coder provided inter-coder agreement for one participant's clip. We used Coder 1 to represent him in this thesis.

Coder 1 was also a PhD student in another university in Sydney. He had no background about gesture and speech annotation before his involvement in our experiment. He learned classification and definitions of gestural phases. We trained him about how to use Anvil and Praat. After a couple of weeks' of training, he was quite confident about how to annotate gestures and speech in Anvil and Praat. What he coded included: the stroke phase for each gesture movement, the lexical affiliate of each gesture represented by stroke phase and marking of each lexical affiliate in Praat. We did the coding agreement test on gesture segmentation, lexical affiliate segmentation and lexical affiliate categories for one participant's clip.

Anvil includes the coding reliability measure Cohen's kappa for quantifying the level of agreement. This statistic is appropriate for testing whether agreement exceeds chance levels for binary and nominal ratings. Anvil offers to compute Cohen's kappa measure for both the degree of agreement in segmentation and classification. We only focused on the segmentation agreement in our test, since Code 1 only annotated the gesture strokes and there was no classification issue. The formula is (Pa - Pe) / (1 - Pe) where Pa is the relative observed agreement among coders, and Pe is the hypothetical probability of chance agreement. Anvil also reports corrected kappa according to Brennen/Prediger [181] where the "chance" term Pe in kappa is replaced by 1/n, where n is the number of categories.

It is likely that two coders have different segmentation numbers for the same annotation task. The challenge in this case is to decide which elements to compare in cases where the segmentation is different. In Anvil, this problem is solved by considering time slices instead of elements. Anvil cuts the annotation file into slices of 0.04 sec and compares categories on each time slice, adding one additional category VOID for the case that no annotation resides on the slice. These counts are put into a confusion matrix used to compute kappa. One can focus on segmentation only using the same method. For this, Anvil uses only two categories, VOID and ANNOTATED, and then performs the same computation as described above, resulting in a segmentation kappa [182].

The resulting kappa values always lie between 0 and 1, where 1 indicates perfect agreement. A kappa between 0.40 and 0.60 is considered as fair, between 0.60 and 0.75 as good and over 0.75 as excellent. Kappa statistics should not be viewed as the unequivocal standard for computing agreement. However they are almost always preferable to simple proportion (percentage) of agreement which does not exclude for chance agreement.

Fig. 3.15 displays the screenshot from Anvil about inter-coding agreement computation for gesture stoke annotation. From this figure we can see that Coder 1 got 58 gesture strokes (track elements in the figure) while Coder 2 had only 52 gesture strokes annotated. There were only two categories (one is for stroke annotation and one is for

parts without annotation), since we only focused on segmentation. The percentage of inter-coder agreement for gesture stroke segmentation was quite high - 90.17%. The Cohen's kappa is about 0.65 which is not excellent but is still fairly high.

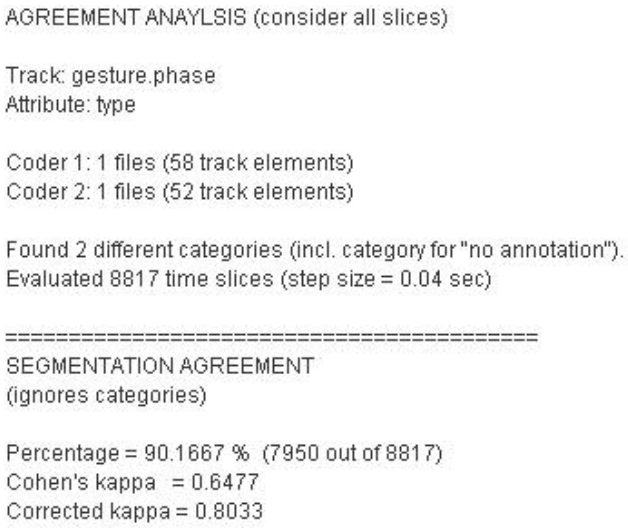


FIGURE 3.15: Inter-coder agreement for gesture segmentation

Intra-coder Agreement

Fig. 3.16 displays the screenshot about intra-coding agreement computation for gesture stoke annotation. For the two annotation files used for intra-coding agreement computation, the second version was generated about two months later than the first version. We can see from Fig. 3.16 that the same coder obtained same gesture stoke numbers (both 58 gestures). Two months later, the percentage of intra-coder agreement was as high as 90.22% and the Cohen's kappa is about 0.67. They were both about the same as the values of inter-coding agreement.

Category Agreement of Lexical Affiliates

The same strategy was used for calculating agreement for lexical affiliates of gestures. We included category agreement for lexical affiliates as well. The degree of coders'

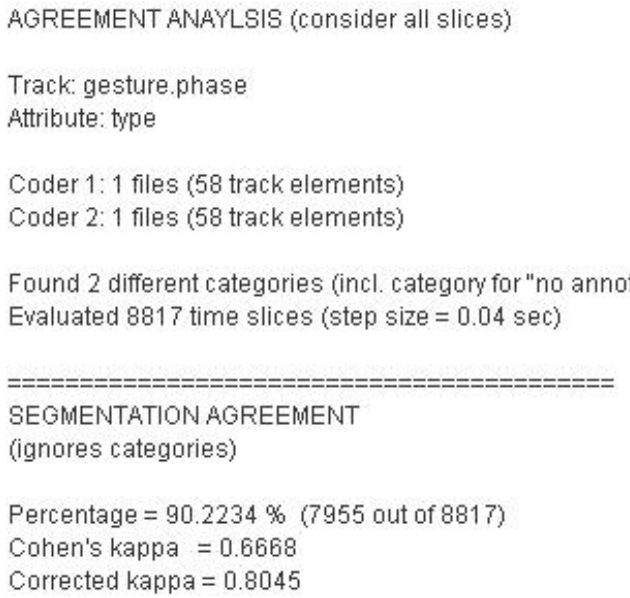


Figure 3.16: Intra-coder agreement for gesture segmentation

agreement on the meaning of the gestures (represented by the categories in coding agreement calculation) will affect the degree of agreement on segmentation of lexical affiliates.

From Fig. 3.17 it is quite clear that the agreement on category of lexical affiliates is perfect (with percentage agreement 100% and Cohen's kappa 1). The segmentation agreement was also excellent (with percentage agreement 99.09% and Cohen's kappa 0.94). As can be seen in this figure, two different coders annotated different numbers of lexical affiliates (58 and 52 respectively), but only 44 different categories were used to calculate category agreement. It was because that some words were used repeatedly by the speaker in the annotated clip. In order to make sure if it was a factor that may affect the degree of category agreement, we named each repeatedly used word with different name only for test purposes. We still obtained perfect category agreement after the simple process.

Fig. 3.18 also indicates that intra-coding agreement was perfect on segmentation and category as well. Percentage agreement was 100% and Cohen's kappa was 1 for

```
AGREEMENT ANAYLSIS (consider all slices)

Track: key word
Attribute: key words

Coder 1: 1 files (58 track elements)
Coder 2: 1 files (52 track elements)

Found 44 different categories (incl. category for "no annotation").
Evaluated 8817 time slices (step size = 0.04 sec)

========================================
SEGMENTATION AGREEMENT
(ignores categories)

Percentage = 99.0927 %  (8737 out of 8817)
Cohen's kappa   = 0.9441
Corrected kappa = 0.9819
========================================
CATEGORY AGREEMENT
(considers 746 slices that both coders annotated)

Percentage = 100 %  (746 out of 746)
Cohen's kappa   = 1
Corrected kappa = 1
```

FIGURE 3.17: Inter-coder agreement for lexical affiliates of gestures

both segmentation and category.

In summary (see Table 3.3), the degree of intra-coding agreement is slightly higher than inter-coding agreement. The agreement on lexical affiliates is higher than on gesture strokes. It may be simply because with the help of the speech intensity curve annotation of lexical affiliates is more direct-viewing than gesture annotation which may be subjective sometimes. On the whole, however, the degree of agreement on gesture annotation and lexical affiliates were both fairly high and acceptable for inter-coders and intra-coders. Even though we did not do the coding agreement test for all the annotated clips, the satisfactory result from the pilot annotation and analysis made us confident for the following annotation and analysis.

```
AGREEMENT ANAYLSIS (consider all slices)

Track: key word
Attribute: key words

Coder 1: 1 files (58 track elements)
Coder 2: 1 files (58 track elements)

Found 44 different categories (incl. category for "no annotation").
Evaluated 8817 time slices (step size = 0.04 sec)

=========================================
SEGMENTATION AGREEMENT
(ignores categories)

Percentage = 100 %  (8817 out of 8817)
Cohen's kappa  = 1
Corrected kappa = 1
=========================================
CATEGORY AGREEMENT
(considers 826 slices that both coders annotated)

Percentage = 100 %  (826 out of 826)
Cohen's kappa  = 1
Corrected kappa = 1
```

Figure 3.18: Intra-coder agreement for lexical affiliates of gestures

Table 3.3: Summary of coder agreement

Agreement	Gesture segmentation	Lexical affiliates
Inter-coder	90.17%	100%
Intra-coder	90.22%	100%

3.4 Cognitive Analysis and Coding

3.4.1 Protocol Analysis

The goal of cognitive analysis is to reveal the cognitive content, structures, and processes in subjects' minds during problem solving. A promising approach to cognitive

analysis is protocol analysis. Protocol analysis is a psychological research method that elicits verbal reports from research participants and analyses the verbal reports to reconstruct what happens in the mind of the participant. The foremost protocols are concurrent verbal reports (think-aloud protocol and talk-aloud protocol), and were developed by Ericsson and Simon based on the claim that these two forms of verbal reports can be the closest reflection of the cognitive processing [183].

Basically in an experiment using think-aloud protocol, participants think aloud while they are performing a set of specified tasks. They are required to say whatever they are attending to, thinking, doing, and feeling as they go through their task, which enables observers to obtain a completed first-hand process of the task besides the final product. Observers at such an experiment are requested to record everything participants say, without interpreting their actions and words. The experimental sessions are normally audio and video recorded so that observers can go back to see what participants said and how they reacted.

A talk-aloud protocol is slightly different from a think-aloud protocol in data-gathering. Think-aloud involves participants only describing what they are doing but not giving explanations. This approach is considered to be more objective compared to think-aloud method since participants merely report how they go about completing a task rather than interpreting or justifying their actions.

Ericsson and Simon claim that *cognitive processes are not modified by these verbal reports, and that task-directed cognitive processes determine what information is heeded and verbalised* [183].

Another type of verbal report defined by Erisson and Simon is the retrospective report. The memory trace of the information heeded successively while completing a task can be assessed from short term memory, at least in part, or retrieved from long term memory and verbalised just after the task is finished. But retrospective reports based on information in long term memory may display the incompleteness that is similar to experimental research on memory. They claim that retrospective reports are also direct verbalizations of specific cognitive processes. Erisson and Simon's protocol analysis has now been broadly used to examine cognitive processes under different

circumstances [184, 185, 1, 186, 187, 188, 189].

Based on the two types of verbal reports in protocol analysis, there are two approaches in protocol data collection: concurrent and retrospective. Retrospective protocols involve interviews with the participant after the problem solving process. Concurrent protocols are generated when the participant verbalises their thoughts while working on a specific task. However, both concurrent and retrospective protocols have been reported to lead to consistent understandings of the problem solving process [190].

In our experiments, the task is to describe two chairs with different structures. We believe that speech and gestures in this process significantly reflect cognitive processing in the participant's mind. We can treat both as the combined concurrent protocols of the object description tasks.

A practical guide is given in [191] that the complete procedures of coding and analyzing verbal data consists of the following eight functional steps, but not all of them are used in some cases:

1. Reducing or sampling the protocols.
2. Segmenting the reduced or sampled protocols.
3. Developing or choosing a coding scheme or formalism.
4. Operationalizing evidence in the coded protocols that constitutes a mapping to some chosen formalism.
5. Depicting the mapped formalism.
6. Seeking pattern(s) in the mapped formalism.
7. Interpreting the pattern(s).
8. Repeating the whole process, perhaps coding at a different grain size.

After the completion of experiments to get the protocols, we used all protocols for the following analysis. After step 2 in the above list, the procedures can be summarised

as segmentation, coding and analysis to identify cognitive processing patterns in our case.

Once the corpus of protocols to be coded is decided, segmentation is conducted to identify the unit of analysis. Normally a change in the participant's intention, or the contents of their thoughts, indicate a new segment. There are two methods to segment data based on two views of the design process: rational problem solving approach and constructivist approach; namely process-oriented [192] and content-oriented [185] methods.

The process-oriented segmentation method focuses on describing the design process in terms of a sequence of problem solving activities, for example, problem recognition, goal setting, solution proposing, solution analysing, or top down vs. bottom up strategies. The rule in this approach for segmentation is to segment protocols based on separate verbalization (e.g pauses, intensity, intonations and syntactic markers).

The target of content-oriented segmentation method is to reveal the content of what participants see, attend to, think of during the task completion. In this approach, protocols are divided by the participant's intention. A change in the participant's intention or the contents of their thoughts or their actions may indicate the start of a new segment [1]. In this case, a single segment can include one sentence or many.

We adopted the coding scheme developed by Suwa et al. [1, 60] to our purpose in this thesis. They used the content-oriented segmentation method in their coding scheme. We thus identified a new segment by the way that there is a change in the speakers' intention or the contents of their thoughts. For example, a participant may have said, 'The seat of this chair is square... and then for the leg part ...' The speaker changes his/her attention from the seat part to the leg part. So we got the start point of a new segment at 'and then ...'. Consequently, a single segment can include one sentence or many. The coding scheme will be introduced in the following section. An example of how we coded the cognitive actions will also be given in the Section 3.4.4.

3.4.2 Suwa's Coding Scheme

Suwa et al. [1, 60] code designers' cognitive actions using a special coding scheme. The scheme identifies various types of cognitive actions and reveals the structure of cognitive actions in the design process. Suwa's code has been applied to different areas in other design domains by many researchers [193, 194]. Other studies using Suwa's code provided us with reliable references, on the basis of which we selected it as our protocol analysis tool.

Visual communication is a process of sending and receiving messages using images. In this thesis, we are primarily interested in visual communication as a process. Visuals are a system of representation and signification that allow us to produce and communicate thoughts and images about reality [195]. We designed the experiments to investigate the differences in visual and oral communication between males and females. The key to this investigation is the meaning attached to the visuals. Visual communication is made up of presentational symbols whose meaning results from their existence in particular contexts. Meaning is formed by seeing and thinking. Therefore, meaning is highly associated with human cognition. The conventions of visual communication are a combination of universal and culturally based conventions. Visual literacy can be defined as the "ability to construct meaning from visual images" [196].

Being visually literate is a combination of syntax and semantics [197]. Syntax is the form or building blocks of an image. The syntax of an image can be regarded as the pictorial structure and organisation. Semantics refers to the way images relate more broadly to issues in the world to gain meaning. The word 'semantic' has a similar origin to the word 'sign'. Semantics are often closely related to Semiotics. Semiotics is the study of signs. In practice, visual semantics refers to the ways images fit into the cultural process of communication. This includes the relationship between form and meaning. Semantics might include looking at the way meaning is created through:

- form and structure
- culturally constructed ideas that shape the interpretation of icons, symbols and representations

- a social interaction with the images

Suwa's coding scheme primarily developed for the interpretation of design descriptions provides a tool for understanding the relationships between syntax, semantics and semiotics in visual communication and can be applied to other domains that utilise visual communication. Syntax refers to Dc (create a new depiction), semantics refer to P and F actions, as well as goals, and semiotics refer to M-actions (gestures) in Suwa's coding scheme.

According to Suwa's coding scheme, cognitive actions of designers are classified into four information categories: physical, perceptual, functional and conceptual. They claim that these four categories are classified according to the levels at which incoming information is thought to be processed in human cognition. Thus, physical actions correspond to the sensory level at which incoming information is first processed sensorially. Then, the incoming information is processed perceptually and semantically which are represented by perceptual actions, functional and conceptual actions respectively. Table 3.4 displays the four categories.

In Suwa's scheme, the levels of information processing have an inherent dependency on each other; processing at an upper level is based on that at lower level(s). At the same time, the relationships among those actions are also coded, i.e., which action(s) are dependent on, suggested by, or triggered by which actions. Table 3.5, Table 3.6 and Table 3.7 give the detailed definition of the cognitive actions belonging to the first three processing levels in Suwa's coding scheme.

In Suwa's coding, the fourth category (conceptual) refers to cognitive actions that are not directly suggested by physical depictions or visuo-spatial features of elements. There are three types of conceptual actions. The first type is the designer's preferential (like-dislike) or aesthetic (beautiful-ugly, good-bad, and so on) evaluation of P-actions or F-actions. It is called an E-action. For example, if a designer evaluated a spatial pattern of the flow of people as 'excellent' the judgement excellent is coded as an E-action. Another type of conceptual action is to set up goals, called G-action. A goal is born in a bottom-up way, triggered by P-actions or F-actions. Once a goal is set up,

Table 3.4: Cognitive Actions Categories (Suwa et al.) [1]

Category	Name	Description	Examples
Physical (Sensory information processing)	D-action	Make depictions	Lines, circles, arrows, words
	L-action	Look at previous depictions	-
	M-action	Other physical actions	Move a pen, move elements, gesture
Perceptual (Perceptual information processing)	P-action	Attend to visual features of elements	Shapes, sizes, textures
		Attend to spatial relations among elements	Proximity, alignment, intersection
		Organise or compare elements	Grouping, similarity, contrast
Functional (Semantical information processing)	F-action	Explore the issues of interactions between artifacts and people/nature	Functions, circulation of people, views, lighting conditions
		Consider psychological reactions of people	Fascination, motivation, cheerfulness
Conceptual (Semantical information processing)	E-action	Make preferential and aesthetic evaluations	Like-dislike, good-bad, beautiful-ugly
	G-action	Set up goals	-
	K-action	Retrieve knowledge	-

it in turn gives birth to other actions in a top-down way. The third type of conceptual action is retrieval of knowledge from memory, called K-action. Retrieval of knowledge and its application involves producing new pieces of information or goals in a top-down way.

TABLE 3.5: Codes of actions belong to 'physical' level (Suwa et al.) [1]

Drf: Revise the shape, size or texture of a depiction	Dc: Create a new depiction
Dts: Trace over a depiction on the same sheet of paper	Dtd: Trace over a depiction on a new sheet of paper
Dsy: Depict a symbol that represents a relation	Dwo: Write sentences or words that express ideas
L: Look at a previous depiction	Mrf: Move a pencil,attending to relationsor features
Mod: Move a pencil over a previous depiction	Ma: Move a depiction against the sheet beneath
Mut: Use tools	Mge: Hand gestures

The coding scheme developed by Suwa et al. was based on the architects' design activities. They detected a wide range of cognitive activities during the design session which was a complex task. They defined sub-classes for each physical, perceptual, functional and conceptual category. The details about the procedures and coding can be found in [1].

3.4.3 Coding Scheme Used in This Thesis

In our experiments, what we are concerned with is the temporal correlation between speech and hand gesture strokes, besides the gender differences in cognitive structures. We adapted the coding scheme to our purpose with some variations. The original

Table 3.6: Codes of P-actions (Suwa et al.) [1]

Psg: discover a space as ground	Pfn: attend to the feature of a new depiction
Pfnp: attend to the feature of a new relation or Psg	Pfp: discover a new feature of an existing depiction, of Pcsg, or of Prsg
Prn: create or attend to a new relation between two new depictions or Psg	Prnp: create or attend to a new relation between a new depiction and an existing one
Prp: discover a spatial or organizational relation	Pcf: continually attend to a feature
Pcr: continually attend to a relation	Pcsg: continually attend to a space as ground
Prf: remember a feature of a depiction	Prr: remember a spatial or organizational relation
Prsg: remember a space as ground	Pipsr: implement a previously mentioned relation by giving new depictions or features

coding scheme was developed by Suwa et al. for a design task with high complexity so that they defined a large number of codes for each of the four categories. In our case, the tasks of object descriptions are simpler than the design task in Suwa's case. So not all types of cognitive actions in Suwa's coding scheme can be found in our case. We, therefore, partly adopted their coding scheme.

Firstly, we coded hand gestures as M-actions in physical category represented by Mge in our coding. We expanded Mge into four sub-classes: Mgei indicating iconic gestures, Mged corresponding to deictic gestures, Mgem for metaphoric gestures and also Mgeb for beat gestures.

Secondly, there are three sub-categories in conceptual category: E-action, G-action

TABLE 3.7: Codes of F-actions (Suwa et al.) [1]

Fnp: think of a function independently of depictions	Fcp: continually think of a function independently of depictions
Fn: associate a new depiction, feature or relation with a new function	Fc: continually think of a function
Fr: remember a function	Fre-i: re-interpretation
Frp: remember a function independently of depictions	Fi: implement a previously explored function by creating a new depiction, feature or relation

and K-action. In our experiment, we seldomly observed these three actions. These categories therefore will not appear in our coding.

Thirdly, for the purpose of analysis and consistency, we will use G-actions to represent gestures which include all M-actions (in [1]) in the coding process. G-actions in this thesis is totally different from G-actions in Suwa's coding, which is a sub-category in conceptual level.

In summary, there are three categories in our coding: Mge (represented by G-action during analysis), P-actions (representing perceptual category) and F-actions (representing functional category). Table 3.6 and Table 3.7 show the sub-classes and codes for P-actions and F-actions respectively. Table 3.8 presents the codes used in the thesis. An example of cognitive action codings will be given in the next section.

After the completion of coding, we analysed the correlations and differences in the occurrence of these three types of cognitive actions.

3.4.4 An Example of Cognitive Action Coding

Cognitive coding was completed mainly by speech and gesture analysis via the annotation tools Praat and Anvil. Praat allows users to rehear any selected part of the audio

Table 3.8: Liu & Kavakli's coding scheme in this work

P		All P-actions defined in Table 3.6
F		All F-actions defined in Table 3.7
(G) Mge-hand gestures coded in Table 3.5	Mgei	Iconic gestures
	Mged	Deictic gestures
	Mgem	Metaphoric gestures
	Mgeb	Beats

(e.g. one segmentation) unlimited times to make the coding more reliable. We illustrate the coding procedures for one participant's protocol as follows and the procedures are applied for all protocols we collected.

For each audio clip, we first listened to it in Praat to find the turning point for segmentation. As described in the previous section, the change in participant attention is an important criterion for segmentation. Each segment was transcribed for later coding.

To give an example, one native participant's actions were annotated as follows. The two paragraphs below were transcribed when this participant described the two objects in Fig. 3.1 and Fig. 3.2.

For the first chair: *"[This is a chair. The **seat** of itself is about this **high** off ground, about that **wide** and about that **deep**.] [Each of leg, the **leg** on each of corner is **square** metal leg. That's running **down**.] [At the back there is a **square**, actually two square wooden **strips**. There is a **shaped** place in between. And there are actually **one**, **two**, **three**, **four** of those with a bit of **gap** to the seat]".*

For the second one: *"[For this chair, is pretty much a **heart shape** seat and comes right the way **down** to a point here,] [and down here is where it's got the feet. It's actually got **four** metal pieces that actually sit **on** the floor.] [And the **seat** itself is about this **high** off ground. It's **circular** seat. It's only about **25cm**, so only about this **big**.] [And the height of **back** above that is only **30cm**, so only about that height at level point.]"*

In the first paragraph, we can see that the participant changed her attention three

times, from the seat part to the leg part then to the back part. So that we got three segments for the first task. Each segment is put in square brackets in the above two paragraphs. The corresponding lexical affiliates of the gestures used by this participant for each segment are marked in bold face in the two description paragraphs.

Based on Suwa's coding scheme, the coding procedures are explained as follows. The cognitive actions extracted from these three segments are shown as below in Table 3.9, Table 3.10 and Table 3.11.

Segment 1: *[This is a chair. The **seat** of itself is about this **high** off ground, about that **wide** and about that **deep**.]*

Based on Suwa's coding scheme, when this participant first mentioned a chair, he actually started a new depiction, we therefore coded it as the first F-action (Fn1). He then talked about the seat, which started another new depiction about the seat. We coded it as another F-action (Fn2). When he said "this high off ground", he thought about the feature "height" the first time, we coded as a P-action (Pfp1). There was also a spatial relation between the seat and the ground mentioned here, we then coded another P-action (Prp). After that, the participant said "that wide and about that deep". "Wide" and "deep" were both about the features of the seat. We coded them as two P-actions (Pcf1 and Pcf2). This participant used hand gestures when he said "seat", "high", "wide" and "deep", therefore, there were four M-actions coded.

TABLE 3.9: Cognitive actions in Segment 1

Mge		P-actions		F-actions	
Mge (Mgei)	seat	Pfp1 (P)	height	Fn1 (F)	chair
Mge (Mgei)	high	Prp (P)	off ground	Fn2 (F)	seat
Mge (Mgei)	wide	Pcf1 (P)	wide		
Mge (Mgei)	deep	Pcf2 (P)	deep		

Segment 2: *[Each of leg, the **leg** on each of corner is **square** metal leg. That's running **down**.]*

In this segment, when the participant mentioned the leg first time, we coded it as

a F-action (Fn). When he said "the leg on each of corner is square metal leg", we coded another F-action (Fc). There was also a organisational relation between the leg and the seat, we coded as P-action (Prp); and another two P-actions (Pfp1 and Pcf) were referred to the the features (square and metal) of the leg. Hand gestures were annotated when the participant said "leg", "square" and "down".

Table 3.10: Cognitive actions in Segment 2

Mge		P-actions		F-actions	
Mge (Mgei)	leg	Prp (P)	on the corner	Fn (F)	leg
Mge (Mgei)	square	Pfp1 (P)	shape(square)	Fc (F)	leg
Mge (Mgei)	down	Pcf (P)	metal		

Segment 3: *[At the back there is a **square**, actually two square wooden **strips**. There is a **shaped** place in between. And there are actually **one**, **two**, **three**, **four** of those with a bit of **gap** to the seat.]*

In this segment, the first time when the participant talked about the back, we coded as a F-action (Fn1). He then said "two square wooden strips". Numeric information (two) was coded as a F-action (Fnp), "square" was about the feature which was coded as a P-action (Pfp1) and "strips" was a part of the chair mentioned first time which was coded as another F-action (Fn2). When he talked about "there is a shaped place in between", there was another feature "shaped" and an organisational relation (in between) mentioned, we therefore coded another two P-actions (Pcf1 and Prp). In the last sentence, the numeric information was coded as another F-action (Fc). Another two P-actions (Pcf2:feature of the strips and Pcr:alignment of the strips and the seat) were also coded. There were eight hand gestures used in this segment.

For the second chair, we segmented the description into 4 segments. As can be seen in the second paragraph, the participant first mentioned the seat part, then the feet part, and then went back to the seat again. The back part was described at the last. The cognitive actions extracted for the second description are displayed in the Table 3.12, Table 3.13, Table 3.14 and Table 3.15.

TABLE 3.11: Cognitive actions in Segment 3

Mge		P-actions		F-actions	
Mge (Mgei)	square	Pfp1 (P)	shape(square)	Fn1 (F)	back
Mge (Mgei)	strips	Pcf1 (P)	shape(shaped)	Fnp (F)	numeric info
Mge (Mgei)	shaped	Prp (P)	in between	Fn2 (F)	stripes
Mge (Mgeb)	1	Pcf2 (P)	alignment(gap)	Fc (F)	numeric info
Mge (Mgeb)	2	Pcr (P)	alignment(to the seat)		
Mge (Mgeb)	3				
Mge (Mgeb)	4				
Mge (Mgei)	gap				

Segment 4: *[For this chair, is pretty much a **heart shape** seat and comes right the way **down** to a point here,]*

In segment 4, when the participant first time mentioned the chair, we coded a F-action (Fn). When he talked about the heart-shape seat, a P-action (Pfp) and another F-action (Fn) were coded. When the alignment of the heart-shaped seat were talked, a P-action was identified (Prp). Two hand gestures were annotated with "heart-shape" and "down".

TABLE 3.12: Cognitive actions in Segment 4

Mge		P-actions		F-actions	
Mge (Mgei)	heart shape	Pfp (P)	shape(heart shape)	Fn (F)	chair
Mge (Mgei)	down	Prp (P)	alignment (comes down)	Fn (F)	seat

Segment 5: *[and down here is where it's got the feet. It's actually got **four** metal pieces that actually sit **on** the floor.]*

In segment 5, "down here" was about the spatial location which had been mentioned in the previous segment, a P-action was therefore coded (Pcr1). He then mentioned the feet, which was coded as a F-action (Fn). In the second sentence, numeric information was coded as another F-action (Fc1), while the feature "metal" was coded as another

P-action (Pcf). "the pieces" were actually the part of the chair, which was coded as a F-action (Fc2). There was also another P-action (Pcr2) coded when the participant talked about the relation of the pieces and the floor. Only two hand gestures were detected in this segment.

Table 3.13: Cognitive actions in Segment 5

Mge		P-actions		F-actions	
Mge (Mgei)	4	Pcr1 (P)	alignment (down here)	Fn (F)	feet
Mge (Mgem)	on	Pcf (P)	metal	Fnp (F)	numeric info
		Pcr2 (P)	alignment(sit on the floor)	Fc2 (F)	pieces

Segment 6: *[And the **seat** itself is about this **high** off ground. It's **circular** seat. It's only about **25cm**, so only about this **big**.]*

In segment 6, there were two F-actions (Fn and Fnp) detected when the participant mentioned the seat and talked the numeric information. Four P-actions were coded as following: "this high" was about the height of the seat (Pfp1); "off ground" was about the alignment of the seat and ground (Prp); "circular" was about the shape of the seat (Pcf1); "this big" was about size of the seat (Pfp2). There were four hand gestures were annotated from this segment.

Table 3.14: Cognitive actions in Segment 6

Mge		P-actions		F-actions	
Mge (Mgei)	seat	Pfp1 (P)	high	Fn (F)	seat
Mge (Mgei)	high	Prp (P)	alignment(from ground)	Fn (F)	numeric info
Mge (Mgei)	circular	Pcf1 (P)	shape(circular)		
Mge (Mgei)	25cm	Pcf2 (P)	size of the seat		

Segment 7: *[And the height of **back** above that is only **30cm**, so only about that height at **level** point.]*

In segment 7, Two F-actions (Fn and Fnp) were coded when the participant mentioned the back of the chair and the numeric information. Three P-actions were detected, while two of them were about spatial or organisational relation (Prp and Pcr) and one is about the feature of the back (Pfp). Three hand gestures were annotated in this segment.

TABLE 3.15: Cognitive actions in Segment 7

Mge		P-actions		F-actions	
Mge (Mgei)	back	Prp (P)	above	Fn (F)	back
Mge (Mgei)	30cm	Pfp (P)	height	Fnp (F)	numeric info
Mge (Mged)	level	Pcr (P)	alignment(at level point)		

Before we started the cognitive coding, the gesture annotation has actually finished in Anvil. So we mainly coded cognitive actions at this stage. In total, seven segments were coded for this participant.

Based on the definition of P-actions and F-actions in Table 3.4, Table 3.6 and Table 3.7, we can see that F-actions are generally associated with functions of the elements of the object (such as seat, back, leg and strips etc.) and also some numeric information (such as 4, 25 and 30 etc.) in our experiment, while the majority of P-actions represent participant's attention to the features of the object (such as square, metal and circular etc.) or the position and alignment of elements (such as in between, from the ground and on etc.) in our experiment.

As introduced in the previous section, the granularity of the coding scheme used in this thesis is different from Suwa's scheme, as many subcategories were not detected in our case. This also can be seen in the example given above. In the final analysis, we did not distinguish the subcategories in P-actions and F-actions. The subcategories in Mge help us understand the characteristics of gestures, but they are categorised as G-action in the post analysis. Our codes used in analysis are included in parentheses in above tables.

3.5 Gender Differences in Speech and Hand Gestures and Their Temporal Alignment

3.5.1 Video Data Corpus

A total of 18 individuals (9 males and 9 females) participated in the first experiment. The age range of them is between 20 and 50. For all female participants, we captured the video records of 1547 seconds in total. The shortest video record is 65s and the longest one is 410s (Mean(M)=171.9s, (Standard Deviation)SD=131.2s). For all male participants, we obtained the video records of 1178 seconds, which ranged in length between 49s and 312s (M=130.9s, SD=77.3s). Fig. 3.19 shows the distribution of video length for female and male participants.

Before detailed analysis of the speech and hand gestures extracted from these video clips, we first present some fundamental differences about the video length, which can be regarded as task time of different participants. From Fig. 3.19 we can see that the median values are quite close. But the task time of female participants are more largely dispersing from the mean value compared to male participants.

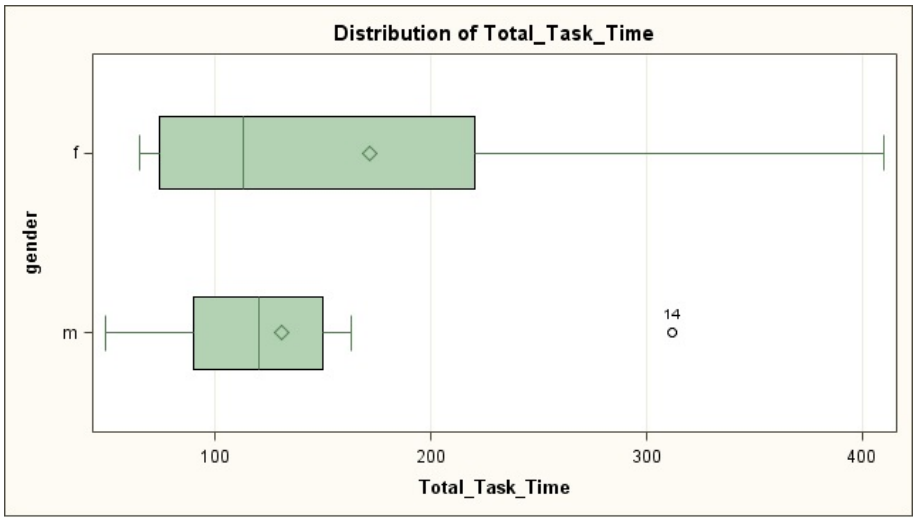


Figure 3.19: Distribution of video length for female and male participants

In order to check if there are significant differences between task time for males

TABLE 3.16: t-Test for total task time of different gender

Gender	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
f	171.9	71.02	272.8	131.2	88.64	251.4
m	130.9	71.43	190.3	77.35	52.24	148.2

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	16	0.81	0.43
Satterthwaite	Unequal	12.96	0.81	0.43

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	8	8	2.88	0.16

and females, we applied two independent samples of t-Test for the variables with the significance level at α=.05. The data set of two variables (task time for males and females) was imported to SAS Enterprise Guide (SAS is used for short in the following) to complete the analysis. Table 3.16 gives the statistical results produced by SAS.

The first sub-table in Table 3.16 displays the means and standard deviations based on the given confidence level (95%) regarding task time.

The second sub-table gives the test statistics, associated degrees of freedom, and p-values using two methods (Pooled and Satterthwaite) in SAS. The Pooled test assumes that the two populations have equal variances and uses degrees of freedom n_1+n_2-2, where n1 and n2 are sample sizes for the two populations. The Satterthwaite test does not assume that the populations have equal variances and uses the Satterthwaite approximation for degrees of freedom.

The third sub-table displays the "Equality of Variances" test reveals insufficient evidence of unequal variances (the Folded F statistic F=2.88, p=0.16). This means that the p-value (p=0.43) obtained by Pooled test reveals no significant statistical differences in terms of task time for males and females. If the "Equality of Variances" test with the result p<.05, the t-Test result generated by Satterthwaite method will be more reliable for decision making.

The two independent samples of t-Test will be used as the fundamental tool in the following sections for the statistical analysis to study gender differences.

After the completion of gesture and speech annotation, a summary was generated in Table 3.17 for females and Table 3.18 for males. In the experiment, there were two tasks (describing two objects) for each participant. We summarized the annotation results for each task respectively. So each table is divided into three parts to show the annotation for each task separately and the summary of two annotations as well.

In each table for each task, we recorded the task time, gesture number used in the task and stroke time. Based on the recorded data, we computed the average stroke time and stroke time proportion from the task time. We then summed them in total.

For example, from the first row of Table 3.17 we can see that a female participant (p1) spend 171 seconds on the first task and had 29 hand gestures. The total stroke time for these 29 hand gestures is 27.93 seconds, while the average stroke time for each gesture is 0.96 seconds. The time proportion for all gesture strokes over the total time in task 1 for this participant is 16.33%. In task 2, this participant took 194 seconds to finish and had 30 hand gestures. The stroke time for all gestures in task 2 is 41.86 seconds and average stroke time for each gesture is 1.40. She spent 21.58% of task time of task 2 on gesture strokes. In total for this female participant, we recorded 365-seconds video clip for her. 59 hand gestures were annotated. The total stroke time of all gestures is 69.79 seconds with average 1.18 seconds for each and time proportion over all task time spent on gestures is 19.12% regarding both tasks. Each column in Table 3.18 has the same meanings of Table 3.17 for male participants.

	Task 1					Task 2					Total				
P	Task time(s)	Gesture no.	Stroke time(s)	Average stroke time(s)	Stroke time proportion	Task time(s)	Gesture no.	Stroke time(s)	Average stroke time(s)	Stroke time proportion	Total task time(s)	Gesture no.	Total stroke time(s)	Average stroke time(s)	Stroke time proportion
p1	171	29	27.93	0.96	16.33%	194	30	41.86	1.40	21.58%	365	59	69.79	1.18	19.12%
p2	230	30	19.59	0.65	9.37%	180	28	40.8	1.46	22.67%	410	58	62.35	1.02	15.21%
p3	34	8	9.24	1.16	37.35%	40	8	23.5	2.94	58.75%	74	16	36.2	1.91	48.92%
p4	50	15	26.6	1.57	53.20%	57	11	22.95	2.09	40.26%	107	26	49.55	1.77	46.31%
p5	49	14	20.2	1.19	41.20%	64	15	20.12	1.34	31.44%	113	29	40.31	1.26	35.67%
p6	53	10	27.46	2.75	51.81%	70	10	40.2	4.02	57.43%	123	20	67.6	3.38	55.01%
p7	33	9	18.13	2.014	54.94%	32	8	12.44	1.56	38.88%	65	17	30.57	1.80	47.03%
p8	112	10	9.7	0.88	8.66%	108	10	6.71	0.67	6.21%	220	20	16.41	0.78	7.46%
p9	34	11	16.62	1.51	48.88%	36	4	15.04	3.76	41.78%	70	15	31.66	2.11	45.23%

Table 3.17: Summary of gesture and speech annotation for females

P	Object 1					Object 2					Total				
	Task time(s)	Gesture no.	Stroke time(s)	Average stroke time(s)	Stroke time proportion	Task time(s)	Gesture no.	Stroke time(s)	Average stroke time(s)	Stroke time proportion	Total task time(s)	Gesture no.	Total stroke time(s)	Average stroke time(s)	Stroke time proportion
p1	71	15	7.75	0.52	10.92%	92	15	10.01	0.67	10.88%	163	30	17.76	0.59	10.90%
p2	28	6	10.04	1.67	35.86%	21	5	6.1	1.22	29.05%	49	11	16.41	1.47	32.94%
p3	68	6	11.03	1.84	16.22%	41	4	9.33	2.33	22.76%	109	10	20.36	2.04	18.68%
p4	59	10	6.56	0.66	11.12%	91	16	8.06	0.50	8.86%	150	26	14.62	0.56	9.75%
p5	129	13	15.87	1.22	12.30%	183	22	27.11	1.23	14.81%	312	35	42.98	1.23	13.78%
p6	55	13	11.62	0.89	21.13%	65	11	9.79	0.89	15.06%	120	24	21.41	0.97	17.84%
p7	35	13	6.45	0.58	9.86%	55	16	10.27	0.95	18.67%	90	29	13.72	0.92	15.24%
p8	60	10	9.57	0.87	15.95%	62	13	9.38	0.78	15.13%	112	23	18.95	0.82	15.53%
p9	27	7	5.41	0.77	20.04%	36	3	5.29	0.66	14.69%	63	10	10.7	0.71	16.98%

Table 3.18: Summary of gesture and speech annotation for males

As displayed in Table 3.16, in total there are no significant differences in the task time for different gender. For different tasks (descriptions of object 1 and object 2), we can obtain the same conclusion based on the t-Test results (p=0.34 for task 1 as in Table 3.19 and p=0.57 for task 2 as in Table 3.20). It means that male participants and female participants spent similar amount of time on the same tasks.

TABLE 3.19: t-Test for task time of task 1

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	16	1	0.33
Satterthwaite	Unequal	10.93	1	0.34

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	8	8	5.27	0.03

TABLE 3.20: t-Test for task time of task 2

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	16	0.58	0.57
Satterthwaite	Unequal	15.12	0.58	0.57

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	8	8	1.64	0.50

3.5.2 Fundamental Differences in Hand Gestures and Corresponding Lexical Affiliates

In this section, we will analyse fundamental differences in speech and gestures used by males and females from the following aspects: 1) the total time they spent on gestures;

2) the average gesture stroke time; 3) the time proportion of total task time on gesture strokes. 4) properties of hand gesture related keywords.

1. Total time Spent on Gestures

During gesture annotation, as stated in Section 3.3.5, we used the 'stroke' phase of each gesture to represent that gesture, since the 'stroke' phase is the 'most energetic part of the gesture movement and also the requisite part of a gesture'. Stroke time in Table 3.17 and Table 3.18 means the stroke time of all gestures for task 1, task 2 and the sum of them respectively.

Fig. 3.20 displays the distribution of total stroke time for task 1 of males and females. For females, the mean value (M=20.10s) is quite close to the median value while for males, the mean value (M=9.03s) is less than the median value. Fig. 3.20 shows clearly the greater skewness of distribution in gesture stroke time spent by females.

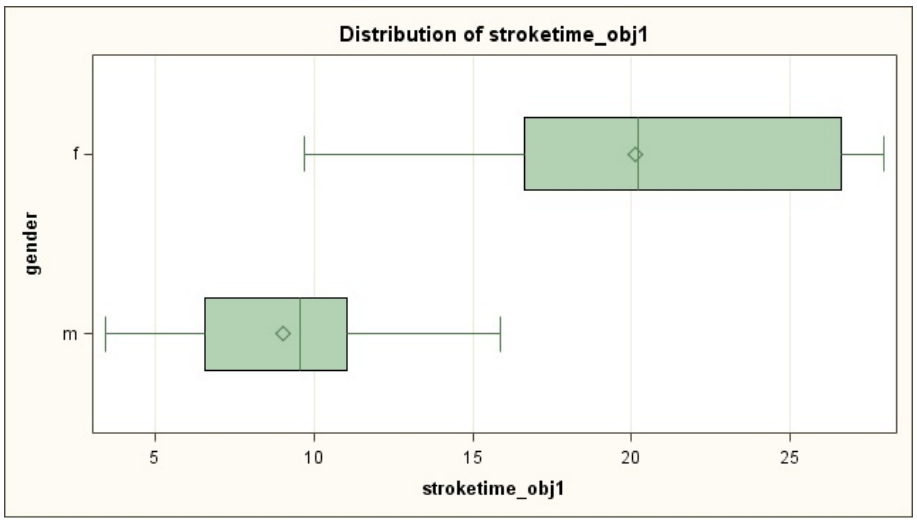


Figure 3.20: Distribution of gesture stroke time of different gender for task 1

Table 3.21 shows the descriptive data for gesture stroke time spent by male and female participants and also the t-Test results. Based on 95% confidence limits for mean value, the gesture stroke time spent by female participants is in the range of 15.10s-25.10s and for male participants the range is between 6.17s-11.90s. The results

disclose a significant statistical difference of the data from two genders ($p<0.0004$).

TABLE 3.21: t-Test for gesture stroke time of different genders for task 1

Gender	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
f	20.10	15.10	25.10	6.51	4.40	12.47
m	9.03	6.17	11.90	3.72	2.51	7.13

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	16	4.43	0.0004
Satterthwaite	Unequal	12.73	4.43	0.0007

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	8	8	3.06	0.13

For task 2, the mean value (M=24.85) is a little bit greater than the median value for females while it (M=10.59) is also greater than median for males. We basically obtained the similar conclusion with p=0.01, which indicates the differences between gesture stroke time for two groups. In Fig. 3.21, there is one extreme outlier of the 9 male participants indicated by box plot generated by SAS. This participant was a male in his 50s. One thing we noted from his video clip was that he repeated some of his descriptions several times in task 2, which resulted in longer gesture stroke time than normal. Nonetheless, when this individual was excluded from the data set, the yielded measures do not change the results in Table 3.22. As the test for equality of variances had an associated p-value<.0001, a modified version of the t-Test was applied. The result of modified test (Satterthwaite test) is also given in Table 3.22, which still shows evidence for a difference in the means of the gesture stroke time of males and females.

In Fig. 3.22 (distribution of gesture stroke time of the sum of the two tasks) and Table 3.23 (the extreme outlier in males was excluded) we can see in total, **gesture stroke time spent by two different genders is significantly different (Finding**

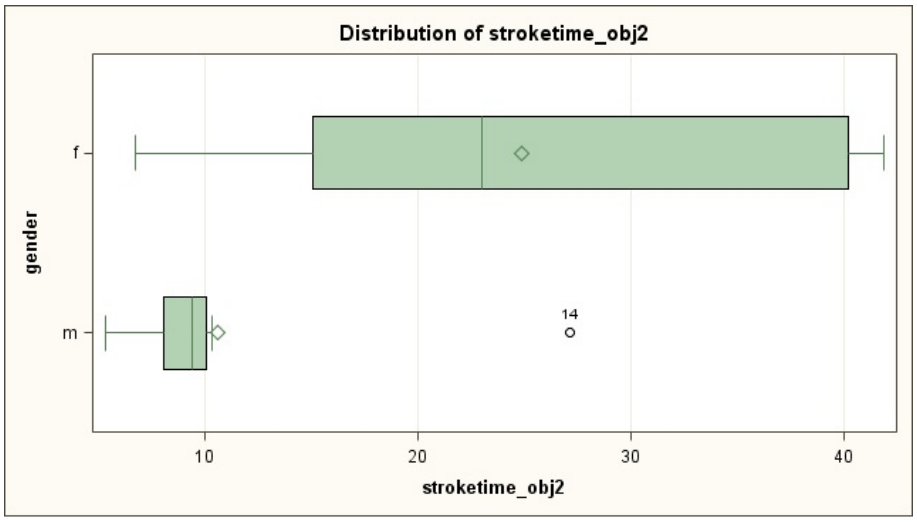


Figure 3.21: Distribution of gesture stroke time of different genders for task 2

Table 3.22: t-Test for gesture stroke time of different genders for task 2

Gender	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
f	24.85	14.73	34.97	13.16	8.89	25.22
m	8.53	6.96	10.10	1.88	1.24	3.83

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	15	3.46	0.003
Satterthwaite	Unequal	8.37	3.68	0.006

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	8	7	48.93	0.0001

1) as equality of variances with p=0.0007 and Satterthwaite test with p=0.0017.

2. Average Stroke Time for Gestures

In Table 3.17 and Table 3.18, the columns with name "Average stroke time" represent the average stroke time for each gesture. The average stroke time was calculated from

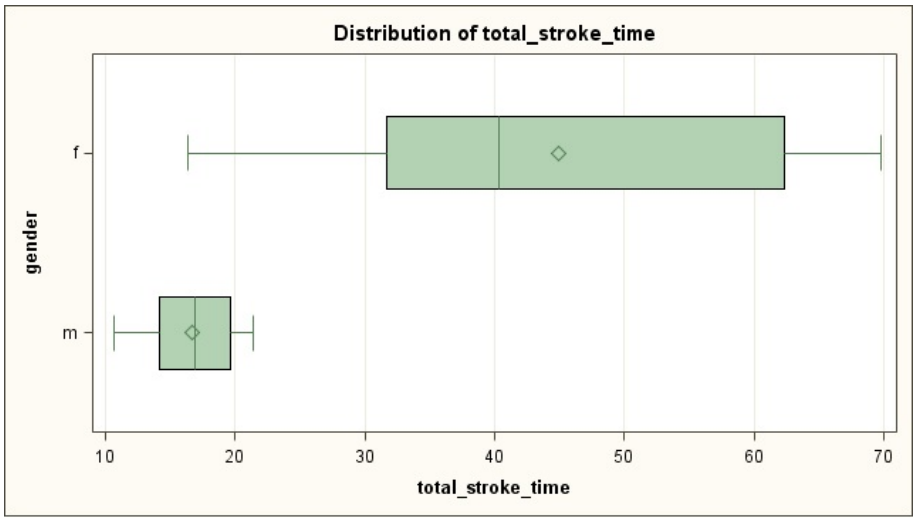


FIGURE 3.22: Distribution of gesture stroke time of the sum of two tasks

TABLE 3.23: t-Test for gesture stroke time of different genders for the sum of two tasks

Gender	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
f	44.94	30.69	59.20	18.55	12.53	35.54
m	16.71	13.69	19.73	3.61	2.39	7.35

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	15	4.22	0.0007
Satterthwaite	Unequal	8.6787	4.47	0.0017

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	8	7	26.39	0.0003

gesture stroke time divided by gesture numbers used in each task and in total. It can also viewed as the average length of "stroke" phase of gestures. We used the same analysis method as we did in Section 3.4.1.

In task 1, the mean value for females is M=1.41 with SD=1.24 while M=1.00 with SD=0.91 is for males. The two independent samples t-Test reveal a marginal difference with p=0.14.

In task 2, the mean value is greater than in task 1 for females (M=2.13, SD=2.42). However these values are quite similar for males (M=1.03, SD=1.05). The p-value of t-Test shows evidence for a difference in average stroke for genders with p=0.02.

In total, the p-value (0.04), although less significant than in task 2 (p=0.02), still indicates that the average stroke time is different for different genders. **On average, the length of gesture strokes is longer for females (M=1.69, SD=1.33) than for males (M=1.03, SD=0.81) (Finding 2)**.

3. Gesture Stroke Time Proportion of Total Task Time

As we found in Section 3.4.1, the statistical analysis revealed no significant differences in task time of males and females but showed evidence for differences in gesture stroke time of males and females. It might be an indication that during the same task male and female participants have different preferences to use gestures and speech. The results of two independent samples of t-Test given in Table 3.24 (with p=0.02), Table 3.25 (with p=0.01) and Table 3.26 (with p=0.008) confirm the assumption. As displayed in these three tables (Table 3.24, Table 3.25 and Table 3.26), no matter in task 1, task 2 or as a whole, the time proportion spent for gesture strokes of males and females are significantly different. Females spent more time on gesture strokes than males. In task 1, on average, females spent 35.73% (vs 17.02% for males) of task time on gesture strokes. In task 2 females spent 35.39% (vs 16.61%) of task time on gesture strokes. In total, the time proportion spent by females and males on gesture strokes are 35.53% and 16.80% respectively.

We also noticed for male and female participants in different tasks they presented similar preferences of using hand gestures and speech. Male participants spent 17.02% of task time on gestures in task 1 on average with a range of 10.78% and 23.25% at 95% confidence limits for the mean, while they spent 16.61% of task time in task 2 on average with a range of 11.89% and 21.32% at the same confidence level.

Female participants spent 35.73% of task time on gestures in task 1 on average with a range of 20.99% to 50.46% at 95% confidence limits for the mean. The average

TABLE 3.24: t-Test for gesture stroke proportion of different genders for task 1

Gender	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
f	35.73	20.99	50.46	19.17	12.95	36.72
m	17.02	10.78	23.25	8.11	5.45	15.54

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	16	2.7	0.02
Satterthwaite	Unequal	10.76	2.7	0.02

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	8	8	5.59	0.03

value changes a little to 35.39% in task 2 and ranges between 22.28% and 48.50% at the same confidence level.

The p-values of t-Test results in Table 3.24, Table 3.25 and Table 3.26 indicate that **the preferences of using speech and hand gestures for males and females in the same task vary**. Even though speech is still the dominant communication medium, **females prefer to spend more time to gesture than males and also their preferences are stable over time even if the tasks are different (Finding 3).**

4. Properties of hand gesture related keywords

Regarding the speech annotation, we recorded the corresponding lexical affiliates of hand gestures. We found that all keywords can be categorised into the following six classes: adjective (adj), noun (n), numeric (nu), adverb (adv), preposition (prep) and verb (v). Fig. 3.23 displays the frequencies chart of these categories for all participants.

From this figure we can see that the majority of keywords correspond to adjectives and nouns both for males and females. **For most females, nouns are the dominant**

Table 3.25: t-Test for gesture stroke proportion of different genders for task 2

Gender	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
f	35.39	22.28	48.50	17.05	11.52	32.67
m	16.61	11.90	21.32	6.13	4.14	11.75

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	16	3.11	0.007
Satterthwaite	Unequal	10.03	3.11	0.01

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	8	8	7.74	0.009

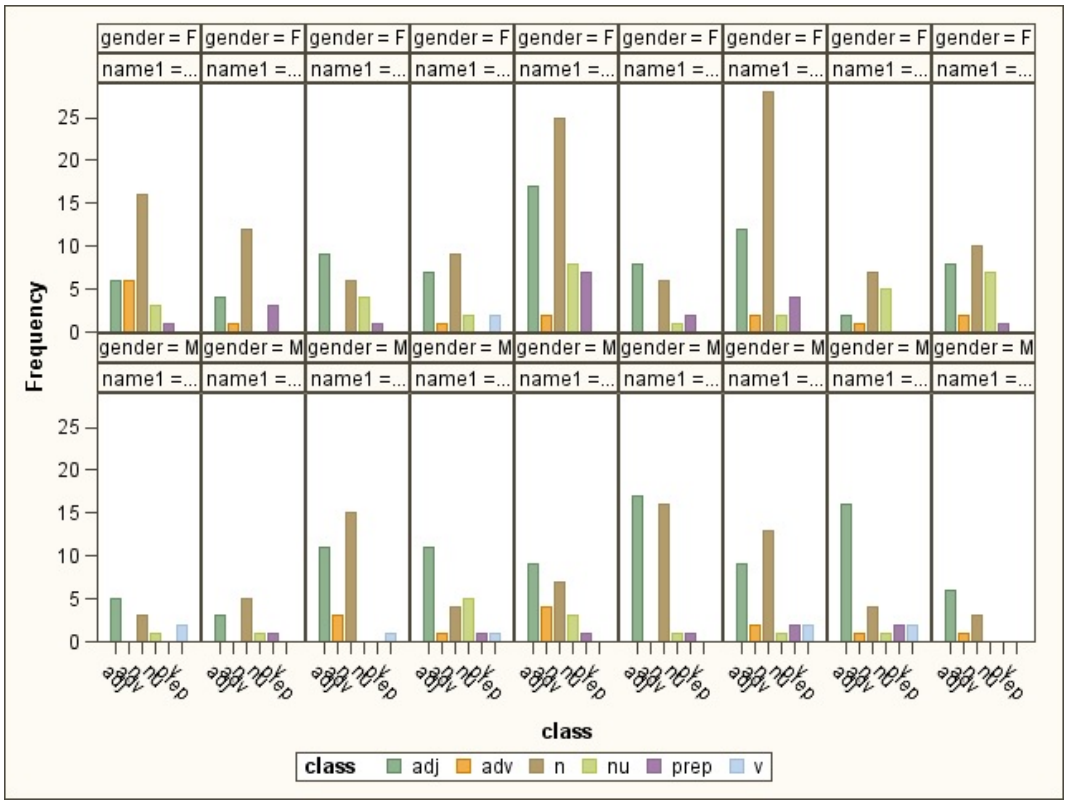


Figure 3.23: Types of gesture corresponding lexical affiliates

part of all other types, while for most males adjectives are the majority (Finding 4). The proportions of other types are quite minor in either gender group. We will interpret these differences in the cognitive analysis in the next section.

TABLE 3.26: t-Test for gesture stroke proportion of different genders in total

Gender	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
f	35.53	22.30	48.76	17.21	11.63	32.97
m	16.80	11.64	21.97	6.73	4.54	12.88

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	16	3.04	0.008
Satterthwaite	Unequal	10.39	3.04	0.01

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	8	8	6.55	0.02

Up to now we demonstrated gender differences in using speech and hand gestures for the same tasks. For instance, females use more gestures than males and the average length of gesture strokes is also longer for females. The dominant types of hand gesture-related keywords are different for males and females. In the following section, we will go into further detail by examining temporal differences in hand gestures and their lexical affiliates used by male and female participants.

3.5.3 Temporal Alignment of Speech and Hand Gestures

As shown in Table 3.17 and Table 3.18, we annotated 136 hand gestures for all females in task 1 and 124 in task 2. There are 260 gesture annotations in total for female participants. For males, 93 hand gestures were annotated in task 1 and 105 in task 2 and 198 in total. The corresponding lexical affiliates for each hand gesture were also transcribed. The temporal parameters (start point and end point for each gesture and its lexical affiliate) were recorded for the post annotation analysis.

1. Lengths of gesture strokes and their lexical affiliates

As demonstrated in Section 3.4.2, on average female participants had longer gesture stroke compared to male participants in summary. The average value of gesture stroke length was calculated through averaging stroke time by total gesture numbers. The average value may not be enough to represent the statistical nature of the length of all gesture strokes. We got the length of each gesture stroke by (end point) - (start point) of gesture. Two independent samples t-Test was applied to study if the means of the length of all gesture strokes are different for males and females. The test results are given in Table 3.27.

Table 3.27: t-Test for the length of all gesture strokes by genders

Gender	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
f	1.45	1.29	1.61	1.32	1.22	1.45
m	1.22	1.12	1.33	0.77	0.70	0.85

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	456	2.12	0.04
Satterthwaite	Unequal	429.27	2.26	0.02

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	259	197	2.94	<.0001

From Table 3.27 we can see that females generally have longer gesture strokes than males (Female: M=1.45, SD=1.45; Male: M=1.22, SD=0.85). Fig. 3.24 shows greater deviation from normal distribution of females' data, which suggests some caution is needed in interpreting the results from our two independent samples t-Test. Fortunately, the t-Test is known to be relatively robust against departures from normality. Even though the Equality of Variances test result ($p<.0001$) shows that the variances equality assumption is invalid, the Satterthwaite test that dropped the equality of variances assumption still reveal a significant difference in the population means about

the gesture stroke length of genders (p=0.02). It is consistent with our **Finding 2** in previous section.

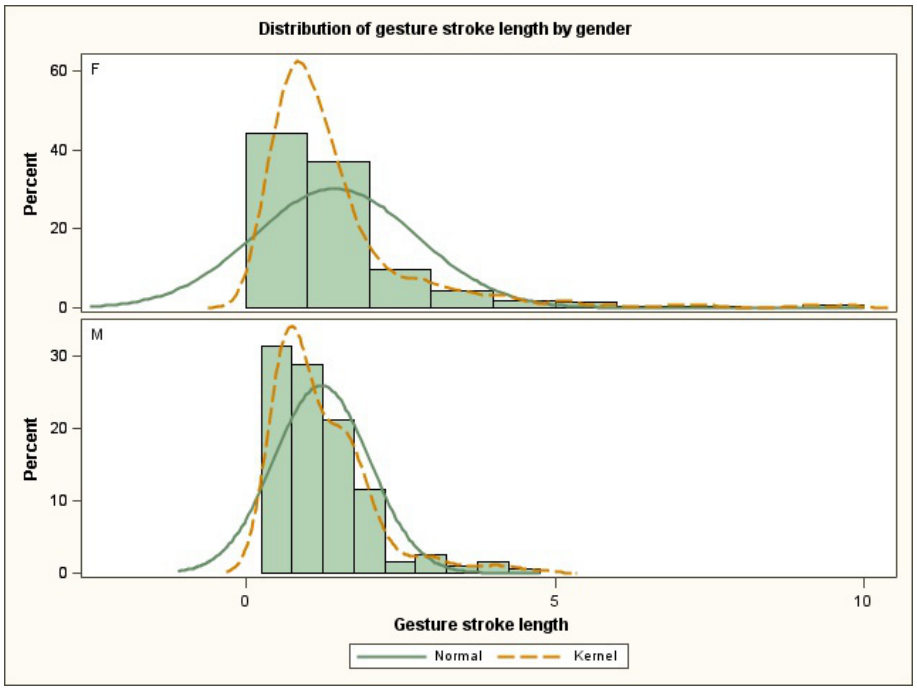


FIGURE 3.24: Distribution of gesture stroke length of genders

We also calculated the length of the corresponding lexical affiliates of gestures by (end point) - (start point) of each lexical affiliate. The results of two independent samples t-Test are given in Table 3.28 and Fig. 3.25. The distribution plots by histogram in Fig. 3.25 indicate that both data of males and females are close to normal distribution. It confirms the t-Test result (p=0.01 in the second sub-table of Table 3.28) that **the means of the length of lexical affiliates of two groups are statistically different (Finding 5)**.

2. Temporal alignment intervals

Due to the time sensitive architecture of MMIS, it is particularly important to find out the temporal alignment of a hand gesture and its lexical affiliate. It is already

Table 3.28: t-Test for the length of lexical affiliates by genders

Gender	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
f	0.55	0.53	0.58	0.20	0.19	0.22
m	0.50	0.47	0.53	0.20	0.18	0.22

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	456	2.59	0.01
Satterthwaite	Unequal	427.4	2.59	0.01

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	259	197	1.03	0.82

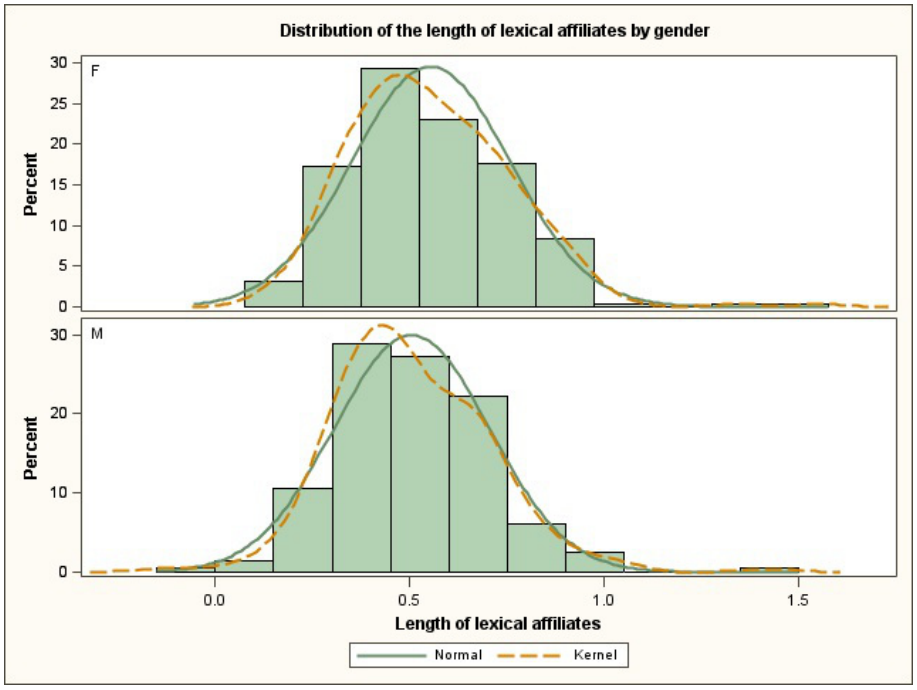


Figure 3.25: Distribution of the length of lexical affiliates of genders

known that gestures normally precede or fully synchronise with their lexical affiliates [39, 40, 41, 42, 43, 44, 45, 23, 46, 180]. The results of our analysis also show that even the onset of gesture strokes precede the related keywords in most cases and only

a small percentage of gesture strokes start behind the corresponding lexical affiliates. Our goal is to study the gender-based differences between the temporal alignment of speech and hand gestures rather than how they temporally align with each other in male and females.

Fig. 3.26 displays the distribution of time intervals of hand gesture strokes and their lexical affiliates of all annotations from male and female participants. The time interval is computed by (the start point of its lexical affiliate) - (the onset of hand gesture stroke). It is apparent from Fig. 3.26 that the majority of data (84.72%) is located in the interval that is greater or equal than zero, which means that **the majority of stroke phases of hand gestures started before the corresponding lexical affiliates (Finding 6)**. In our annotations for female participants, 81.15% of hand gesture strokes preceded the related lexical affiliates. For male participants, it reaches to 89.39%.

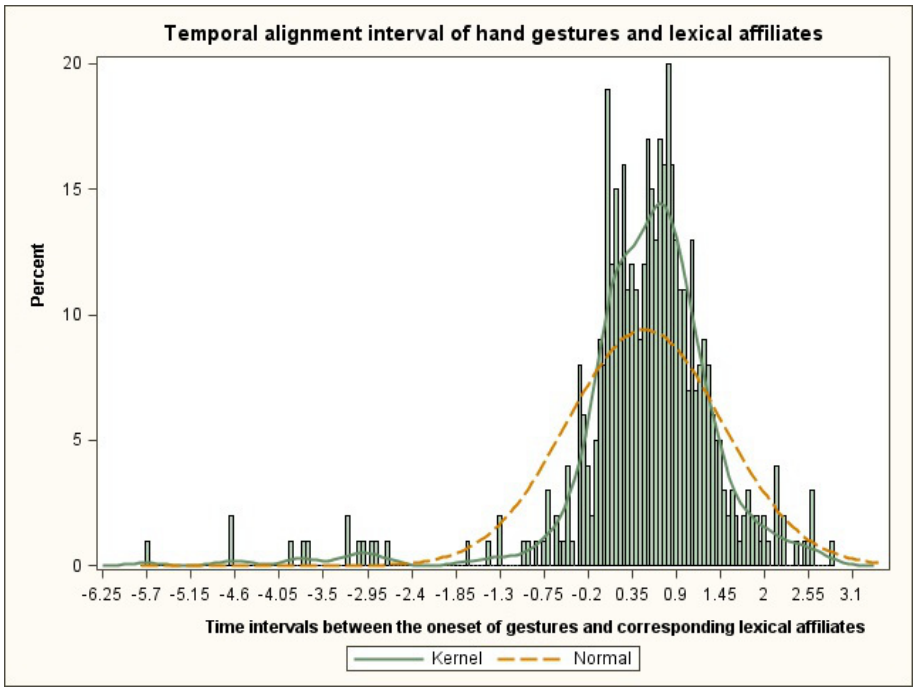


FIGURE 3.26: Distribution of temporal alignment of hand gestures and lexical affiliates

Beside the time intervals between the onset of gesture strokes and corresponding

lexical affiliates, we also checked the overlaps of hand gestures and corresponding keywords. Overall, for females 79.62% of hand gestures overlap with related keywords, while for males 70.71% cases have overlaps. Given the previous results about start time intervals, we can conclude that males and females have similar integration patterns, that is, gesture stroke phases precede the corresponding lexical affiliates with overlaps. The dominant patterns can be illuminated in Fig. 3.27, which shows that speech synchronises with hand gestures dominantly. However, the characteristics of the integration are varied as presented in the previous section as well as the temporal alignment.

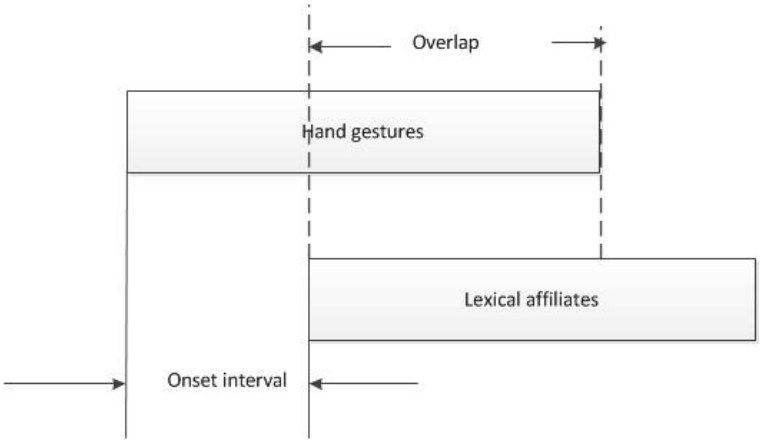


Figure 3.27: Dominant integration patterns of speech and hand gestures in males and females

In the following section, for males and females, we will mainly analyse the differences in temporal alignment intervals of males and females which are greater than or equal to zero.

The distribution plots of the time intervals that are greater than or equal to zero in Fig. 3.28 show no significant departures from normality for both genders where T-test is applicable.

From the summary in Table 3.29 we can clearly see that on average the time interval is longer for males (M=0.87s) than for females (M=0.70s). The time intervals fall between 0.79s and 0.95s at the 95% confidence level for males while for females the

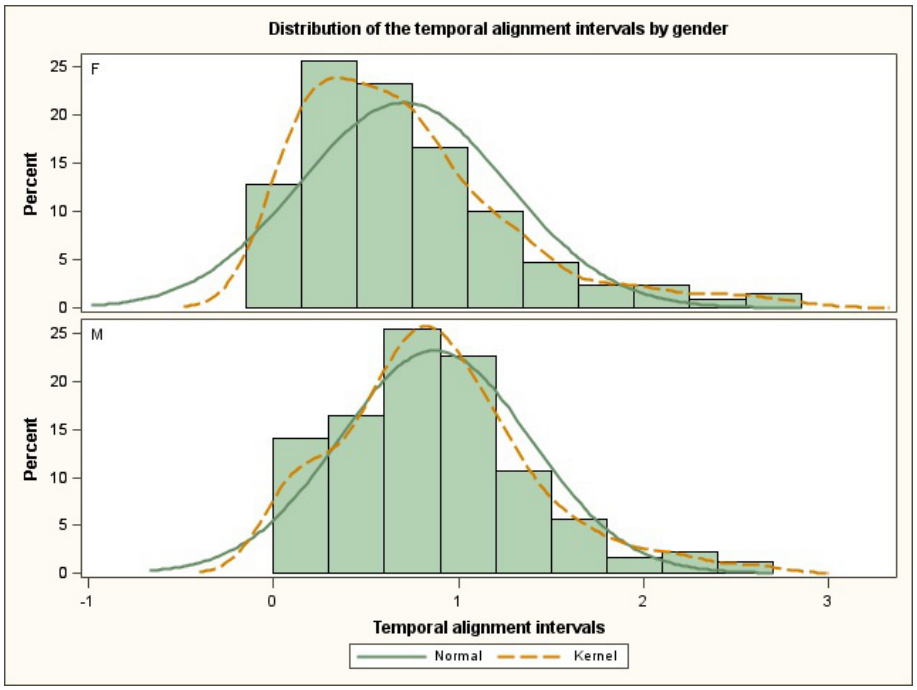


FIGURE 3.28: Distribution of the temporal alignment intervals of genders

time intervals are between 0.63s and 0.77s at the same confidence level. The two independent samples t-Test in this table also shows evidence for a difference in the population means of the time intervals between hand gesture strokes and their lexical affiliates of genders. The initial results from our experiments indicate that **males may need longer time to integrate speech and co-occurring hand gestures than females (Finding 7)**.

3.5.4 Evaluation of Results in Gender Differences of Speech and Hand Gestures and Their Temporal Alignment

Up to now, we studied the general gender differences in speech and hand gestures.

We found for the same tasks, males and females spend similar amount of time to finish. However, we found females spend more absolute time on gestures than males regardless of the total task time (**Finding 1**). For females, the amount of absolute time spent on gestures also takes a greater proportion of the total task time than males.

Table 3.29: t-Test for the temporal alignment intervals by genders

Gender	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
f	0.70	0.63	0.77	0.56	0.51	0.62
m	0.87	0.79	0.95	0.51	0.47	0.57

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	386	-3.08	0.002
Satterthwaite	Unequal	383.09	-3.11	0.002

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	210	176	1.20	0.22

This indicates that females may prefer to use more gestures than males. We also found that their preference of using speech and hand gestures are stable over the time from one task to another (**Finding 2**).

We also investigated the properties of hand gesture related keywords. We found that for most females, nouns are the dominant part of all other types, while for most males adjectives are the majority (**Finding 4**). This is relevant to cognitive processing which we will investigate in the next section.

Regarding to the design of MMIS, the ultimate goal of MMIS is to eliminate the gap between HCI and human-human communication by providing users with choice and switch of input. The differences we found in two gender groups may indicate that different speech and gesture vocabularies can be designed for different groups which may improve the performance of the system by allowing users with different choices.

As introduced in Chapter 2, the performance of each single channel as well as the integration of all channels in MMIS can affect the overall performance of the whole system. We found the length of gesture strokes and the length of corresponding lexical affiliates are significantly different for males and females (**Finding 3 and Finding 5**). This can be influential for the performance of speech and gesture channels in MMIS.

As for the integration of speech and hand gestures, we found that males and females have similar integration patterns, that is the majority of stroke phases of hand gestures started before the corresponding lexical affiliates with overlaps (**Finding 6**). However, there are significant differences in the integration time intervals of speech and hand gestures for males and females. On average, the integration time interval is longer for males than for females (**Finding 7**). This may imply that males need longer time to integrate speech and co-occurring hand gestures than females. If adaptive processing strategies are used for the integration of speech and hand gesture channel for different gender groups, the overall performance of the system can be potentially improved.

3.6 Post Analysis of Cognitive Actions

3.6.1 Correlation of Different Cognitive Actions

For the cognitive action codings, we used the same video/audio clips for speech and hand gestures analysis. The coding results are given in Table 3.30 and Table 3.31. In these two tables, G, F and P represent the gestures (G that corresponds to M-actions in the coding scheme in Table 3.4), F-actions and P-actions respectively.

For example, from the first row of the Table 3.30 we can see that a female participant had 29 gestures, 34 F-actions and 27 P-actions in Task 1. Her description of the object was divided into 8 segments. On average, she had 11.25 actions per segment. In total, we annotated 136 gestures, 165 F-actions and 128 P-actions in task 1 for females. For males in task 1, 93 gestures were annotated, 78 F-actions and 101 P-actions were coded. In task 2 the corresponding number of gestures, F-actions and P-actions are 124, 139 and 118 for females and 105, 87 and 117 for males. In total, we obtained 260 gestures, 304 F-actions and 246 P-actions for females and 198 gestures, 165 F-actions and 218 P-actions for males.

	Task 1					Task 2					Total				
P	G1	F1	P1	Seg1	Ave	G2	F2	P2	Seg2	Ave	G	F	P	Seg	Ave
P1	29	34	27	8	11.25	30	35	25	6	15	59	59	62	14	12.85
P2	30	35	34	12	8.25	28	25	20	9.13	8	58	55	59	20	8.6
P3	8	9	6	3	7.67	8	7	6	2	10.5	16	15	13	7	6.28
P4	15	9	13	3	12.33	11	12	10	4	8.25	26	19	25	7	10
P5	14	15	13	4	10.5	15	15	14	4	11	29	29	28	8	10.75
P6	10	17	12	4	9.75	10	12	11	5	6.6	20	28	24	9	8
P7	9	17	8	3	11.33	8	10	7	3	8.33	17	24	18	6	9.83
P8	10	22	12	5	8.8	10	18	19	5	9.4	20	41	30	10	9.1
P9	11	7	3	3	7	4	5	6	2	7.5	15	13	8	5	7.2
Ave	15.11	18.33	14.22	5	9.53	13.78	15.44	13.11	4.33	9.77	28.89	31.44	29.67	9.56	9.42

Table 3.30: Summary of the number of cognitive actions for females

	Task 1					Task 2					Total				
P	G1	F1	P1	Seg1	Ave	G2	F2	P2	Seg2	Ave	G	F	P	Seg	Ave
P1	15	20	14	4	12.25	15	10	15	4	10	30	30	29	8	11.25
P2	6	5	6	3	5.67	5	4	3	2	6	11	9	9	5	5.8
P3	6	5	8	2	9.5	4	8	10	3	7.33	10	13	18	5	8.2
P4	10	9	13	3	10.67	16	14	18	6	8	26	23	31	9	8.89
P5	13	10	14	3	12.33	22	17	21	7	8.57	35	27	35	10	9.7
P6	13	8	12	3	11	11	12	16	4	9.75	24	20	28	7	10.28
P7	13	4	9	2	13	16	8	11	3	11.67	29	12	20	5	12.2
P8	10	10	15	4	8.75	13	9	13	3	11.67	23	19	28	7	10
P9	7	7	10	3	8	3	5	10	2	9	10	12	20	5	8.4
Ave	10.33	8.67	11.22	3	10.07	11.67	9.67	13	3.78	9.08	22	18.33	24.22	6.78	9.52

Table 3.31: Summary of the number of cognitive actions for males

Before investigating the differences between the cognitive actions of males and females, we studied the correlation between the occurrences of G-actions, F-actions and P-actions in the same task for both male and female participants.

Table 3.32 and Table 3.33 display the correlation analysis of the occurrences of different types of cognitive actions for males and females. As seen in the tables, generally, there are high correlations between different types of cognitive actions.

Table 3.32(a) and Table 3.32(b) give the results of the correlation coefficients of the occurrences of different cognitive actions for males in task 1 and task 2 separately. Even though the correlations of different cognitive actions in task 1 (0.61 for G-actions and F-actions; 0.65 for G-actions and P-actions; 0.71 for F-actions and P-actions) are slightly weaker than that in task 2 (0.82 for G-actions and F-actions; 0.79 for G-actions and P-actions; 0.95 for F-actions and P-actions) for males, their correlations are still strong on average as shown in Table 3.32(c). Generally, the weakest correlation is seen between G-actions and F-actions (0.77) and the strongest correlation is seen between F-actions and P-actions (0.88) for males.

(a) task 1

	G1	F1	P1
G1	1.00	0.61	0.65
F1	0.61	1.00	0.71
P1	0.65	0.71	1.00

(b) task 2

	G2	F2	P2
G2	1.00	0.82	0.79
F2	0.82	1.00	0.95
P2	0.79	0.95	1.00

(c) total

	G2	F2	P2
G2	1.00	0.77	0.79
F2	0.77	1.00	0.88
P2	0.79	0.88	1.00

Table 3.32: Correlation coefficients of different cognitive actions for males

Table 3.33(a) and Table 3.33(b) show the correlation analysis for females' cognitive action occurrences in task 1 (0.83 for G-actions and F-actions; 0.94 for G-actions and

P-actions; 0.91 for F-actions and P-actions) and task 2 (0.94 for G-actions and F-actions; 0.86 for G-actions and P-actions; 0.96 for F-actions and P-actions). In the two tables we can see different actions are highly correlated with each other in females. On average, the correlation coefficients for different cognitive actions are all greater than 0.9 (0.91 for G-actions and F-actions; 0.95 for G-actions and P-actions; 0.97 for F-actions and P-actions).

(a) task1

	G1	F1	P1
G1	1.00	0.83	0.94
F1	0.83	1.00	0.91
P1	0.94	0.91	1.00

(b) task 2

	G2	F2	P2
G2	1.00	0.94	0.86
F2	0.94	1.00	0.96
P2	0.86	0.96	1.00

(c) total

	G2	F2	P2
G2	1.00	0.91	0.95
F2	0.91	1.00	0.97
P2	0.95	0.97	1.00

TABLE 3.33: Correlation coefficients of different cognitive actions for females

We can see from Table 3.32 and Table 3.33 that generally different groups of cognitive actions are highly correlated with each other for both male and female participants. However, higher correlations between the three different types of cognitive actions are observed in females not only on average as a whole, but also in both tasks. The relative weak correlations are observed for males in task 1.

3.6.2 Differences in Occurrences of Cognitive Actions

1. In overall speech and hand gestures

In Table 3.33 and Table 3.32 we observe stronger correlations between the occurrences of different cognitive actions in females. In addition to the differences in the strength

of correlations, we can also see the traces for different patterns of cognitive processing used by male and females.

In Table 3.30 of results for females, we can see that most of the female participants have more F-actions than P-actions. When seeing the results in Table 3.31 for males, most of them have more P-actions than F-actions. In total, still more F-actions than P-actions were observed by most females (8 out of 9) and more P-actions than F-actions in males (8 out of 9). On average, females have 31.44 F-actions that is more than 29.76 P-actions, while males have 18.33 F-actions that is less then 24.2 P-actions. The occurrences of G-actions are between F-actions and P-actions for both females (28.89) and males (22.0). In general, we can conclude **the differences in the occurrences of cognitive actions used by males and females as that P > G > F for males and F > G > P for females (Finding 8)**.

As for the occurrences of different actions of males and females, we can see on average females presented more cognitive actions than males in both task 1 (females: G(15.11), F(18.33), P(14.22); males: G(10.33), F(8.67), P(11.22)) and task 2 (females: G(13.8), F(15.4), P(13.1); males: G(11.67), F(9.67), P(13)). As a whole, the same pattern was observed (females: G(28.89), F(31.44), P(29.67); males: G(22.0), F(18.33), P(24.22)). As presented in Section 3.5.1, generally males and females spend similar amount of time in each task, but **females' protocols include more cognitive actions than males for the same task in our experiments.** This might be an implication of that **females give more attention to details on different parts of the objects compared to males (Finding 9)**.

Another interesting difference we found from Table 3.33 and Table 3.32 are the segment numbers we obtained during cognitive coding. On an average, more segments were extracted from females' data clips. As explained in Section 3.2.1, a new segment indicates a change in the speaker's intention or the contents of their thoughts. More segments found for female participants may indicate that females alter their attention on different parts of the chair more frequently than males. For the average number of cognitive actions (including G, F and P) per segment, females and males do not show significant differences (female: task 1 (9.53), task 2(9.77), total (9.42); male: task 1

(10.07), task 2 (9.08), total (9.52)).

2. In spoken words accompanying gestures

We found that females use more nouns than adjectives to accompany hand gestures, while males use more adjectives. We list the six types of keywords used with hand gestures again: adjective, noun, numerical info, adverb, preposition and verb.

Generally, nouns are used to describe the parts of the chairs, such as legs, back, base, bars, frame, etc., which associate with functional cognitive actions in the coding scheme (see Table 3.7). Adjectives represent the features of these parts, such as round, rectangular, square, heart-shape and curved etc., which can be coded as perceptual cognitive actions (see Table 3.6). Numerical information is also regarded as functional actions as defined in the coding scheme [1, 60]. Prepositions and adverbs usually are in relation to the position or alignment of parts of chairs, which are in line with perceptual actions. Only a few verbs were detected (e.g. cross, locate, raise, intersect, etc.) also for the alignment description that correspond to perceptual actions.

We can assume that the number of all nouns and numerical information may be consistent with the number of F-acitons, while the number of all adjectives, prepositions, adverbs and verbs may be in line with P-actions. We classify the spoken words accompanying hand gestures into two classes: "f" that represents the F-actions related words and "p" that represents the P-actions related words. Fig. 3.29 shows the distribution of these two classes in males and females. In this figure, we can also observe that males have more perceptual actions while females present more functional actions generally.

Reviewing the definition of perceptual and functional actions in Table 3.4, we can see perceptual actions are more associated with shapes, sizes or alignment of objects (using type "p" words), which are observed more frequently in males, functional actions are more associated with the structural description of objects (using type "f" words) in our case that dominate over others in females. This might be an implication that **different cognitive processing patterns may be the reason for the differences**

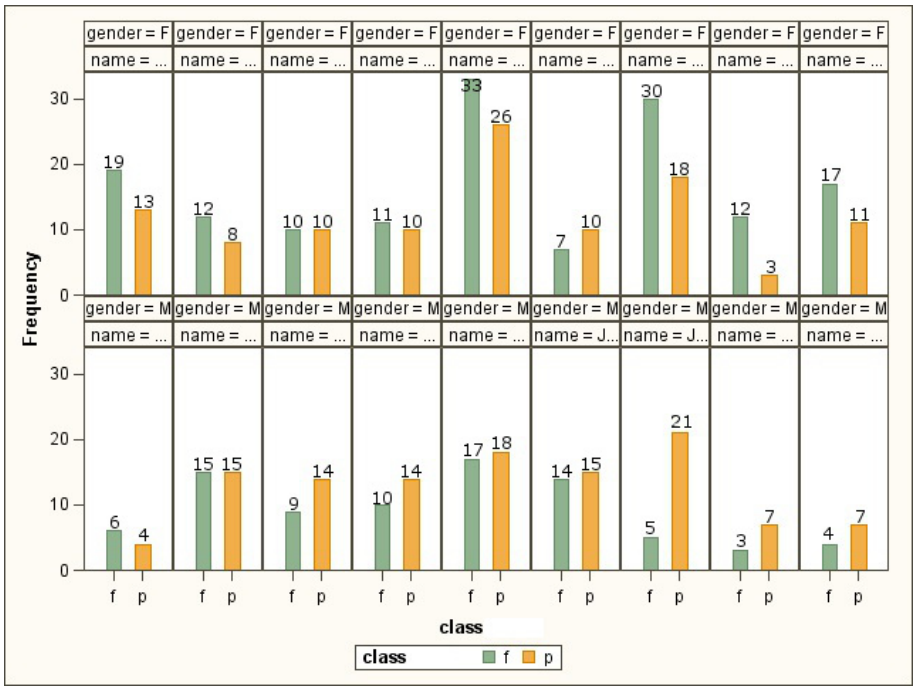


Figure 3.29: Distribution of two classified types of spoken words

in the distribution of word types used with hand gestures for males and females (Finding 10).

3.6.3 Evaluation of Results on Cognitive Processing

In Section 3.4 of this chapter, we demonstrated gender differences in the presentation of speech and hand gestures, such as they have different preferences in using speech and hand gestures and in corresponding lexical affiliates of gestures, females use more nouns while males use more adjectives. We also investigate gender differences in cognitive processing in this section, which may be the reason for the different presentation of speech and hand gestures in genders.

Based on the cognitive analysis, we found that males and females use different cognitive patterns when they describe the same objects. Generally, females have more functional actions than perceptual actions, while males have more perceptual actions than functional actions (**Finding 8**). On average, we also found that females protocol also include more cognitive actions than males (**Finding 9**).

Regarding the definition of each type of cognitive action, perceptual actions are more associated with visuospatial description of objects (using type "p" words) which are found superior in males, functional actions are more relevant to the structural description of objects (using "f" type words) which dominate over others in females. To put it another way, different cognitive processing patterns may be the reason for different distribution of word types used with hand gestures for males and females (**Finding 10**).

3.7 Conclusion

In this chapter, we explored the gender differences in use of speech and hand gestures internally (the cognitive processing) and externally (temporal alignment). We conducted an experiment to collect the multimodal data. In the experiment, we filmed participants when they described two objects using speech and hand gestures together. Gestures and the corresponding lexical affiliates were annotated from the collected video clips. Protocol analysis were also conducted to study the gender differences in the cognitive processing of speech and hand gestures. Based on Suwa's coding scheme, we coded cognitive actions (P-action, F-action and G-action (hand gestures)) from the video clips.

From the analysis of the data in our experiment, we found that

- for the same tasks, males and females spend similar amount of time to finish. However, females spend more absolute time on gestures than males regardless of the total task time (Finding 1)

- the preference of using speech and hand gestures in male and female is stable over the time from one task to another (Finding 2)

- the length of gesture strokes and the length of corresponding lexical affiliates are different for males and females (Finding 3 and Finding 5).

- regarding the properties of hand gesture related keywords, for most females,

nouns are the dominant part of all other types, while for most males adjectives are the majority (Finding 4)

- males and females have similar integration patterns, that is the majority of stroke phases of hand gestures start before the corresponding lexical affiliates with overlaps (Finding 6)
- on average, the integration time interval is longer for males than for females (Finding 7)
- females have more functional actions than perceptual actions, while males have more perceptual actions than functional actions (Finding 8)
- on average, females protocol also include more cognitive actions than males (Finding 9)
- different cognitive processing patterns may be the reason for different distribution of word types used with hand gestures for males and females (Finding 10)

In our experiment, we found that gestures preceded the related lexical affiliates both for males and females. This is consistent with the findings in other studies [39, 40, 41, 42, 43, 44, 45, 23, 46]. More than 90% gestures started before the related words within 2 seconds in our experiment. Some other studies claimed different time intervals between gestures and the spoken utterances (e.g. 1 to 2 seconds [40], 0 to 3.75s [39], within 4s [134]). While measuring the temporal synchrony of speech and co-occurring gestures, different measurement points (e.g. gestural onset (the start point of a gesture) to speech onset (the start point of the lexical affiliates), the apex of gestural stoke and the stressed point in related keywords) might be a an explanation for the different conclusions. Other potential factors affecting the results could be different gesture types used in experiment, different tasks selected for experiment and user-related factors (e.g. user gender, age, cultural background and other individual differences).

We found even the onset of the gesture stroke starts before the onset of the lexical affiliates within 2 seconds, but the time interval between them is longer for male than

for female. In future, more work may need to quantify the specific time interval of hand gestures and corresponding speech for males and females.

4

Experiment 2 and Analysis

As introduced in Chapter 3, the objectives of the thesis include studying gender differences in three aspects: 1) using speech and hand gestures, 2) cognitive processing, 3) brain activities associated with using speech and hand gestures. We have studied gender differences in the first two aspects in Chapter 3. In this chapter, we will introduce the methodology used and the experiment conducted for the third purpose. The procedures of this experiment and the evaluation of the experimental results will be given in detail.

4.1 Introduction to Emotiv Neuroheadset

The Emotiv EEG offers a high resolution, multi-channel and wireless portable EEG system. 14 EEG channel names based on the International 10-20 locations are[1]: AF3, F7, F3, FC5, T7, P7, O1, O2, P8, T8, FC6, F4, F8, AF4. The scalp locations covered by Emotiv Neuroheadset is shown in Fig. 4.1.

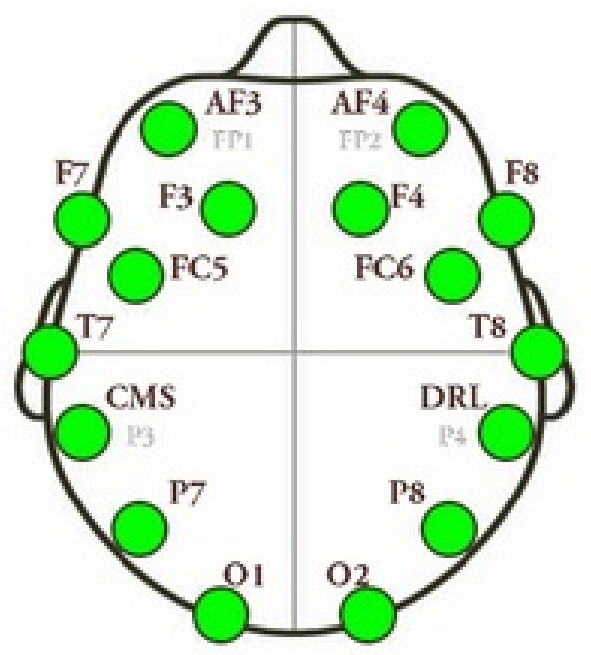


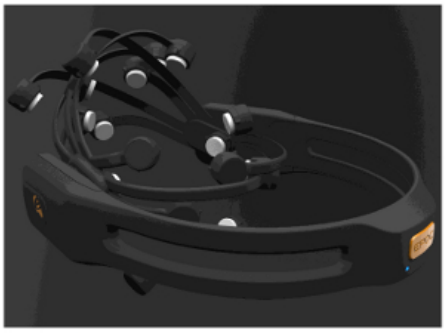
FIGURE 4.1: Scalp locations covered by Emotiv Neuroheadset

In this experiment, we collected EEG signals through Emotiv Neuroheadset. Some details about Emotiv Neuroheadset are displayed in Fig. 4.2. Emotiv EEG system has been the most prevalent low-cost EEG system which is totally affordable today for researchers in different areas. In a recent study different low-cost EEG systems were compared and rated by usability (see Fig. 4.3) [198]. In their findings, the EmotivEpoc scored best in terms of usability, but its price was only in the middle level among other

[1]http://www.emotiv.com

EEG systems. Other studies using Emotiv EEG system also provided us with reliable references, on the basis of which we selected it as our research tool.

One important element of this experiment is the saline solution used for filling the pads of the sensors. The key requirements are: saline content between 0.5% and 4%, preferably at the lower end to reduce salt build-up, non-allergenic, fitted with anti-microbials. Multi-purpose contact lens solution normally works well. The sensors contact the skin of participants directly, so every participant has to be tested with the saline solution to ensure that they are not allergic to it.



	EEG HEADSET
Number of channels	14 (plus CMS/DRL references, P3/P4 locations)
Channel names (International 10-20 locations)	AF3, F7, F3, FC5, T7, P7, O1, O2, P8, T8, FC6, F4, F8, AF4
Sampling method	Sequential sampling. Single ADC
Sampling rate	128 SPS (2048 Hz internal)
Resolution	14 bits 1 LSB = 0.51μV (16 bit ADC, 2 bits instrumental noise floor discarded)
Bandwidth	0.2 - 45Hz, digital notch filters at 50Hz and 60Hz
Filtering	Built in digital 5th order Sinc filter
Dynamic range (input referred)	8400μV (pp)
Coupling mode	AC coupled
Connectivity	Proprietary wireless, 2.4GHz band
Power	LiPoly
Battery life (typical)	12 hours
Impedance Measurement	Real-time contact quality using patented system

Figure 4.2: Details about Emotiv Neuroheadset from Emotiv EEG specifications

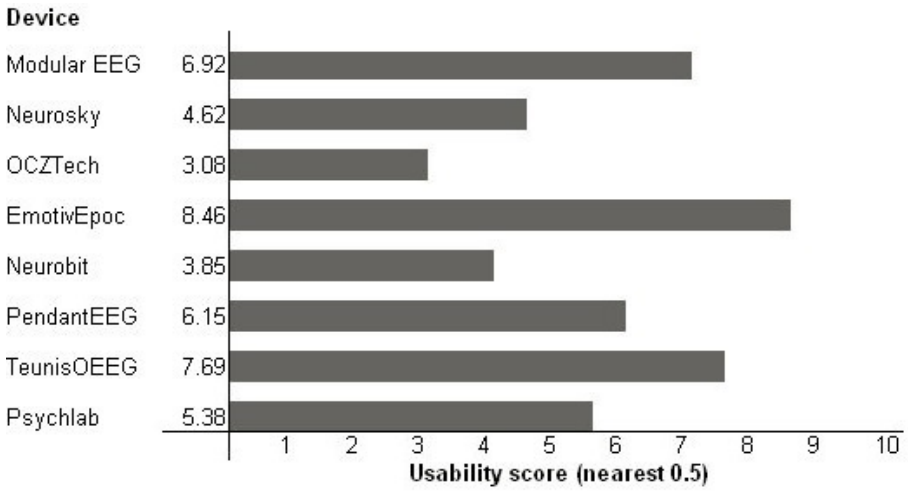


FIGURE 4.3: Usability rating of low-cost EEG devices

The Emotiv Neuroheadset can be wirelessly connected to PCs through a USB bluetooth receiver. A program named TestBench runs on a PC and independently collects data packets from the bluetooth receiver and processes them to display, analyse, record and play back time-dependent EEG signals. The TestBench application provides a user friendly interface for us to collect and view EEG signals. A screenshot of TestBench control panel is displayed in Fig. 4.4.

The left side of the TestBench Panel is known as the TestBench Status Pane. This pane shows sensor contact quality of Emotiv neuroheadset and the location of each sensor as well. There are four different levels of contact quality displayed in different colors in TestBench Status Pane (from best to worst: green, yellow, orange and red). If the sensor is totally disconnected, the color turns black. The sampling frequency and battery status displays under the status pane when the program runs. The right hand side of control panel reports real-time brain wave signals when the sensors are connected. It allows users to select single or multiple channels to be displayed.

There is an important item "Marker" under "Menu" on the top left hand side of the control panel. Markers are used to indicate specific events or areas of interest in EEG files as they are recording, so users can find the events later during playback or

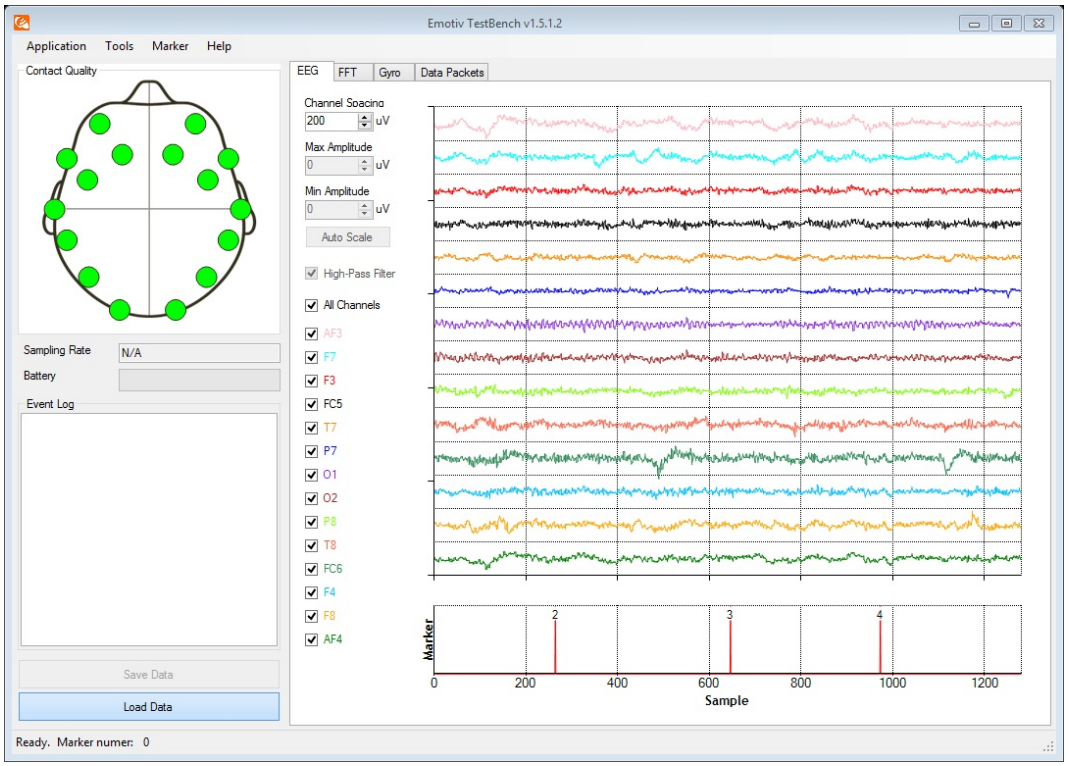


Figure 4.4: A screenshot of TestBench control panel showing all EEG channels

analysis. Marker can be set manually or connected to a serial port (or virtual serial port) to allow other applications to mark the events automatically. The markers will be saved together with EEG signals after the completion of each experimental section. One can view the markers by playing back the recorded EEG signals and see the markers displayed at the bottom right hand side as shown in Fig. 4.4.

Before we move to our experiment, it should be noted that EEG signals are quite sensitive to artifacts and noises. It is impossible to completely remove these artifacts and noises, but we can design an experiment to reduce them as much as possible, in order to extract the EEG signals we are interested in.

The most common artifacts that might affect the brain wave signals are the following:

- Electro-Oculographic (EOG) artifacts that arise from movement of the eye and blinks cause changes in the eye's electric fields. Blinks and eye-movements not

only show up in an EOG recording, but also affect the brain wave signals. Even though there is no sensor placed directly around eyes on Emotiv neuroheaset, some sensors (like AF3 and AF4) might be close to eyes when placed on the participant's head. To prevent this, we requested our participants to close their eyes when they were speaking using hand gestures to reduce this kind of effect.

- Electro-Myographic (EMG) artifacts are generated from muscle contractions which represent neuromuscular activities not brain activities. The EMG signals influence the EEG recordings. It is therefore important to weaken the influence of muscle contractions. We need participants to move their hands during the experiment, but we asked them to sit on a chair, relax the other parts of their body and try to move other parts of their body as little as possible, while they are moving their hands.

- Electrocardiogram (ECG) artifacts caused by the heart's electrical activity can also be visible in EEG recordings. ECG influence can be ignored, because no sensor is placed close to the heart or upon a pulsating blood vessel in our experiment. However we still requested the participants to relax to eliminate the potential noise.

4.2 Experiment 2

4.2.1 Participants for Experiment 2

Though the participants for this experiment were independent from the participants of Experiment 1, the inclusionary and exclusionary criteria for the participants were similar in terms of recruitment and the consent form (see Appendix A). For example, only people who can speak English fluently without any history of disorders in language, speech, hearing or development could participate in the EEG experiment. Two more requirements are particularly specified for the second experiment. One is that any participant should be right-handed, which is assumed to eliminate the possibility of the influence on the brain wave signals by participant's handedness. Another one is

the participant must not be sensitive to the saline solution used for Emotiv sensors. We used multi-purpose contact lens solution that is safe even for sensitive skin and also works well with the Emotiv headset. None of our participants dad any allergic reaction to the multi-purpose contact lens solution.

Finally, fourteen participants (7 males and 7 females) participated in our experiment. None of them participated in the first experiment.

4.2.2 Task and Data Collection

In the first experiment, we used the video/audio clips that recorded participants' speech and hand gestures when they described two objects, and extracted the top 10 lexical affiliates of gestures used by all participants. These top 10 keywords were then used in the second experiment. The aim of the second experiment was to study the differences in brain activities of males and females who use speech and hand gestures together. Therefore the participants were required to speak the ten words obtained from the previous experiment and use gestures to accompany their speech spontaneously.

For these ten keywords, it is obvious that different participants may have different gesture styles. For example, when asked to express "round" by speech and hand gesture, one may use two fingers to draw a circle in the air, while another may hold two bent palms together with a hole in between to express "round". The individual differences in rendering hand gestures are quite common and natural. We did not restrict the gesture styles that each participant preferred to use, but we did ask all participants to use two hands for all gestures, since the movement of different hands (right or left) may cause significant differences in brain wave signals for the same person. Apparently, if one uses left hand and another uses right hand, the brain activities can be significantly different. The consistency of which hand(s) are used across all participants is crucial to study the differences in brain activities of genders.

Prior to the commencement of the experiment, the author showed each participant what has been done in the first experiment and also the pictures used previously. The author let them know the 10 keywords extracted from other participants' descriptions of the two objects in pictures. During the experiment, each participant was asked to sit

in a comfortable chair in front of a desk. The 10 keywords were printed out on ten A4 white pages and displayed vertically on a small display board right ahead each participant, so they can view these words clearly and easily. The PC hosting the TestBench application was placed next to the display board and the experimental conductor sat in front of the PC to guide the participant to complete the whole procedure. The arrangement of the devices used in this experiment is displayed in Fig. 4.5.

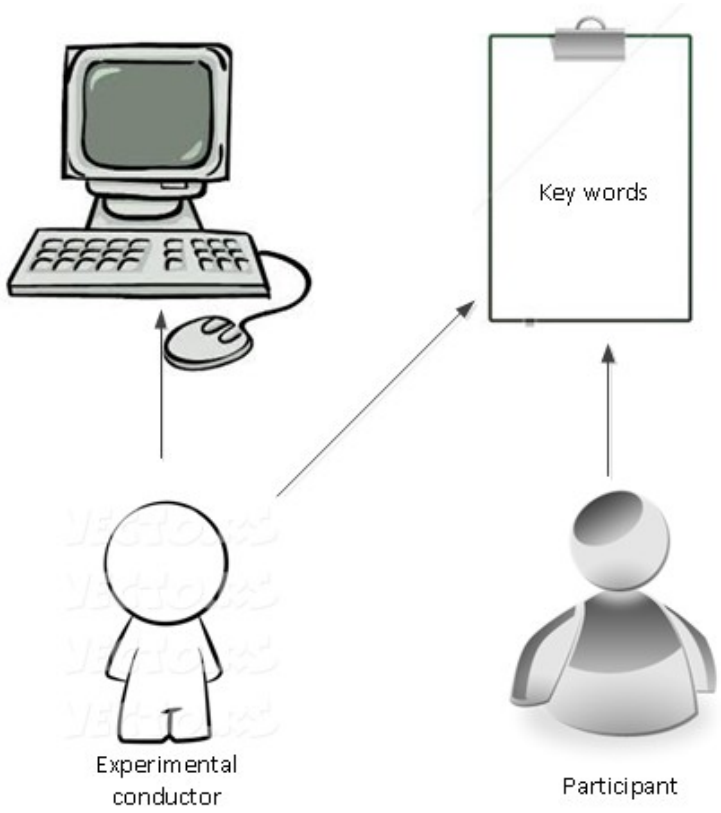


FIGURE 4.5: Arrangement of the devices used in experiment

As mentioned in the previous section, participants were requested to relax, close their eyes, only use hand gestures, as well as not to move (or move as little as possible) other parts of their body. The experimental conductor helped the participant put on the Emotiv headset at the beginning and sometimes adjust the sensors slightly to ensure best connection for each sensors. During the whole procedure, the participant sat on the right hand side of the conductor, while the conductor was in charge of two

things:

1. Mark each section (start point and end point) of participant's speech and hand gesture through the "Marker" menu of the TestBench control panel manually, which can be seen in Fig. 4.6. One section of data is defined for the completion of one hand gesture with the spoken keyword. For each participant, we obtained one set of data which consist of all sections for ten keywords.

2. Turn page for the participant. Since the 10 white pages that have the keywords printed on are put together on the display board.

In our experiments, we sent markers manually. Markers can be edited and saved prior to the start of experiment. From Fig. 4.6 we can see, each marker has two parameters: name and value. The name is almost meaningless for us because we will extract each section from the continuous record based on the marker value. We preset the value of markers from 1 to 11. "1" indicates the onset of the first section and "11" means the end of the last section. All values between "1" and "11" represent the end of the previous section and also the beginning of the following section.

The whole procedure of one set of data collection was conducted as follows:

First, the conductor opened the TestBench control panel and assisted the participant in putting on the Emotiv headset to make sure that all sensors were correctly connected and stable.

Second, the conductor opened the marker menu and meanwhile the participant took a look at the first keyword and then closed their eyes to relax.

Third, the conductor prepared to collect data and gave the participant an indication to commence by saying "start". At the same time the conductor clicked the "send" button next to the marker value "1" (see Fig. 4.6).

Fourth, the "send" button next to the marker value "2" was clicked, once the participant finished the first keyword with associated hand gesture. The participant was told to gesture slowly but naturally and end a gesture with "retract" phase to the greatest extent. Each participant repeated ten times one keyword fluently without any

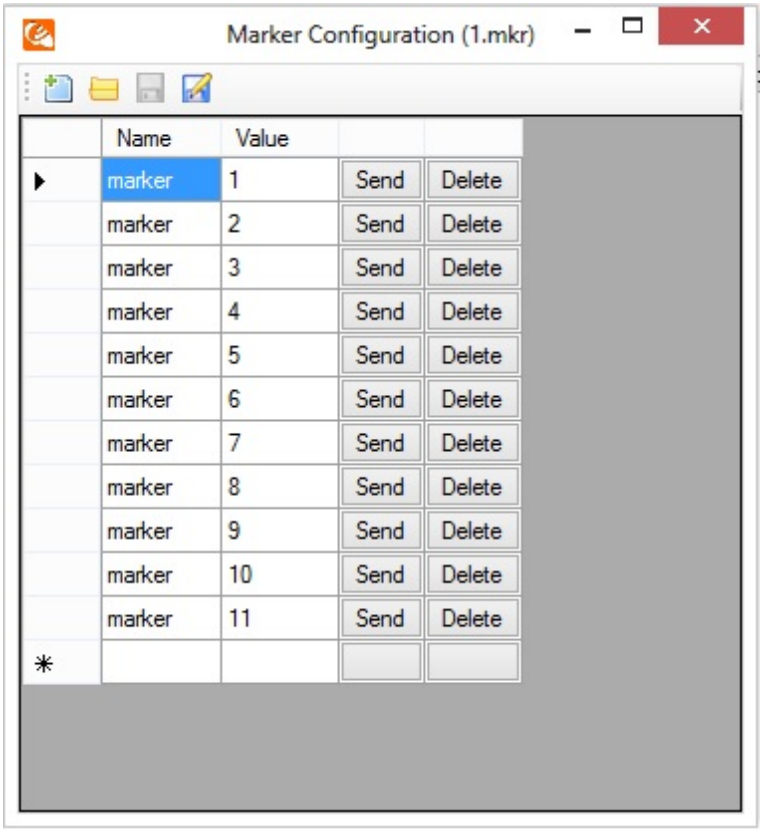


FIGURE 4.6: Marker menu of TestBench

breaks. The marker with value "2" indicates the end of the first section and also the start of the second section, by parity of reasoning for other markers. At the end, each participant produced 100 sections of data. The participant was requested not to count the number of sections as counting in mind can dramatically influence the EEG signals we receive.

Fifth, after the completion of the tenth section, the conductor clicked the marker "11" and said "OK" to inform the participant to stop. The participant then opened his/her eyes to look at the next keyword and replicated the foregoing procedures from the second to the fifth. The data collection continued during the transition from one keyword to another. We extracted sections corresponding to each keyword in post processing.

The data collected from TestBench are saved with ".edf" extension that is a special

format for EEG signals. They can be converted to CSV format (used in the post-processing) through "Launch EDF to CSV converter..." menu.

The final data set is analysed by the methods introduced in the next section and the results will be discussed at the end of this chapter.

4.3 Emotiv Headset Signal Processing

Even though Emotiv EEG system is still new, it has already been applied to different areas by some developers in recent years [199, 200, 201, 202, 203, 204]. A survey about the usability of different EEG acquisition devices shows that Emotive headset gets the highest score [198] among others.

The EEG is a dynamic noninvasive method to trace the state of the brain. An EEG signal is recorded with electrodes placed on the scalp and consists of different waves with different characteristics. The greatest advantage of EEG is that it is an instantaneous indicator of brain activities. EEG signal analysis involves using computational and mathematical tools to analyse and interpret the signals. Broadly speaking, EEG signal analysis can be classified into two categories: time domain methods and spectral methods [205].

Time domain methods include measurements of the raw signal characteristics. The commonly used time domain method is simply visual inspection that refers to observing amplitude distribution of EEG signals and its respective mean and variance [206]. Period or interval analysis refers to counting the number of incidences where the EEG crosses the zero voltage line [207]. Time domain techniques are usually fast and responsive to immediate signal analysis. These techniques are typically used to detect and decipher typical states of brain (e.g. sleep, fully anesthetized and burst suppression etc.) or abnormal brain injury or disease states (e.g. epileptic spikes) [205].

Spectral methods are by far the most prominent form of EEG analysis [208]. It takes EEG and normally converts it to a frequency domain (a power spectrum). The power spectrum can be subdivided into different frequency bands (delta, theta, alpha and beta) which can correlate with normal brain activities associated with bodily

functions. These frequency bands corresponding to different bodily functions will be briefly introduced in the next section. Due to the complexity and the nonstationarity of EEG signals, spectral analysis methods are widely used to analyse and interpret the trends in EEG, particularly in research. We adopt spectral analysis in the thesis and the procedure of our analysis will be introduced in the following sections.

4.3.1 Spectral Analysis of EEG

Traditional power spectrum analysis normally classifies EEG into the following four spectral bands [205]:

- Delta (0.5-4Hz): Delta waves are the lowest frequency component and are usually seen in the frontal lobe. They normally appear during sleep, in infancy or in serious organic brain diseases. They are seldom seen in healthy awake adults. Animals are known to have more widespread activity in delta range.

- Theta (4-7Hz): Theta waves happen mainly in parietal and temporal lobes of children's brains. They generally appear in healthy and awake adults only when they are drowsy or emotionally stressed (e.g. disappointed, frustrated) or at certain stages of sleep. The appearance of theta waves normally indicates the central inhibitory state of the brain.

- Alpha (8-13Hz): Alpha waves normally occur at a frequency between 8 and 13Hz. About 85% of adults demonstrate the frequency of alpha waves between 9.5 and 10.5. Alpha waves are the basic rhythmic waves in a healthy adult brain. Brain activity in this frequency range is distributed across the whole brain, but is often recorded from occipital and frontal regions symmetrically (see a diagram of brain areas in Fig. 4.7). Alpha waves are often seen with high amplitude in an awake but relaxed person and typically when their eyes are closed. When their eyes are opened or there is visual stimulation or the person has mental activities, alpha waves will be inhibited and replaced by beta waves quickly, as a normal brain activity.

- Beta (14-30Hz): Beta waves are found in the frontal and parietal areas and are affected by mental activity and tension. Basically beta waves are defined within 14-30Hz, but sometimes rise to 50Hz.

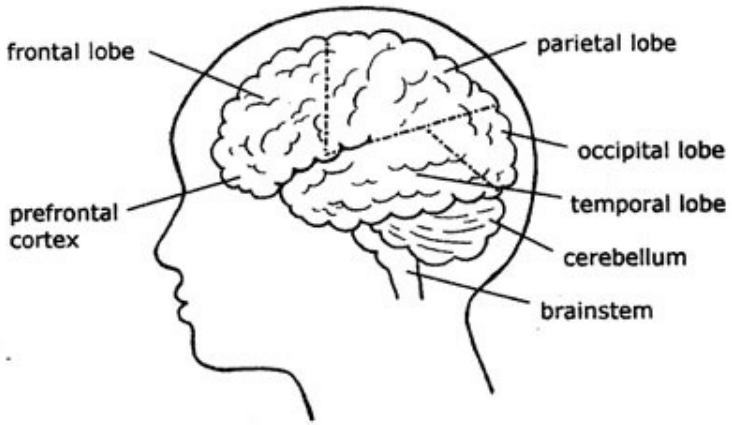


Figure 4.7: The brain areas

Human sensorimotor and cognitive behavior is associated with changes in the oscillatory activity of the brain. Specifically, motor activity is associated with changes in beta frequency oscillations, which has a range of 14-30 Hz and peaks at about 20 Hz [209]. In some articles, the beta frequency band is further classified into two subclasses: beta I (13-20Hz) and beta II (20-50Hz) [205]. Beta I, almost twice the alpha wave in frequency, accompanies the occurrence of alpha wave. Beta II occurs only during intense mental activity and tension. Generally alpha waves can be inhibited by visual stimulation or by voluntary/passive movement of the body. The appearance of alpha waves in the central area is also accompanied by beta waves in the frequency band 18∼26Hz. The energy change in EEG caused by the event related desynchronizing/synchroniztion (ERD/ERS) of beta rhythms appears with the sensorimotor performance in brain or the real movement of body.

The aim of our experiment is to study brain activities when speech and hand gestures are used in a relaxed situation. So alpha and beta frequency bands will be our target in the following analysis. We will use power spectrum analysis to calculate the

spectral power in alpha and beta frequency bands, since the changes in spectral power and phase can characterise the changes in the oscillatory dynamics of ongoing EEG [210].

The data analysis procedure for EEG signals collected from an Emotiv headset is shown in Fig. 4.8. The signal processing procedure starts with a baseline removal section. It removes the included DC offset in EEG signals since the data is transmitted as an unsigned integer, so that the values of the signal will be distributed around zero. Other steps in this procedure are detailed below.

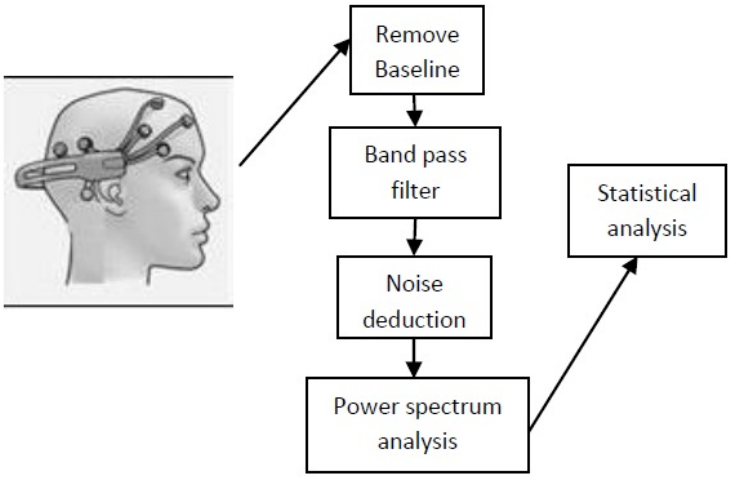


FIGURE 4.8: Data processing procedures for EEG signals

4.3.2 Band Pass Filter

The Emotiv Neuroheadset internally samples at a frequency 2048Hz which then gets downsampled to 128Hz approximately. The following preprocessing steps are also done in the hardware [211]:

- Low-pass filter with a cutoff at 45Hz to remove high frequency noise, since there are very few brain signals which are significant and distinguishable from mains interference above about 45Hz.

- High-pass filter with a cutoff at 0.16Hz to take the internal offset and any slow drift out of the signals.
- Notch filter at 50Hz and 60Hz to remove electrostatic noise.

The collected records from Emotiv Neuroheadset are then band pass filtered between 4Hz and 40Hz in order to remove artifacts related to higher frequencies and also the frequencies lower than 4Hz, since we are only interested in alpha and beta frequency bands.

4.3.3 Noise Reduction

Detecting and removing artifacts (muscle activity, eye blinks and electrical noise etc.) in EEG records is an important process in EEG signal research. Even though we can eliminate some high frequency noises by a band pass filter, there is noise in EEG signal frequency bands. We used Independent Component Analysis (ICA) to further clean the EEG signals after band pass filtering.

ICA was originally proposed to recover independent source signals from blind source signals [212]. Bell et al. proposed a simple neural network algorithm to carry out ICA that blindly separates mixtures of independent sources using information maximization (infomax) [213]. This method has been demonstrated to be suitable for performing blind source separation on EEG data [214]. ICA has been used as an effective method for eliminating artifacts and separating individual sources of brain signals from EEG recordings [215, 210].

The ICA algorithm is highly effective at performing artifact correction for EEG data by linear decomposition. This method is based on the assumptions that EEG signals recorded from the scalp [216]:

- are spatially stable mixtures of the activities of temporally independent cerebral and artifactual sources,
- the summation of potentials arising from different parts of the brain, scalp, and body is linear at the electrodes, and propagation delays from the sources to the

electrodes are negligible,

- the number of sources is no larger than the number of EEG electrodes (14 in Emotiv EEG headset).

In EEG signals, multichannel EEG records can be regarded as mixtures of underlying brain and artifactual signals. The first assumption is reasonable because the sources of eye and muscle activity, line noise, and cardiac signals are not generally time locked to the sources of EEG activity which is thought to reflect synaptic activity of cortical neurons. As volume conduction, which is commonly used to describe the transmission of electric or magnetic fields from an electric primary current source through biological tissue towards measurement sensors, is thought to be linear and instantaneous, the second assumption is satisfied. The third assumption is questionable, because we do not know the effective number of statistically independent signals contributing to the scalp EEG. However, numerical simulations have confirmed that this does not affect the effectiveness of applying ICA in EEG analysis [217].

The purpose of ICA is to linearly decompose multidimensional data vectors into statistically independent components. Given a set of observations of random variables (x_1(t), x_2(t),..., x_n(t)), where t is the time or sample index, n is the number of observed variables (n=14 in our case), they are assumed to be expressed as linear combination of independent components (s_1(t), s_2(t),..., s_n(t)) as follows:

$$\mathbf{x}(t) = \mathrm{A}\mathbf{s}(t) + \mathbf{v}(t)$$

where $\mathbf{A}$ is the $n*n$ mixing matrix, and $\mathbf{v}$(t) is the additive noise. After ICA analysis, the polluted $\mathbf{x}$(t) is estimated as clean signal $\mathbf{As}$(t), which is used in the power spectrum analysis in the next step.

4.3.4 Power Spectrum Analysis

The Fast Fourier Transform (FFT) was used to calculate the spectral power in the EEG frequency bands for alpha and beta waves. It should be noted here that each

participant spent different amounts of time on each trial in our experiment. That is to say the lengths of time or sample index vary from trial to trial and participant to participant as well.

Moving-average spectral analysis was applied to the recorded EEG signals with various lengths, in order to get smoother power spectrum curve. Each EEG record corresponding to each trial in the experiment was divided by a 128-point window with 64 point overlap and the 128-point windowed data was calculated by 256-point FFT to estimate its power spectrum, resulting in a frequency resolution near 0.5Hz in power spectrum density estimation. No record was less than 128 points in our case. For each record, the average spectral power of all 128-point windowed data was calculated to represent the spectral power of this record. We also normalised the average power spectrum of each record to a logarithmic scale, since the power of the EEG spectral amplitudes tend to change more linearly in a logarithmic scale than in the normal scale [210]. We then extracted the mean of power in the alpha and beta frequency bands as one of the power spectrum features.

We also use a feature in terms of the power spectral moments which has been used in some studies and proven to be a promising approach for EEG characterisation [218, 200]. The m^{th} order spectral moment is defined as

$$\mathbf{M}_m = \sum_{k=0}^{W} k^m P[k]$$

where W is the bandwidth of the spectrum. The implementation of spectral moments is accomplished through the normalization by low-order moments as illustrated in [200], since low-order moments are more stable to noise. The ratio of the second moment to the zero moment of each record is defined as below and used as another extracted power spectrum feature for each of the EEG band alpha and beta.

$$Alpha = |\frac{\sum_{k=8}^{13} k^2 P[k]}{\sum_{k=8}^{13} P[k]}|$$

$$Beta = |\frac{\sum_{k=14}^{30} k^2 P[k]}{\sum_{k=14}^{30} P[k]}|$$

The mean of the spectral power and the spectral moment of alpha and of beta bands are finally used to study the differences in brain activities of males and females who use speech and hand gestures together.

In order to study the differences in brain activities of males and females, we collected EEG signals by Emotiv Neuroheadset from 14 participants (7 females and 7 males). In the next section, we will discuss our experimental results.

4.4 Brain Activities Related with Speech and Hand Gestures

After we find differences in the use of speech and hand gestures, we are unavoidably concerned with the reasons for these differences. Inspired by the findings reported by other researchers [156, 157, 58] that gender differences in language processing is related to functional hemispheric brain asymmetry and female brain is less lateralised with functions spread over both hemispheres of the brain, we hypothesised that there are gender differences in the brain activities when using speech and hand gestures together. In this section we study the EEG signals captured by Emotiv Neuroheadset to validate hypothesis.

4.4.1 EEG Data Corpus

Fourteen (14) participants (7 females and 7 males) participated in our second experiment involving EEG signal collection. None of these 14 participant participated in the first experiment. Each participant was required to speak a number of keywords extracted from the first experiment using hand gestures while they were wearing Emotiv Neuroheadset (Fig. 4.2) with their eyes closed. In total, 10 keywords were used in the second experiment.

After the participant finished the repetition of each word as well as its corresponding hand gesture, a record of EEG signal was captured by TestBench and saved in .edf file, which can be converted to CSV format. For each participant, 10 records were captured

from TestBench. Within each record, 10 repeated sections were then extracted from CSV file through Matlab. In total for each participant, 100 EEG files were recorded including 1400 sections for all participants.

4.4.2 Spectral Moment Analysis

Each section of EEG signals was pre-processed by baseline remover, band pass filter and noise deduction (as illustrated in Fig. 4.8) before spectral analysis. As described in Section 3.3.3, we extracted the spectral moment as the feature to study the differences in brain activities of males and females. The implementation of spectral moment within alpha and beta frequency bands is defined as follows:

$$Alpha = |\frac{\sum_{k=8}^{13} k^2 P[k]}{\sum_{k=8}^{13} P[k]}|$$

$$Beta = |\frac{\sum_{k=14}^{30} k^2 P[k]}{\sum_{k=14}^{30} P[k]}|$$

P$[k]$ means the power spectrum at frequency k. For the 14 columns of one section EEG data (which comes from the 14 sensors of Emotiv headset), the moving-average FFT was applied to estimate its power spectrum. Fig. 4.9 displays a typical periodogram of estimated power spectrum. In order to smooth the periodogram of the power spectrum, we further averaged the estimated power spectrums of ten repeated sections for each keyword used by each participant. The averaged power spectral periodogram is given in Fig. 4.10 which will be used in spectral moment analysis in the following sections. As we can see, the periodogram is much smoother than it is in Fig. 4.9.

First, we extracted the spectral moment in alpha and beta frequency bands as well as the spectral moment change from alpha band to beta band (Beta - Alpha) for one section of data in the averaged power spectrum.

The top two figures in Fig. 4.11 give the spectral moment curves of 14 channels in alpha and beta frequency bands of two section data collected from one female and one

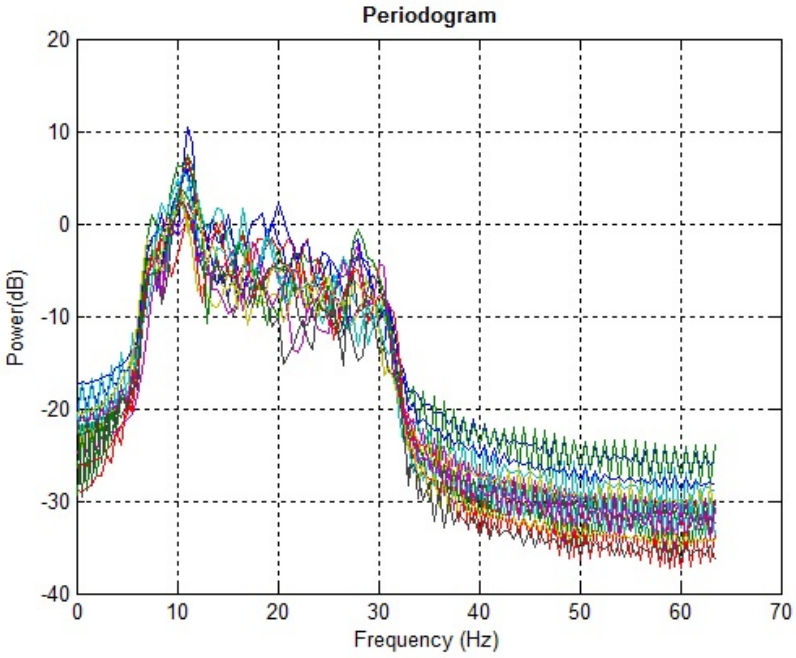


FIGURE 4.9: Power spectral estimation of one section EEG signal of one participant

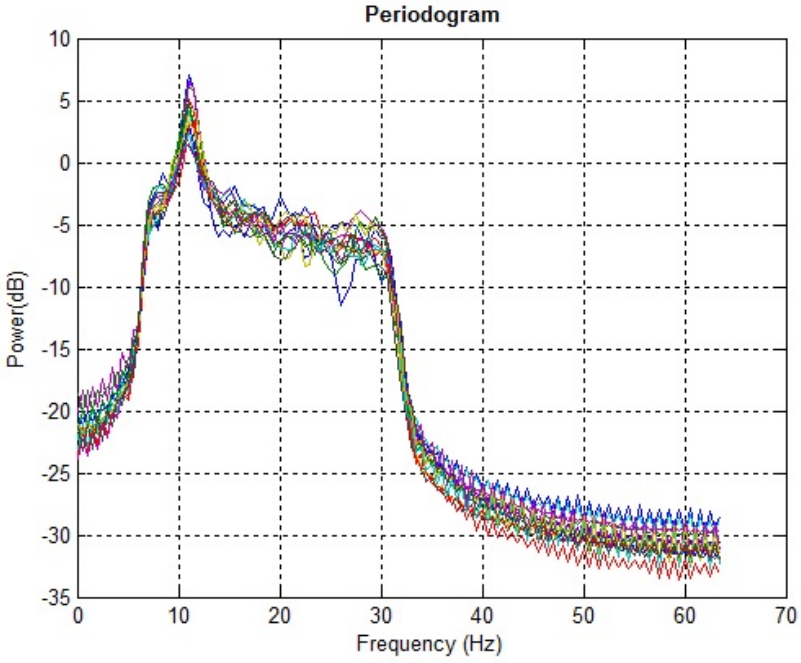


FIGURE 4.10: Average Power spectral estimation

male participant respectively, but for the same keyword. Both male and female participants spoke the same keywords using their hand gestures during the data collection. The bottom figure in Fig. 4.11 shows the spectral moment change from alpha to beta band of 14 channels for the selected example.

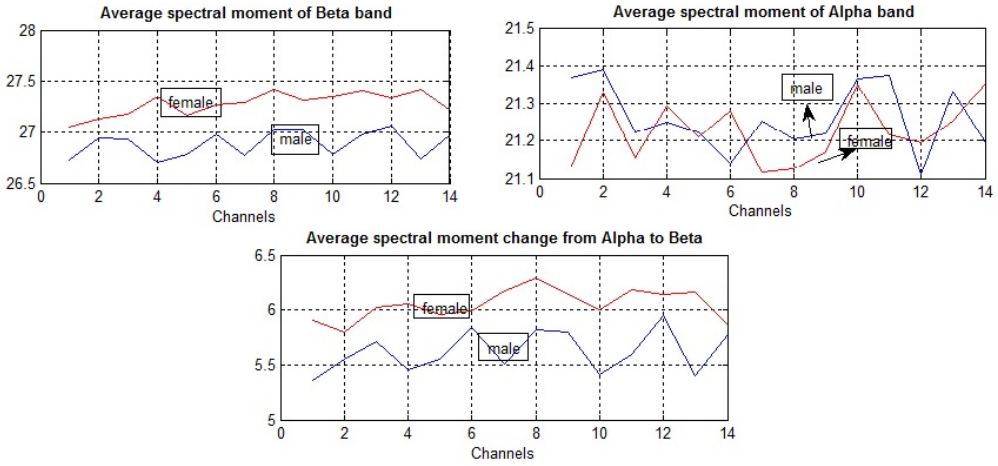


Figure 4.11: An example of spectral moment in different frequency band and the change between alpha and beta band

We can see clearly in this example that spectral moments in beta band are higher for female participants than male participants for each corresponding channel. However, we cannot see the consistency in alpha spectral moment. For some channels, it is higher for females, while for some other channels, it is higher for males. When looking at the bottom figure about the changes from alpha to beta bands, we can see the similar phenomena happened as for beta frequency band. The spectral moment change for females is always greater than for males in each corresponding channel.

In Fig. 4.11 we observe that the average spectral moment of all channels in beta frequency band is higher for females than for males. The average change from alpha to beta frequency band is also higher for females.

We illustrate the average spectral moment in beta frequency band for all 10 keywords for all participants in Fig. 4.12. We can see the values of average spectral moments across all data for males and females are quite close and range from 26.5 to 27.5. However, we also observe that the average spectral moment waves for females

are always located above the waves for males except for one participant. For most of the participants, the range of spectral moments in beta frequency for females is higher than for males.

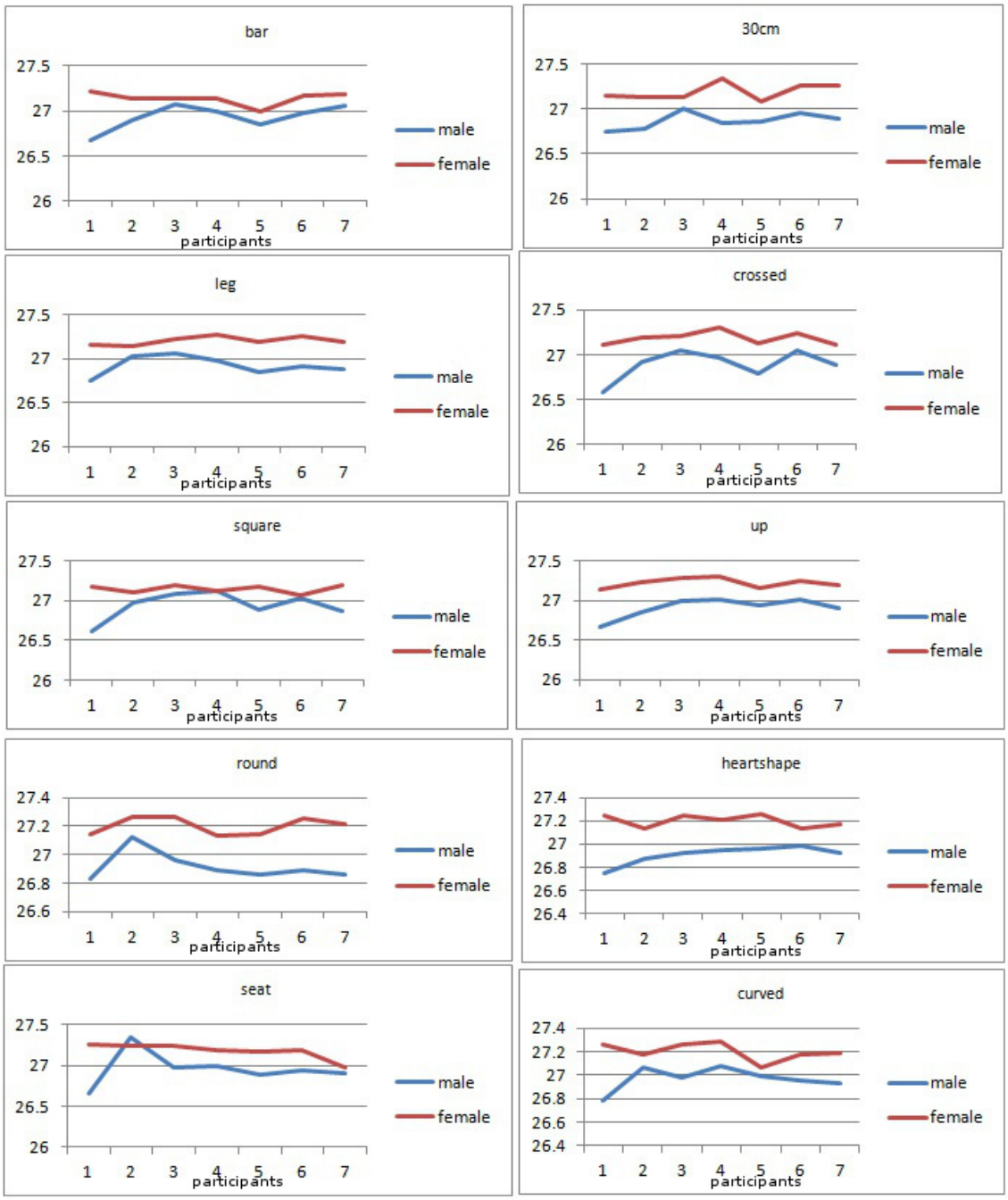


FIGURE 4.12: Spectral moment in beta frequency band for males and females

Fig. 4.13 shows the changes from alpha to beta frequency band for all 10 keywords for all participants. Even though, in alpha frequency band, the average spectral moment waves of males and females for some sections intersect with each other (as in Fig. 4.11), the spectral moment changes (Beta-Alpha) for females are still greater than males.

Generally, alpha waves are the basic rhythmic waves in healthy adult brain waves and are often seen when they are awake, relaxed and typically the eyes are closed [205], which is the initial status of our participants. It has also been stated in some studies that the appearance of beta waves inhibit the alpha waves and beta waves happen with sensorimotor activities in brain or the real movement of body [205]. The initial results from the spectral moment analysis in our experiment show that females and males with eyes closed in relaxed status do not present significant differences in brain activities, since there is no significant differences in the averaged spectral moment in alpha band. **However, when they use speech and hand gestures coordinated together, we observe that beta spectral moment waves are stronger in females and the changes of spectral moment from alpha to beta bands are more significant for females (Finding 11)**.

The significant spectral moment in brain waves may imply faster brain activities for females when use speech and hand gestures coordination, which may be the reason for shorter integration time of speech and hand gestures for females.

4.4.3 Analysis of Balance in Brain Activities

In the previous section, we found that generally beta wave spectral moments and the changes of spectral moment are stronger in female brain. As introduced before, female brains were found to be less lateralised than male brains in language processing with functions spread over both hemispheres of their brains. This finding leads us to consider whether the female brain is less lateralised than the male brain when speech accompanies hand gestures. In this section we study the gender differences in spectral moment of left and right hemispheres of the brain, related to speech and hand gestures.

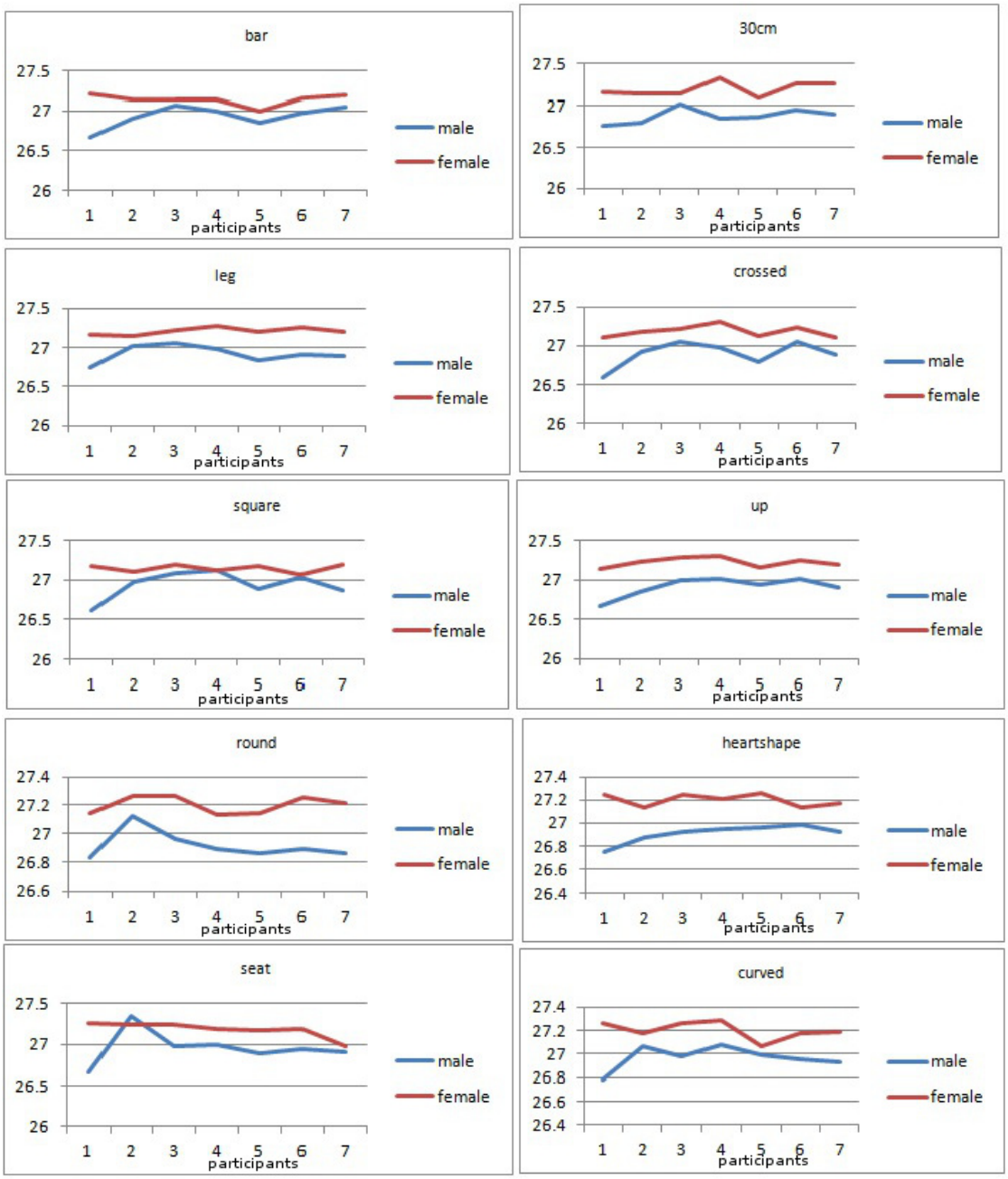


FIGURE 4.13: Spectral moment change from alpha to beta frequency band for males and females

As we know, the energy change in beta rhythm may appear with the sensorimotor activities in brain or the real movement of body [205]. We first checked beta spectral moment in both hemispheres of the brain of males and females. The beta spectral moments were calculated for the signals from both hemispheres of brain separately.

We used the mean value of beta spectral moments in all 7 channels to represent the beta spectral moment in each hemisphere. The differences of beta spectral moment in both hemispheres may not always have the same tendency. That is to say, for some records, beta spectral moments were stronger in the left hemisphere, while others may be weaker. Then, the absolute values of the differences in beta spectral moment of both hemispheres were computed for males and females. Fig. 4.14 gives the absolute differences of beta spectral moment in left and right hemispheres of the brain for males and females.

As seen in Fig. 4.14, the differences in beta spectral moment of left and right hemispheres of the brain are quite small and not consistent within gender groups. The values across all cases and participants fluctuate between 0 and 0.3. If brain activities present different lateralization patterns in speech and hand gesture coordination for males and females as demonstrated in language processing [156, 157, 58], the spectral moment in beta or alpha frequency bands should be different in left and right hemispheres of the brain for males and females. **Inconsistencies and small differences in beta spectral moment waves in left and right hemispheres of the brain for two gender groups imply that brain activities may be balancing the speech and hand gesture coordination (using two hands in our experiment) (Finding 12)**.

We also applied the two independent samples t-Test to the 10 cases across all participants. The results show no significant statistical differences in beta spectral moments of the two hemispheres of brain for males and females (p=0.12, 0.79. 0.85, 0.59, 0.63, 0.08, 0.18, 0.83, 0.82, 0.24). Similar results were also obtained in alpha frequency bands and the changes from alpha to beta band.

4.4.4 Evaluation of Results on EEG Analysis

As a summary, our experimental results indicate no significant differences in lateralisation in brain activities associated with speech and hand gestures between males and females (**Finding 11**). However, we found that beta spectral moment waves are

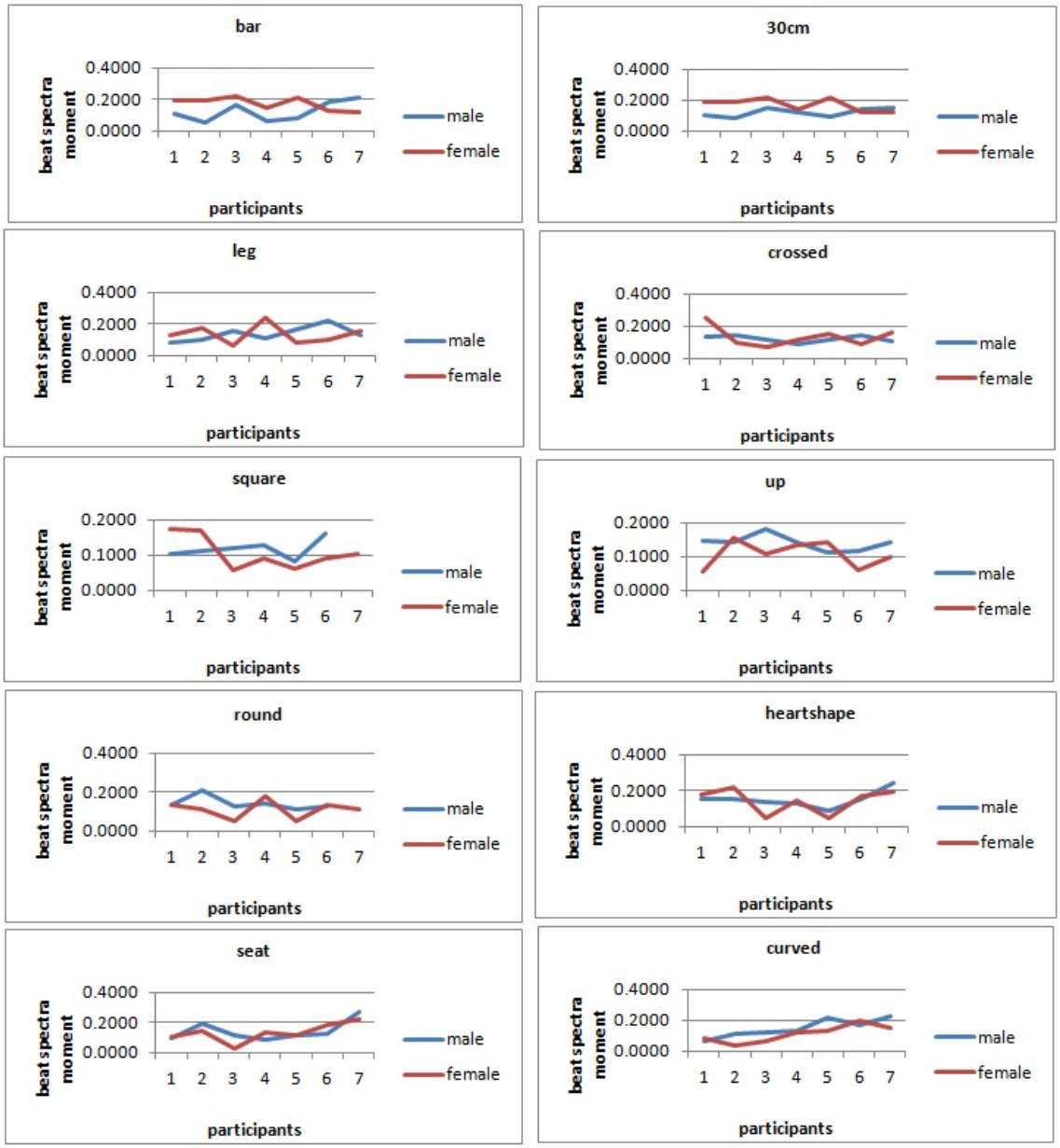


FIGURE 4.14: Spectral moment differences between left and right hemispheres of brain in beta frequency band for males and females

stronger in females and the changes of spectral moment from alpha to beta bands are more significant for females as well when they use speech and hand gestures together (**Finding 12**).

In Multimodal Resource Theory [49], cognitive resources have limited capacity and are shared by multiple tasks being completed at once (e.g. speech and gestures). Task

interference occurs when two concurrent tasks requiring the same resource compete or interfere with one another. When speech and hand gestures happen together, there might be some resources required both by verbal and motor systems in brain, which may eliminate the lateralisation that is observed in language processing.

In cognitive analysis, we found that normally females have more cognitive actions for same tasks than males. Females give more attention to details on different parts of the objects compared to males. More cognitive actions may indicate more frequent brain activities, which can cause strong brain waves with significant changes. The significant spectral moment in the brain for females may also imply faster brain activities associated with speech and hand gestures, which may be the reason for shorter integration time of speech and hand gestures for females.

4.5 Conclusion

In this chapter, we conducted another experiment to collect EEG signals through Emotiv Neuroheadset, in order to study gender differences in brain activities associated with using speech and hand gestures. Participants in the experiment were requested to use speech and hand gestures to describe 10 words when they sit on a chair in relax status with eyes closed.

Power spectrum analysis were applied for the collected data. The experimental results show that

- no significant differences in lateralisation in brain activities associated with speech and hand gestures between males and females. (Finding 11)

- beta spectral moment waves were stronger in female and the changes of spectral moment from alpha to beta bands were more significant for females as well when they used speech and hand gestures together (Finding 12)

Regarding the brain activities associated with speech and hand gestures, the findings from this experiment do not provide any evidence to support the lateralisation

hypothesis for males as found by others in phonological tasks [157] or in language processing [158].

However, study about the activated brain areas associated with different tasks is still controversial. Some other studies suggest males and females show different brain activation strength linked to word generation [58] and in mental rotation task [160]. We obtained some evidence from our experiment that there are gender differences in brain activation strength when speech coordinates with hand gestures.

We found stronger beta spectral moment waves in female when only using speech and hand gestures. As beta are affected by mental activity and tension, stronger beta spectral moment may indicate stronger brain activation strength. We also found female produced more cognitive actions in the tasks involving speech and hand gestures in Chapter 3, which may be linked to the stronger brain activation. As we observed shorter integration time of speech and hand gestures in female in Chapter 3, we suggested stronger brain activation may account for the shorter time interval.

5

Gender Prediction Modeling

5.1 Introduction

As introduced in Chapter 1, we are exploring three aspects in gender differences through our experiments. The first two aspects were studied in the first experiment and the third one was covered in the second experiment. The experimental results in Chapter 3 and Chapter 4 demonstrate that there are gender differences in the speech and hand gestures. The ultimate goal of the study in this thesis is to benefit MMIS. If the MMIS can recognise gender based on gender differences in their actions, the user will have more immersive experience and the system can achieve better performance. In this chapter, we will explore the possibility to predict gender based on their differences in the presentation of speech and hand gestures.

In the previous chapters we demonstrated findings from the experiments and evaluated results. Our findings suggest that there are gender differences in speech and hand gestures both internally and externally. Internal gender differences in cognitive processing and brain activities can be regarded as the intrinsic reasons for the external differences in the presentation of speech and hand gestures. These external differences can potentially affect the performance of MMIS using speech and hand gestures as input. The performance of a MMIS can be potentially improved if adaptive processing strategies are used for different gender groups. How to identify the gender groups is another issue that needs to be explored. In this section, we describe an attempt to build models that can predict gender based on their differences in the presentation of speech and hand gestures. It is difficult and even impossible to make a machine to analyse the internal differences, like cognitive processing at the moment.

To the best of our knowledge, there are only a few studies about gender prediction using speech and hand gestures. Some studies explore modeling methods to recognise gender using their speech. The modeling methods used in these articles are complex, like combining more than three different approaches [34, 35, 151, 152, 153]. In this thesis, we explore if gender can be predicted from speech and hand gestures using a simple but effective approach. The modelling methods we adopted in this thesis include logistic regression, decision tree, and neutral network. An introduction to them will be provided in the following section.

5.2 Gender Prediction Methods

The general hypothesis in this thesis is that there are gender differences in speech and hand gestures internally and externally. It may be possible to use differences in multimodal actions to predict gender. In this thesis we rely on the statistical analysis to investigate gender differences for gender prediction. In a broad sense, models are built from historical event records and are used to predict future occurrences of these events. If the multimodal actions can be predicted before being integrated together in MMIS, it will vastly improve the performance of MMIS by adopting the different integration

strategies for male and female users respectively. In this section we will introduce three statistical modeling methods (decision tree, neural network and logistic regression) we used to predict gender. We will compare the results of the three modeling approaches in the next section.

5.2.1 Decision Tree

Decision trees [219, 220] are a simple, but powerful tool for multiple variable analysis. They possess unique capabilities to supplement and complement for traditional statistical forms of analysis and a variety of data mining tools. The appeal of decision trees lies in their relative power, ease of use, robustness with a variety of data and levels of measurement, and ease of interpretability.

Decision trees are produced by split-search algorithms that identify various ways of splitting a data set into branch-like segments. These segments are organised as an inverted tree structure (a root node at the top of the tree.) The object of analysis is reflected in this root node as a simple, one-dimensional display in the decision tree interface. The name of the field of data that is the object of analysis is usually displayed, along with the spread or distribution of the values that are contained in that field. Decision tree can reflect both a continuous and categorical object of analysis.

For simplicity, we assume a binary target here, but the algorithm for interval targets is similar.

The first part of the algorithm is called split search. This split search starts by selecting an input for partitioning the available data. If the measurement scale of the selected input is an interval, each unique value serves as a potential split point for the data. If the input is categorical, the distinct value of each categorical input level is assigned to either side of the child nodes as a potential tried split.

For a selected input and fixed split point, two groups are generated. Cases with input values less than the split point are called branch left. Likewise, cases with input values greater than the split point are called branch right. The groups, combined with the target variable, form a 2×2 contingency table. A Pearson chi-squared statistic is adopted to quantify the independence of counts in columns. Large values for the

chi-squared statistic suggest that the proportion of zeros and ones in the left branch is different from the proportion in the right branch, indicating a good split.

This split search continues within each leaf and the data is partitioned according to the best split, which creates a second partition rule. The process repeats in each leaf until there are no more splits unable to satisfy termination conditions.

A sample decision tree is illustrated in Fig. 5.1. The top node reflects the data set variables or fields in the analysis. The bottom nodes represent a best-split according to the algorithm described above. Splitting rules are applied one after another, resulting in a hierarchy of branches nested with branches underneath. For each leaf, the decision rule provides a unique path for data to enter the class that is defined as the leaf. It is worth noting that all nodes, including the bottom leaf nodes, have mutually exclusive splitting rules. Therefore, each record or observation from the parent falls into one child node only.

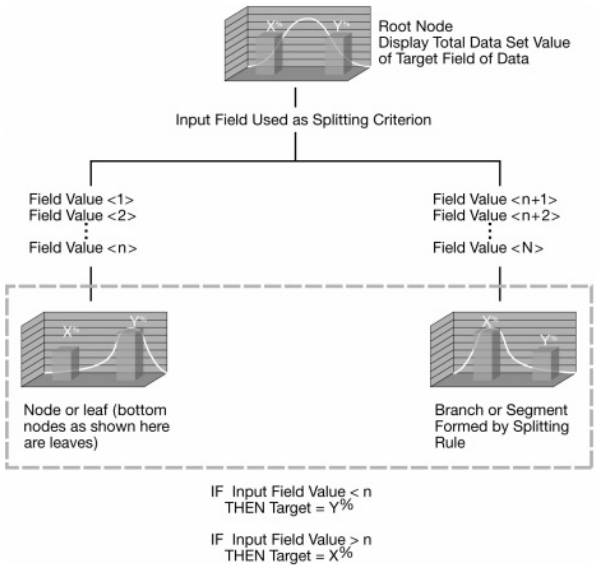


FIGURE 5.1: Illustration of the decision tree

Once such a tree structure is derived, then decision rules can be created to describe the relationships between the inputs (the leaf nodes) and targets (the root node). The rules in Fig. 5.1 are shown below the dashed box at the bottom. Once the decision

rules have been determined, it is possible to use the rules to predict new node values based on new or unseen data. In predictive modeling, the decision rule yields the predicted value.

Although decision trees have been invented and used for around 60 years, many new forms of decision trees are evolving to provide exciting new capabilities in the areas of data mining and machine learning. For example, researchers take advantage of decision tree for variable selection and transformation in developing logistic regression models. In business analytics and business intelligence, decision trees can be used to explore and clarify data for dimensional cubes as well.

5.2.2 Logistic Regression Model

Regression analysis describes the relationship between a response or outcome variable and another set of (one or more) explanatory variables. The intent is to study the effect different covariates (these independent variables are often called covariates) have on a quantitative response. There are many scenarios when the main question of interest involves a dichotomous response: Yes/No, Success/Fail, Sick/Well, etc.

What distinguished a logistic regression model from the well-known linear regression model is that the outcome variable in logistic regression is binary or dichotomous. The difference between logistic and linear regression is reflected both in the choice of a parametric model and in the assumptions. Once the difference is accounted for, the methods employed in an analysis using logistic regression (also called a logit model) follow the same general principles used in linear regression.

In the logistic model, the log odds of the outcome variable is modeled as a linear combination of the predictor variables. Logistic regression fits a model using formula:

$$\log(\frac{\hat{p}}{1-\hat{p}}) = \hat{W}_0 + \hat{W}_1 x_1 + \hat{W}_2 x_2 + \cdots + \epsilon$$

where $x_1, x_2, ...$ are the covariates of interest. W_0 is the intercept for the regression equation. W_i (i=1,2,...) is the regression coefficient. To compare the odds of having an accident between different groups, we often use odds ratios. Let us say p is the

probability of an accident, then the odds of having an accident are given by

$$Odds = \frac{p}{1-p}$$

Say the chance of having a wreck among people with poor vision is 20% and among people with good vision it is 10%. The corresponding odds are 0.25 and 0.11, the ratio of the odds is 2.27. So the odds of an accident are more than double for people with poor vision problems.

Sometimes we find it important to go from odds back to probabilities so also note

$$p = \frac{Odds}{1+Odds}$$

Logistic regressions are closely related to linear regressions. In logistic regression, the expected value of the target is transformed by a link function to restrict its value to the unit interval. In this way, model predictions can be viewed as primary outcome probabilities. A linear combination of the inputs generates a logit score, the log of the odds of primary outcome, in contrast to the linear regression's direct prediction of the target.

Assumptions of logistic regression includes:

- The regression equation should have a linear relationship with the logit form of the dependent variable.
- The dependent variable must be a binary or dichotomy (only including two categories).
- Independent variables (predictors) need not be interval, nor normally distributed, nor linearly related, nor of equal variance within each group.
- The categories (groups) must be mutually exclusive and exhaustive; a case can only be in one group and every case must be a member of one of the groups.
- The error terms need to be independent. Logistic regression requires that each observation should not be from any dependent samples design, e.g., before-after measurements, or matched pairing.

5.2.3 Neural Network

With its exotic sounding name, a neural network model (multi-layer perceptions) is often regarded as a mysterious and powerful predictive weapon. However, the most typical form of the model is a natural extension of a regression model.

The idea of a neural network model is similar to a regression model, but with an interesting and flexible addition. This addition enables a properly trained neural network to model virtually any association between input and target variables. A neural network can be thought of as a regression model on a set of derived inputs, called hidden units. Fig. 5.2 displays a typical structure of a simple neural network.

The nodes can be seen as computational units. In turn, the hidden units can be a thought of as regressions on the original inputs. The hidden units include a default link function, which is also named the activation function. They receive inputs, and process them to obtain an output. This processing might be as simple as summing the inputs, or as complex as containing another network within one node. The connections control the information flow between nodes. The interactions of nodes through the connections lead to a global behaviour shift, which cannot be observed in the elements of the network. This means that the abilities of the network supersede the ones of its elements, making networks a very powerful tool. Therefore, it ensures its ability to approximate virtually any continuous association between the inputs and the target. This usually requires the practitioner to specify the correct number of hidden units and find reasonable values for the weights (the network parameter estimates). Specifying the correct number of hidden units involves some trial and error. Finding reasonable values for the weights is done by least squares estimation.

Multi-layer perception models were originally inspired by neurophysiology and the interconnections between neurons, and they are often represented by a network diagram instead of an equation. Natural neurons receive signals through synapses located on the dendrites or membrane of the neuron. When the signals received are strong enough (surpass a certain threshold), the neuron is activated and emits a signal though the axon. This signal might be sent to another synapse, and might activate other neurons.

An ‘artificial neuron’ is a computational model inspired by the natural neurons,

but highly abstracted with complexity. Basically artificial networks consist of inputs (like synapses), which are multiplied by weights (strength of the respective signals), and then computed by a mathematical function which determines the activation of the neuron. Another function (which may be the identity) computes the output of the artificial neuron (sometimes independent of a certain threshold). The typical model in Fig. 5.2 arranges neurons in layers. The first layer, (input layer) connects to a layer of neurons (hidden layer), which in turn, connects to a final layer (output layer). Each element in the diagram has a counterpart in the network equation.

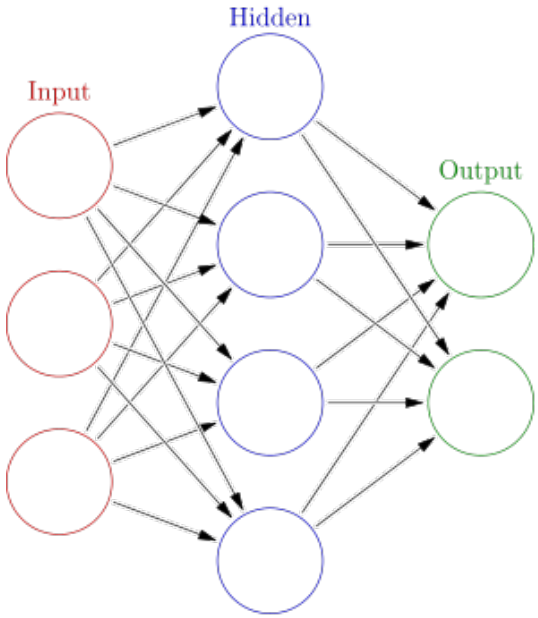


FIGURE 5.2: Illustration of a simple neural network

The basic neural network prediction formula displayed as follows:

$$\hat{y} = \hat{W_{00}} + \hat{W_{01}} \cdot H_1 + \hat{W_{02}} \cdot H_{02} + \hat{W_{03}} \cdot H_{03}$$

$$H_1 = tanh(\hat{W_{10}} + \hat{W_{11}} x_1 + \hat{W_{12}} x_2)$$

$$H_2 = tanh(\hat{W_{20}} + \hat{W_{21}} x_1 + \hat{W_{22}} x_2)$$

$$H_{23} = tanh(\hat{W_{30}} + \hat{W_{31}} x_1 + \hat{W_{32}} x_2)$$

The iteration of learning process is explained here. The higher a weight of an artificial neuron is, the stronger the input which is multiplied by it will be. Weights

can also be negative, so we can say that the signal is inhibited by the negative weight. Differences on the weights generates different computation of the neuron. By adjusting the weights of an artificial neuron we can obtain the output we want for specific inputs. But obviously it would be quite complicated to find by hand all the necessary weights if hundreds or thousands of neurons are present. Algorithms adjust the weights of the neurons in order to obtain the desired output from the network. This process of adjusting the weights is named as learning or training.

Hundreds of different neural network models have been developed to meet different needs. They largely vary in terms of functions, the topology, the learning algorithms, etc. Neural network models are not only used to model real neural networks, and study animal behaviour, but also for engineering purposes, such as pattern recognition, data compression and forecasting.

The performance of these models and the comparison of these models performance will be introduced in the next section.

5.3 Model Building

The models are implemented with the help of SAS Enterprise Miner V7.1. The overall structure is depicted in Fig. 5.3. Decision tree, neural network and logistic regression are built individually at the initial stage. Afterwards, the performance of these three models are evaluated.

Basically the modeling processes in SAS can be divided into three stages as follows:

1. Data preparation: As the first stage, it is the most important step for the following stage. In this stage, all information that can be used to build models is collected. In our case, that means the information we get from the presentation of speech and hand gestures. The information can be either implicit or explicit. We will introduce all the potential factors that we collect from the presentation of speech and hand gestures in the next section.

2. Data exploration: In the second stage, we use SAS to analyse the potential factors

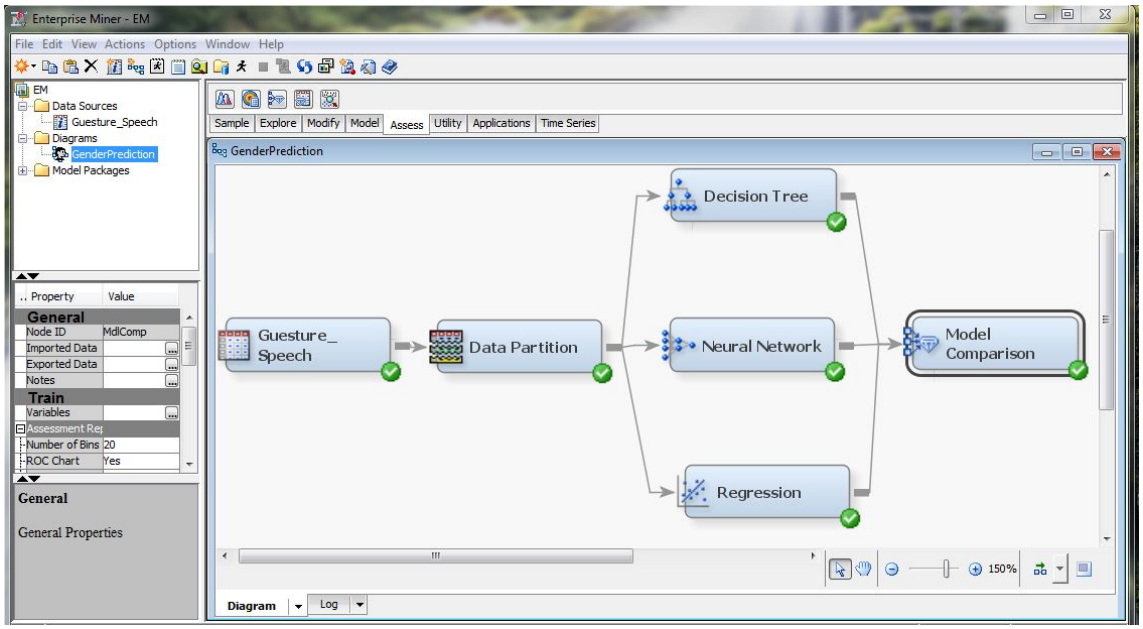


FIGURE 5.3: Overall structure of modeling in SAS

and remove any irrelevant ones. Final factors are selected as input for the next stage.

3. Model building: The third stage is to decide parameters for each models based on the selected factors from the second stage. This is done by SAS modeling functionalities. The optimised parameters are derived to obtain the best performance for this model.

Finally, the performance of these three models is compared. We will introduce these three stages respectively in the following sections.

5.3.1 Data Preparation

Based on the results demonstrated in the previous sections, we found gender differences in the presentation of speech and hand gestures, like the time intervals between the onset of hand gestures and the corresponding lexical affiliates, the time length of gesture strokes and the time length of lexical affiliates. These are explicit factors that might be influential in model building. But there are some other factors that are implicit. We

may not find their contribution in gender difference analysis, but it does not mean they are not useful in model building. We list all the potential factors that we derive from the experimental data, which include 458 paired actions (hand gestures and corresponding keywords) from 9 males and 9 females.

start: The start time point for each keyword of gestures in an audio clip. It represents time information for a keyword.

end: It corresponds to *start* and represents the end time point for a keyword in an audio clip. It also represents time information for a keyword.

start1: Start time point for a gesture in an video clip. This represents time information for a gesture.

end1: End time point for a gesture in an video clip. This also represents time information for a gesture.

keywords: The important words recorded in an audio clip. This represents semantic information.

gender: Gender indicator. This is also our target variable, that is, what we want to predict.

fileN: The file number for the data of each participant. We assign a file number for each data set to de-identify the private information of each participant.

id: It is used to identify the order of each action in time axis of each data set.

Obviously, the absolute time values would be meaningless to be used directly for building models, due to limitations such as coloration and predictive power.

To facilitate the predictive modeling purpose, derived variables are created at the very beginning:

d1: The time length of each keyword (end-start).

d2: The time length of each gesture (end1-start1).

deltaD: The difference between the time length of a gesture and corresponding keyword (d1-d2).

startd: The time interval between the onset of a gesture and its lexical affiliate (start1-start).

endd: The time interval between the end time point of a gesture and its lexical

affiliate (end1-end).

sumD: The sum of gesture stroke time length (d2) and keyword time length (d1) (d1+d2).

absDeltaD: The absolute value of deltaD. The time length of a gesture stroke could be either longer or shorter than the keyword, so the value of *d1-d2* can be positive or negative. *absDeltaD* represents absolute value of *d1-d2*.

overlapF: An Boolean indicator representing whether having an overlap between a gesture stroke and the keyword. The valid values can only be 0 or 1.

overlap: The time span of an overlap. If there is no overlapping (overlapF=0), overlap is set to 0;

overlaprate: The time proportion of overlap over the total time for a paired gesture and keyword (overlap/abs(d1+d2-overlap)).

startflag: A boolean indicator representing whether a gesture starts before its lexical affiliate for each paired action. The valid values can only be 0 or 1.

endflag: A boolean indicator representing whether a gesture ends before its lexical affiliate for each paired action. The valid values can only be 0 or 1.

laststart: The time interval between the start time point of a keyword and the previous one. For the first keyword of every participant, it is set to missing.

lastend: The time interval between the end time point of a keyword and the previous one. For the first action of every participant, it is set to missing.

laststart1: The time interval between the start time point of a gesture and the previous one. For the first action of every participant, it is set to missing.

lastend1: The time interval between the end time point of a gesture and the previous one. For the first action of every participant, it is set to missing.

last2start: The time interval between the start time point of a keyword and the one before the previous one. For the first action of every participant, it is set to missing.

last2end: The time interval between the end time point of a keyword and the one before the previous one. For the first action of every participant, it is set to missing.

last2start1: The time interval between the start time point of a gesture and the one before the previous one. For the first action of every participant, it is set to missing.

last2end1: The time interval between the end time point of a gesture and the one before the previous one. For the first action of every participant, it is set to missing.

last3start: The time interval between the start time point of a keyword and the one before the previous two. For the first action of every participant, it is set to missing.

last3end: The time interval between the end time point of a keyword and the one before the previous two. For the first action of every participant, it is set to missing.

last3start1: The time interval between the start time point of a gesture and the one before the previous two. For the first action of every participant, it is set to missing.

last3end1: The time interval between the end time point of a gesture and the one before the previous two. For the first action of every participant, it is set to missing.

starthalfN: The number of keywords appearing within 0.5 second before a keyword for same participant.

startoneN: The number of keywords appearing within 1 second before a keyword for same participant.

startoneandhalfN: The number of keywords appearing within 1.5 seconds before a keyword for same participant.

starttwoN: The number of keywords appearing within 2 seconds before a keyword for same participant.

startthreeN: The number of keywords appearing within 3 seconds before a keyword for same participant.

startfiveN: The number of keywords appearing within 5 before a keyword for same participant.

start1halfN: The number of gestures happening within 0.5 second before a gesture for the same participant.

start1oneN: The number of gestures happening within 1 second before a gesture for the same participant.

start1oneandhalfN: The number of gestures happening within 1.5 seconds before a gesture for the same participant.

start1twoN: The number of gestures happening within 2 seconds before a gesture for the same participant.

start1threeN: The number of gestures happening within 3 seconds before a gesture for the same participant.

start1fiveN: The number of gestures happening within 5 second before a gesture for the same participant.

From the definition of these factors, we can see that most of them are time-related factors and some of the them are relevant to the integration pattern of speech and hand gestures (e.g. *overlap* and *overlaprate*). As demonstrated in [14] that users keep their habitual integration pattern across the whole session, we extracted these time-related factors to indicate the patterns of the presentation of speech and hand gestures for different participants.

5.3.2 Data Exploration

After data preparation, all the extracted factors are fed into SAS models. Frequency analysis is implemented in SAS to find dependence structure underlying these factors (in SAS, factors are also called variables). A screenshot of data exploration is given in Fig. 5.4. Keywords show large dispersion and vary from one participant to another. *ID* and *fileN* are also excluded for model building, since they do not have any predictive power. In total, 5 time-related variables and 2 nominal variables are rejected by SAS automatically. The final input variables for each model include 3 binary and 45 time-related variables all together. Our target is to predict gender (Female and Male).

In predictive modeling (also known as supervised prediction or supervised learning), the standard strategy for honest assessment of model performance is data splitting. A portion is used for fitting the model, that is, the training data set. The remaining data are separated for empirical validation. Predictive modeling starts with a training data set. The samples in a training data set are known as training cases (also known as example or instances). The variables are called inputs (also known as predictors, features, or independent variables) and targets (also known as outcome, or dependent variables). For a given sample, the input reflects the state of knowledge before measuring the target. For data partition purpose, we split 458 records into a training set and

	Variable Name	Type	Percent Missing	Minimum	Maximum	Mean	Number of Levels	Mode Percentage	Mode
1	gender	CLASS	0	.	.	.	2	56.76856	F
2	keywords	CLASS	0	.	.	.	128+	8.674699	BACK
3	name	CLASS	0	.	.	.	18	12.8821	MANOLYA
4	absdeltaD	VAR	0	0	9.17	0.871581	.	.	
5	d1	VAR	0	0	9.77	1.350721	.	.	
6	d2	VAR	0	-0.14	1.55	0.531646	.	.	
7	deltaD	VAR	0	-0.88	9.17	0.819074	.	.	
8	end	VAR	0	3	349.63	87.61238	.	.	
9	end1	VAR	0	2.74	351.02	87.303	.	.	
10	endd	VAR	0	-8.22	2.62	-0.30938	.	.	
11	endflag	VAR	0	0	1	0.50655	.	.	
12	id	VAR	0	1	458	229.5	.	.	
13	last2end	VAR	0.436681	-330	65.77	0.088465	.	.	
14	last2end1	VAR	0.436681	-332.72	64.88	0.084956	.	.	
15	last2start	VAR	0.436681	-332.4	63.8	0.084583	.	.	
16	last2start1	VAR	0.436681	-332.77	64.87	0.085	.	.	
17	last3end	VAR	0.655022	-326.1	68.27	0.106176	.	.	
18	last3end1	VAR	0.655022	-326.56	67.73	0.101341	.	.	
19	last3start	VAR	0.655022	-328.04	67.6	0.101758	.	.	
20	last3start1	VAR	0.655022	-326.66	67.27	0.101736	.	.	
21	lastend	VAR	0.218341	-333	45.1	0.0507	.	.	
22	lastend1	VAR	0.218341	-335.74	44.47	0.048446	.	.	
23	laststart	VAR	0.218341	-336.6	43.46	0.048074	.	.	
24	laststart1	VAR	0.218341	-335.81	44.69	0.048468	.	.	
25	overlap	VAR	0	0	1.55	0.300677	.	.	
26	overlapF	VAR	0	0	1	0.757642	.	.	
27	overlaprate	VAR	0	0	0.933333	0.23342	.	.	
28	start	VAR	0	1.8	349.03	86.26166	.	.	
29	start1	VAR	0	2.36	350.39	86.77135	.	.	
30	start1five	VAR	0	-2.64	345.39	81.77135	.	.	
31	start1fiveN	VAR	0	1	5	1.917031	.	.	
32	start1half	VAR	0	1.86	349.89	86.27135	.	.	
33	start1halfN	VAR	0	1	2	1.026201	.	.	
34	start1one	VAR	0	1.36	349.39	85.77135	.	.	
35	start1oneN	VAR	0	1	2	1.058952	.	.	
36	start1oneandhalf	VAR	0	0.86	348.89	85.27135	.	.	
37	start1oneandhalfN	VAR	0	1	3	1.183406	.	.	
38	start1three	VAR	0	-0.64	347.39	83.77135	.	.	
39	start1threeN	VAR	0	1	4	1.521834	.	.	
40	start1two	VAR	0	0.36	348.39	84.77135	.	.	
41	start1twoN	VAR	0	1	4	1.28821	.	.	
42	startd	VAR	0	-5.72	2.84	0.509694	.	.	
43	startfive	VAR	0	-3.2	344.03	81.26166	.	.	
44	startfiveN	VAR	0	1	6	1.89083	.	.	
45	startflag	VAR	0	0	1	0.847162	.	.	
46	starthalf	VAR	0	1.3	348.53	85.76166	.	.	
47	starthalfN	VAR	0	1	2	1.026201	.	.	
48	startone	VAR	0	0.8	348.03	85.26166	.	.	
49	startoneN	VAR	0	1	3	1.082969	.	.	
50	startoneandhalf	VAR	0	0.3	347.53	84.76166	.	.	
51	startoneandhalfN	VAR	0	1	4	1.187773	.	.	
52	startthree	VAR	0	-1.2	346.03	83.26166	.	.	
53	startthreeN	VAR	0	1	6	1.504367	.	.	
54	starttwo	VAR	0	-0.2	347.03	84.26166	.	.	
55	starttwoN	VAR	0	1	5	1.31441	.	.	
56	sumD	VAR	0	0.58	10.41	1.882367	.	.	

Figure 5.4: Data exploration

validation set in the proportion of 80:20 (364 samples for training set and 94 samples for validation set created by SAS).

The validation data set is used for monitoring and tuning the model to improve its generalization. The tuning process usually involves selecting of different types and complexities among models. It optimises the selected model based on the validation data. These steps are carried out by SAS automatically by and large.

5.3.3 Evaluation the Performance of Models

In this section we will evaluate the performance of the three models respectively:

- Decision tree
- Neural network
- Logistic regression

1. Decision Tree

The output of decision tree modeling in SAS can be viewed as a tree map, depicted in Fig. 5.5. Each leaf of the tree map represents a generated business rule for prediction.

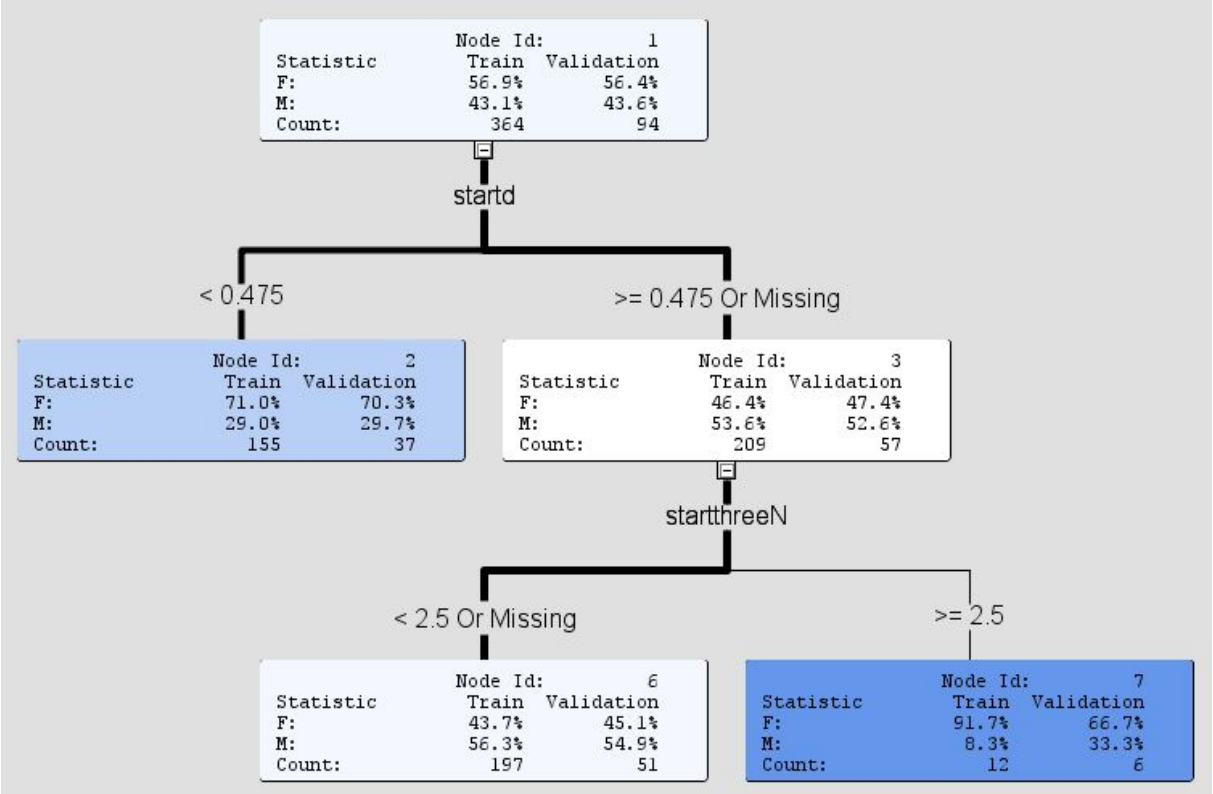


FIGURE 5.5: The full tree map for the decision tree models

The business rules generated from the decision tree model are shown in Fig. 5.6.

English Rules

```
1  *------------------------------------------------------------*
2   Node = 2
3  *------------------------------------------------------------*
4  if startd < 0.475
5  then
6   Tree Node Identifier   = 2
7   Number of Observations = 155
8   Predicted: gender=M  = 0.29
9   Predicted: gender=F  = 0.71
10
11 *------------------------------------------------------------*
12  Node = 6
13 *------------------------------------------------------------*
14 if startthreeN < 2.5 or MISSING
15 AND startd >= 0.475 or MISSING
16 then
17  Tree Node Identifier   = 6
18  Number of Observations = 197
19  Predicted: gender=M  = 0.56
20  Predicted: gender=F  = 0.44
21
22 *------------------------------------------------------------*
23  Node = 7
24 *------------------------------------------------------------*
25 if startthreeN >= 2.5
26 AND startd >= 0.475 or MISSING
27 then
28  Tree Node Identifier   = 7
29  Number of Observations = 12
30  Predicted: gender=M  = 0.08
31  Predicted: gender=F  = 0.92
```

Figure 5.6: Business rules generated from the decision tree models

We can interpret this rule as below:

If *startd* (the time interval between the onset of a gesture and its corresponding keyword) is less or equal than 0.475 seconds, the gender of this participant is likely to be female rather than male.

If *startthreeN* (the number of keywords appeared within 3 seconds before a keyword for same participant) is greater or equal than 2.5 and *startd* (the time interval between the onset of a gesture and its corresponding keyword) is less or equal than 0.457 seconds, the gender of this participant is highly likely to be female rather than male.

The model selection process is shown in Fig. 5.7. We can see that if the number of leaves is greater than 5 then, the misclassification rate has a large drop in the training

set (the lower curve) but in the validation set (the upper curve), it levels. Probably, this indicates an over-fitting in the model. With the increase of the number of leaves from 3 to 5, the misclassification rate on the validation data set does not change. But more leaves generate more complex business rules. So the optimal number of leaves is 3.

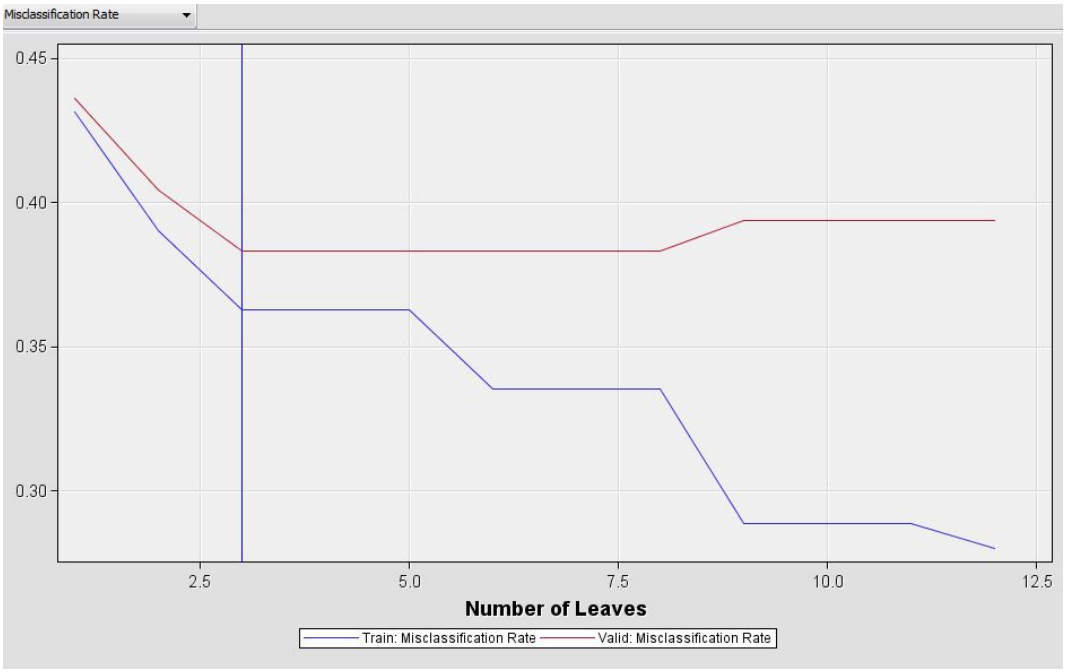


FIGURE 5.7: The model selection process for the decision tree

The model fit statistics are listed in Table 5.1. In this table, what we are concerned with is the accuracy rate based on the validation data sets (0.62).

	Accuracy	Misclassification Rate	Maximum Absolute Error	Sum of Squared Errors	Average Squared Error	Root Average Squared Error
Training	0.64	0.36	0.92	162.62	0.22	0.47
Validation	**0.62**	0.38	0.92	44.16	0.23	0.48

TABLE 5.1: Classification rate for decision tree

Table 5.2 shows the confusion matrix of the decision tree model for gender prediction. Confusion matrix is a widely used method to allow visualization of the performance of the predictive models. For example, false negative represents the number of females were incorrectly marked as male. True Negative represents the number of male correctly classified as male. False Positive represents the number of males that were incorrectly labeled as female. True Positive represents the number of female that were correctly classified as female. Obviously the sum of True Negative and True Positive indicate the numbers of the correctly classified samples.

	False Negative	True Negative	False Positive	True Positive
Training	46	121	86	111
Validation	13	30	23	28

Table 5.2: Confusion matrix for decision tree event classification

Variable importance rank (degree of the factors' contribution to the final gender prediction) is also shown in Table 5.3. We can see from this table that *startd* (the time interval between the onset of a gesture and its corresponding keyword) plays a critical role to distinguish the two gender groups. It is consistent with our previous findings.

Name	NRULES	IMPORTANCE	VIMPORTANCE	RATIO
startd	1	1.00	1	1
startthreeN	1	0.70	0	0

Table 5.3: Variable importance for decision tree

2. Neural network

An artificial neural network can be defined as a computer application that attempts to mimic the neurophysiology of the human brain in the sense that it learns from examples to find patterns in data. By finding complex non-linear relationship in data, neural networks can help to make predictions about real-world problems.

For the neural network modeling, SAS uses all the samples in training set to compute and optimise the parameters for each neuron as introduced in Section 5.2.3.

The training process of neural network includes 50 times iterations as indicated in the x-axis of Fig. 5.8. Y-axis of this figure shows misclassification rate corresponding to specific iteration times.

FIGURE 5.8: Misclassification rate for neural network

The iteration is stopped when the best performance is achieved that is optimised by SAS. The model fit statistics are listed in Table 5.4. From this table we can see the accuracy rate in validation data set is 0.65 which is a slightly higher than decision tree model. In this model, all input variables contribute to final model building.

	Accuracy	Misclassification Rate	Maximum Absolute Error	Sum of Squared Errors	Average Squared Error	Root Average Squared Error
Training	0.68	0.32	0.90	135.84	0.19	0.43
Validation	**0.65**	0.35	0.96	46.86	0.25	0.50

TABLE 5.4: Classification rate for neural network

Confusion matrix of neural network is shown in Table 5.5. The sum of correctly

classified samples in this model is more than those found in the decision tree model.

	False Negative	True Negative	False Positive	True Positive
Training	49	140	67	108
Validation	13	33	20	28

Table 5.5: Confusion matrix for neural network event classification

3. Logistic regression

As presented in Section 5.2.2, the important thing for logistic regression modeling is to determine the regression coefficient.

SAS uses different variable selection methods to generate a regression model. In the modeling process, gender is a binary value and we mark female as the 1 (target variable) and male as 0. After all variables are fed into logistic model in SAS, each sample in data set will be given a probability score (called logit, refer to Section 5.2.2) to represent the likelihood for being female's action (in contrast to male). That is, the higher the score is, the higher chance that the sample is from females.

The final model is investigated in the cumulative captured response graph in Fig. 5.9. The cumulative captured response is a measure of how many target events (which are females in our model, for simplicity) are identified in each percentile. Given a data has been ranked from lowest to highest, a n percentile means n% of the data is below it. For example, the 50th percentile (also called the median) is point that half of the data below it.

In Fig. 5.9, it shows that more than one-third (33.8%) of females have been identified/correctly predicted in the first 20% of samples ranked by the predicted score. This will be useful for model comparison in the next section.

The model fit statistics are listed in Table 5.6. From this table we can see the accuracy rate of logistic regression model is 0.70 which is the highest among these three models.

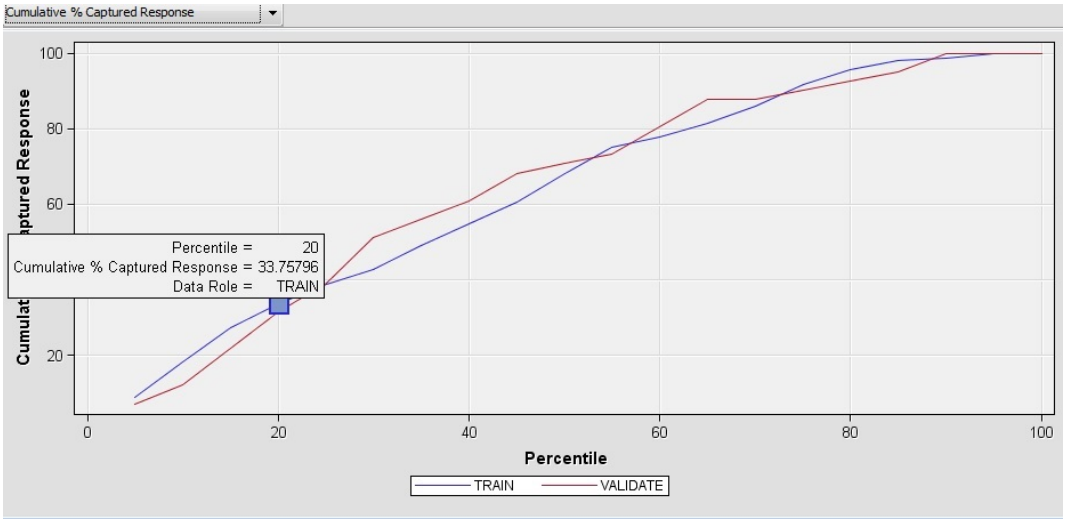


FIGURE 5.9: Final model for logistic regression model

	Accuracy	Misclassification Rate	Maximum Absolute Error	Sum of Squared Errors	Average Squared Error	Root Average Squared Error
Training	0.64	0.36	0.90	150.46	0.21	0.45
Validation	**0.70**	0.230	0.88	39.40	0.21	0.46

TABLE 5.6: Classification rate for logistic regression

Maximum likelihood estimates are calculated in Table 5.7. Maximum likelihood selects the set of values of model parameters that maximised the likelihood estimation. The idea of maximum method is illustrated by an example: one may be interested in the heights of all adult male penguins, but be unable to measure the height of every single penguin in a population due to cost or time constraints. Assuming that the heights are normally distributed with some unknown mean and variance, the mean and variance can be estimated with maximum likelihood while only knowing the heights of some sample of the overall population. Maximum likelihood accomplishes this by taking the mean and variance as parameters and finding particular parametric values that make the observed results the most probably given the model.

Parameter	DF	Estimate	Pr>ChiSq	Exp(Est)
Intercept	1	11.81	0.89	999.00
absdeltaD	1	-0.85	0.31	0.43
d1	1	1.21	0.17	3.34
d2	1	-2.340	0.08	0.09
...	...	...	...	...

Table 5.7: Analysis of Maximum Likelihood Estimates

In Table 5.7, "Parameter" represents the variables fed into modeling process and "Estimate" represents the regression coefficients estimated by SAS for the corresponding parameter. The positive value indicates that positive correlation between the parameter and the target variable, which is female in our case.

Confusion matrix for this model is shown in 5.8. As introduced in the previous section, the sum of "True Negative" and "True Positive" indicates the number of correctly identified samples. We can see that this number is the greatest for logistic regression model compared to the previous two models.

	False Negative	True Negative	False Positive	True Positive
Training	71	148	59	86
Validation	16	41	12	25

Table 5.8: Confusion matrix for logistic regression event classification

5.3.4 Model Comparison

The cumulative captured response graph is depicted in Fig. 5.10. Logistic regression is shown to outperform the neural network and the decision tree. In the first 20% validation dataset as ranked by prediction score, logistic regression correctly identified 31.7% of gender information, which is above the counterparts (decision tree and neural network). This proves the performance of the regression model is better than a random

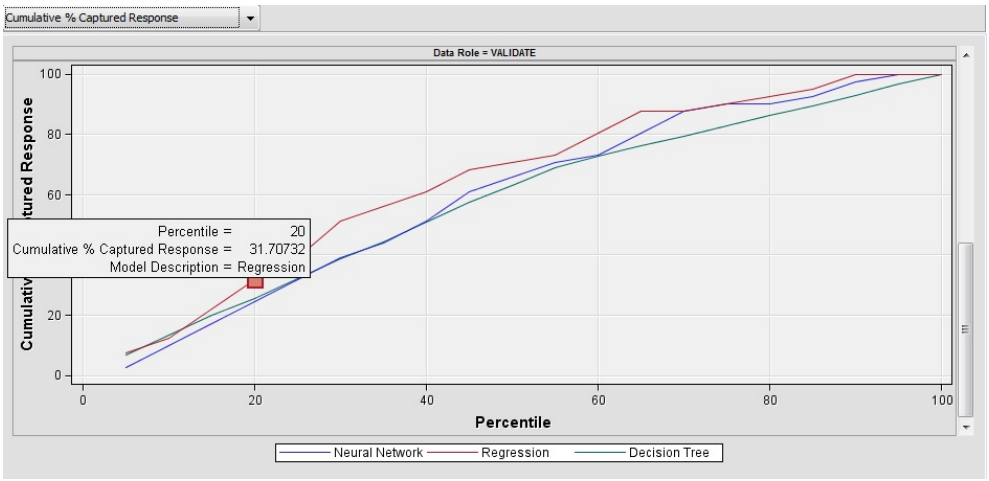


FIGURE 5.10: The cumulative of captured response graph

guessing model.

ROC chart (a plot of true positive rate against false positive rate) [221] in Fig. 5.11 is used to compare model performance by SAS. The larger area under ROC curve of a model indicates a better performance. It shows *neural network* is the best performed model in the training set, however, second to *regression* in the validation set. This possibly indicates neural network is suffering over-fitting issue.

Fitting a model to data requires searching through the space of possible models. Constructing a model with good generalization requires choosing the right complexity. Selecting model complexity involves a trade-off between bias and variance. An insufficiently complex model might not be flexible enough, which can lead to under-fitting, that is, systematically missing the signal (high bias).

A naive modeler might assume that the most complex model should always outperform the others, but this is not the case. An overly complex model might be too flexible, which can lead to over-fitting, that is, accommodating nuances of the random noise in the particular sample (high variance). A model with the right degree of flexibility gives the best generalization.

Model selection is based on lowest misclassification rate in the validation set (see Table 5.9). In our case, the logistic regression model demonstrates the trade-off between

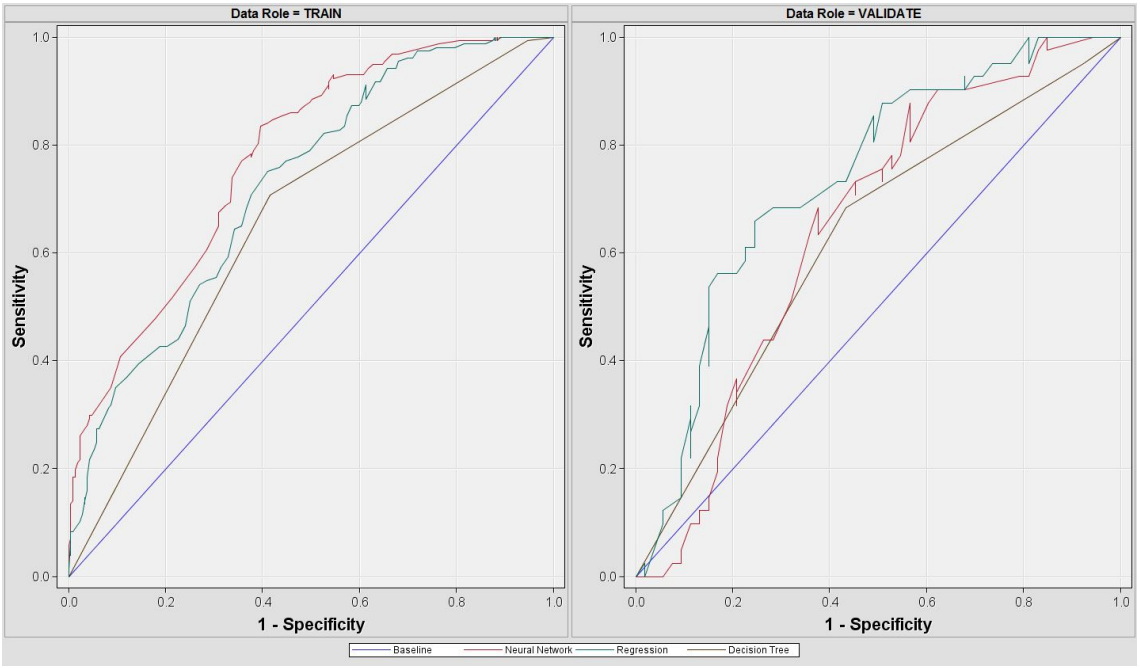


Figure 5.11: ROC graph of the three models

the complexity and classification rate.

Model Description	Validation: accuracy Rate	Validation: misclassification Rate
Logistic Regression	**0.70**	0.30
Neural Network	0.65	0.35
Decision Tree	0.62	0.38

Table 5.9: Model classification accuracy comparison

5.4 Conclusion

In this chapter, we explored the possibility to predict gender using the gender differences in speech and hand gestures. Adopting the different integration strategies for male and female users in MMIS will be feasible if the multimodal actions can be predicted before being integrated together.

Even though there a few studies addressing gender prediction using speech and hand gestures, the methods used are complicated to achieve good performance [34, 35, 151, 152, 153]. We attempted to achieve the acceptable prediction accuracy with a simple but effective approach. Three statistical modelling methods (decision tree, neural network and logistic regression) were studied and compared. The performance of these three methods was evaluated in terms of the classification accuracy.

The results showed that the logistic regression model can achieve better performance with the trade-off between the complexity and classification rate.

6

Conclusion and Future Work

6.1 Conclusion

In this thesis, we have analysed gender differences in speech and hand gestures for the integration in MMIS. We reviewed current work in the area of MMIS under the following headings: the input modalities that can be used in MMIS, the integration strategies for input modes and the frameworks for multimodal information fusion. The ultimate goal of MMIS is to eliminate the gap between HCI and human-human communication by providing users with a choice and switch of input and accommodating adaptive strategies for different users.

We also reviewed the cognitive theories regarding multimodal processes. These theories support that our brain can accommodate multimodal processing of information and our working memory also deals with multimodal input in a coordinated manner.

This allows HCI systems to simulate natural ability to process and produce multimodal information. Thus, MMIS will be able to make flexible use of the entire gamut of modal productions (e.g. gaze, gesture and speech) to improve system performance.

Although there has been a large amount of work in the area of MMIS, with some in depth, there is still a dearth of attention to several high-level problems, such as user-related factors in the design of MMIS, and more specifically the gender of the user. As depicted in Chapter 2, the performance of each input branch in the framework in Fig. 1.1 may significantly affect the overall performance of the system. In other words, user-related factors can be significantly influential for the design of MMIS. Gender differences have been studied in other areas, however, only a few studies have paid attention to gender differences in speech and hand gestures as well as the reasons for these differences.

In this thesis we studied gender differences focusing on the following three aspects:

- presentation of speech and hand gestures;
- cognitive processing in the coordination of speech and hand gestures;
- brain activities when speech and hand gestures are used together.

Fig. 6.1 shows a summary of our work. As indicated in this figure, our findings suggest that gender differences in the coordination of speech and hand gestures occur externally (in the presentation of speech and hand gestures) and internally (in cognitive processing model and brain activities).

The first hypothesis in this thesis was established as follows:

H1:There are gender differences in the coordination of speech hand gestures as well as their temporal alignment in mutimodal information processing.

Based on this hypothesis, we raised some questions in Chapter 1 to study the gender differences in the external presentation of speech and hand gestures. We list them again as follows:

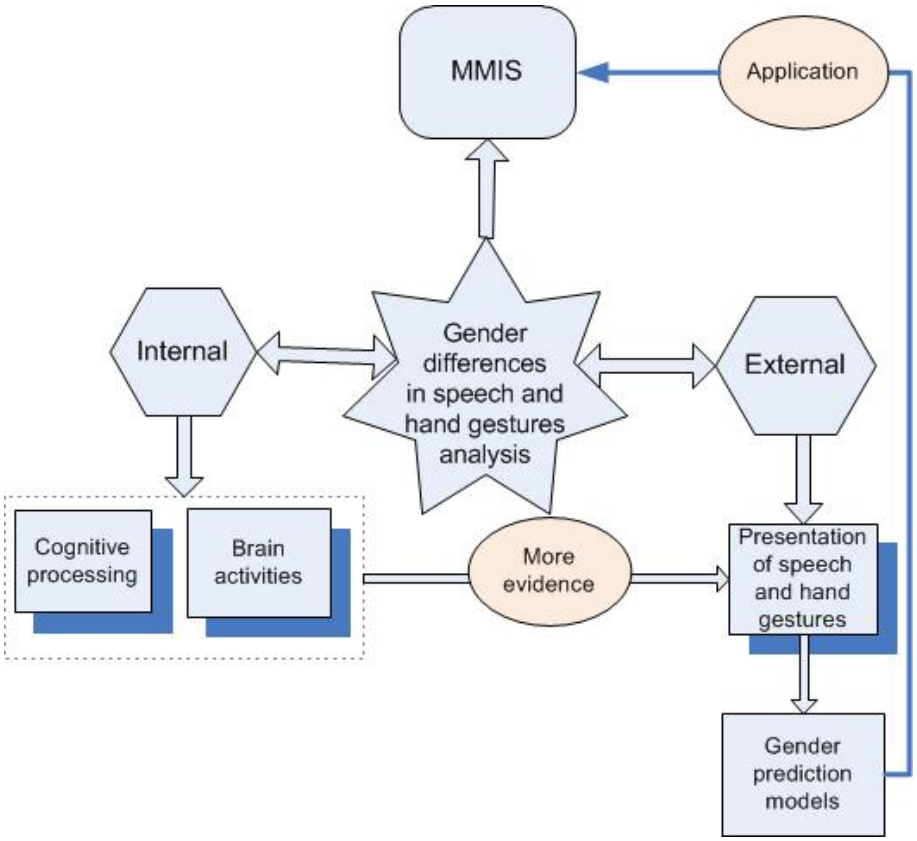


Figure 6.1: Future work

RQ1: Are there any gender differences in using speech and hand gestures? What are the similarities and differences between males and females in the coordination of speech and hand gestures given the same tasks? Are there any gender differences in the integration or temporal alignment patterns of speech and hand gestures?

Statistical analysis of the task time spent by males and females suggests no significant differences in them. However, our **Finding 1** suggests that females spend more time on gestures than males regardless of the total task time and also for females, the amount of time spent on gestures also takes greater proportion of the total task time than males. This indicates that males and females have their own preferences in the choice of input modes regarding the use of speech and hand gestures. We also found that the preference of using speech and hand gestures for males and females are stable over time even for the different tasks (**Finding 2**).

As reviewed in Chapter 2, males and females use different words and gestures in communication. Some researchers suggest that multimodal language does not differ linguistically from unimodal language. We found in our experiments that, for object description tasks, the corresponding lexical affiliates of hand gestures are adjectives and nouns for both females and males. However, for females, nouns are the dominant type, while for males adjectives are the majority (**Finding 4**). Our findings suggest that the performance of MMIS may be improved with the use of different vocabularies of speech and hand gestures for different gender groups.

Regarding the differences in the integration or alignment of speech and hand gestures, we found that, generally speech and hand gestures are tightly synchronised with each other. Males and females actually have similar integration patterns in which gestures precede the related speech within 2 seconds and have overlaps with corresponding lexical affiliates on the time axis (**Finding 6**). In our annotations for female participants, 81.15% of hand gesture strokes precede the related lexical affiliates. For male participants, it is even higher (89.39%). However, the temporal alignment of speech and hand gestures varies for males and females. The time lags between speech and co-occurring hand gestures are shorter for females than males (**Finding 7**). Also our findings showed that the duration of gesture strokes and related keywords are significantly different in males and females (**Finding 3 and Finding 5**). These findings suggest that gender is a significant factor in the integration of speech and hand gestures for the design of MMIS. Adaptive integration strategies for different gender groups may improve the performance of systems.

In Chapter 2, we provided a review on cognitive theories related to multimodal processing. These theories state that our brain processes information multimodally and working memory deals with multimodal modes in a coordinated manner. User-centered interface design can free up mental resources and further improve user performance [54, 55]. Based on these, we suggested that speech and hand gestures are integrated systems, but there are gender differences in processing these two types of input.

The second hypothesis we established in this thesis regarding gender differences in cognitive processing was as follows:

H2: Males and females employ different cognitive processing models in the coordination of speech and hand gestures. Gender differences in cognitive processing might be a reason for the differences in the presentation of speech and hand gestures.

In order to explain these external differences in the presentation of speech and hand gestures, we studied the cognitive processes of males and females in speech and hand gestures. Regarding this hypothesis we formed two questions:

RQ2: Do males and females employ different cognitive processing models in the coordination of speech and hand gestures? Are the differences in the presentation of speech and coordinated hand gestures driven by the differences in cognitive processing?

In our experiments, we observed that males and females present different cognitive processing models. Males were found to have more perceptual actions than functional actions while females have more functional actions than perceptual actions in their protocols (**Finding 8**). We also found that females' protocols also include more cognitive actions than males (**Finding 9**). Our findings support the fact that cognitive activity is higher in the females object description protocols and the use of speech and co-occurring hand gestures.

According to the definition of cognitive actions, perceptual actions reflect the information processing in mind about visual features of objects, spatial relations among elements and the comparison between elements. Functional actions represent the attention to the structure and function of elements. We found that perceptual actions are externalised by adjectives, prepositions and adverbs in speech and co-occurring gestures these are preferred by males, while functional actions are more likely to be reflected by nouns and numerics and these are preferred by females (**Finding 10**). On other words, gender differences in cognitive processing models might be the reason for the differences in the distribution of word types accompanying hand gestures.

In Chapter 2, our literature review demonstrated that there are gender differences in brain activities associated with language processing and female brain is less lateralised with functions spread over both sides of the brain and left-hemispheric dominance of language functions is greater in males than it is in females.

The third hypothesis in this thesis was established as follows:

H3: There are gender differences in brain activities in the coordination of speech and hand gestures.

Our research question about this hypothesis was:

RQ3: Are there any gender differences in brain activities in speech and coordinated hand gestures? Is the male brain more lateralising in the coordination of speech and hand gestures?

In our experiments, we did not find enough evidence to support the lateralisation hypothesis for males as found by others. However, we still found some gender differences in brain activities in the coordination of speech and hand gestures.

We examined the spectral moment in beta and alpha frequency bands for EEG signals of the left and right hemispheres of the brain associated with speech and gesture coordination and found that the gender differences are quite minor (**Finding 11**). Due to the limitation of the EEG device (Emotiv Neuroheadset), we could not examine the functional brain activities in speech and hand gesture coordination in detail. This might also be the reason for lack of evidence for the lateralisation hypothesis.

However we found that females show stronger beta spectral moment and more significant changes in spectral moment from alpha to beta band (**Finding 12**). This may require shorter integration intervals (fast connection) of speech and hand gestures for females in MMIS design. Actually there are studies which found that males showed more lateralisation of emotional activity, and females showed more brainstem activation in affective paradigms [222]. Strong brainstem activation may cause strong spectral moment in brain activities of females. Gender differences in grey and white matter are also reported as *"In general, men have approximately 6.5 times the amount of gray matter related to general intelligence than women, and women have nearly 10 times the amount of white matter related to intelligence than men. Gray matter represents information processing centres in the brain, and white matter represents the networking of - or connections between - these processing centres"* [223]. Those connections may allow a woman's brain to work faster than a man's.

In summary, our findings demonstrated in this thesis support the three hypotheses

regarding gender differences in the coordination of speech and hand gestures. The internal and external gender differences we suggest that gender could have significant impact on the design of MMIS using speech and hand gestures as input modes. MMIS could potentially gain better performance by accommodating gender differences with gender adaptive processing strategies.

Using our findings regarding gender differences in the presentation of speech and hand gestures, we developed models to predict the users' gender evaluating their multimodal actions. We compared three different modeling methods (decision tree, neural network and logistic regression) to predict users' gender, and found that a reasonable performance can be achieved by logistic regression model with an accuracy over 70%. Thus, we demonstrated that various gender prediction models can successfully be implemented using our findings and our results are promising for the design of gender adaptive MMIS.

6.2 Future Work

In Fig. 6.1, we show two major directions for future work. First is to further investigate the correlation of internal and external differences in speech and hand gestures of males and females. The Emotiv Neuroheadset used in our studies has 14 sensors. Though it is reliable to a certain degree, more precise device measurement may allow us to study the functional brain activities associated with speech and hand gestures. The annotation of speech and hand gestures were mainly conducted manually by coders in this thesis. With the new technology (e.g. Kinect), automated or semiautomated processing of speech and hand gestures may be achieved in future. Larger group samples will also be helpful for the further verification of our findings.

Another direction for future research is the application of our findings. Most current studies do the offline analysis. This makes sense given the relatively young nature of MMIS involving speech and manual hand gestures. However, as better technologies are developed and algorithms mature, the need for gender adaptive processing methods for different users in MMIS will be increasing. How to implement predictive and adaptive

gender based information processing models in MMIS is still an open issue.

A

Appendix

INFORMATION SHEET AND CONSENT FORM

Name of the project:

Gender Differences in Visual Cognition for Multimodal Systems Design

This research project will study gender differences in

- speech and hand gestures,
- cognitive processing, and
- brain activities.

This experiment will be recorded, either by a digital camera and/or by a microphone embedded in the camera.

The tools that are going to be used for this experiment are a digital camera, microphone, computer and Emotiv Headset.

These tools are safe and publicly available. The non-standard saline solution may cause allergy to sensitive skin. The likelihood is minimal in this study. However, you CANNOT participate if you are sensitive to the saline. A verbal warning will be given prior commencing the study.

Please read the following points carefully:

- Should you decide to participate, you may quit anytime during the study but please remain in the VR lab and wait until the researcher has removed the Emotiv Headset.

- Whenever you feel uncomfortable during the study, please immediately let the conductor know. Conductor will be in the VR lab during the entire session.

- Should you decide to stay during the study and experience severe discomfort, we will refer you to on-campus medical service.

The location of on-campus medical service is included in this information statement and consent form.

At the beginning of the experiment, you will be given a fifteen-minute tutorial on the purpose of the experiment and how to use the necessary applications, during which time you will be introduced to the system. Feel free to ask any questions you may have about the experiment or about the system. The total time commitment involved is estimated to be 30 minutes.

All material, including video recordings will be kept strictly confidential and will not be made available to any persons outside this project. The researchers have no material interest in the outcome of this experiment. The results will be presented at departmental research seminars, peer-reviewed Australian and International conferences and via peer-reviewed journal articles. We will only use the images and speech in the video clips after the participants identity is obscured in presentations and publications. The participants faces will not be exposed under any circumstances. The de-identified data would be retained for inclusion in related research by the investigators in the future.

Participation in this study is entirely voluntary. You are under no obligation to participate, and may withdraw your consent to participate at any time without consequence to you. If you are interested in this study, A/Prof. Manolya Kavakli and Jing Liu will be happy to discuss it further with you and answer any queries you may have. Please feel free to contact on (02) 98509572.

Participants can obtain feedback regarding the results of the project from the Interactive Systems and Virtual Reality Research Group website located at

http://web.science.mq.edu.au/groups/visor/

Thank You.

www.research.mq.edu.au/researchers/ethics/human_ethics/forms/

Medical Service on campus:

Suite 305, Level 3

Macquarie University Clinic Building (F10A)

2 Technology Place

Macquarie University NSW 2109 Tel: (02) 9812 3944 or (02) 9812 3096

For further queries about this study, please contact:

Dr. Manolya Kavakli (Chief Inv.)	02 9850 9572	manolya.kavakli@mq.edu.au
Jing Liu (PhD student)	02 9850 9548	jing.liu21@students.mq.edu.au

I, _____have read (or, where appropriate, have had read to me) and

understood the information given and any questions I have asked, have been answered

to my satisfaction. I agree to participate in this research study, entitled **Gender Differences in Visual Cognition for Multimodal Systems Design**, which is conducted by

Dr. Manolya Kavakli (A/Pro., Dept. of Computing, Macquarie University),

Jing Liu (PhD student, Dept. of Computing, Macquarie University),

knowing that participation is entirely voluntary and I can

withdraw from further participation in the research at any time without consequence.

I allow ☐ / do not allow ☐ the de-identified data to be retained for inclusion in related research by the investigators in future.

I have been given a copy of this signed form to keep.

Participants Name:_____ (block letters)

Participants Signature: _____ Date:_____

Investigators Name:_____ (block letters)

Investigators Signature: _____ Date: _____

The ethical aspects of this study have been approved by the Macquarie University Human Research Ethics Committee. If you have any complaints or reservations about any ethical aspect of your participation in this research, you may contact the Committee through the Director, Research Ethics (telephone (02) 9850 7854; email ethics@mq.edu.au). Any complaint you make will be treated in confidence and investigated, and you will be informed of the outcome.

References

[1] M. Suwa, T. Purcell, and J. Gero. Macroscopic analysis of design processes based on a scheme for coding designers' cognitive actions. Design studies **19**(4), 455 (1998).

[2] P. C. Yuen, Y. Y. Tang, and P. S. Wang. *Multimodal Interface for Human-Machine Communication*, vol. 48 (World Scientific Publishing Company, 2002).

[3] D. Tan and A. Nijholt. Brain-Computer Interfaces and Human-Computer Interaction. In *Brain-Computer Interfaces*, pp. 3–19 (Springer, 2010).

[4] B. Dumas, D. Lalanne, and S. Oviatt. Multimodal interfaces: A survey of principles, models and frameworks. Human Machine Interaction pp. 3–26 (2009).

[5] A. Özyürek, R. M. Willems, S. Kita, and P. Hagoort. On-line integration of semantic information from speech and gesture: Insights from event-related brain potentials. Journal of Cognitive Neuroscience **19**(4), 605 (2007).

[6] A. S. Dick, S. Goldin-Meadow, U. Hasson, J. I. Skipper, and S. L. Small. Co-speech gestures influence neural activity in brain regions associated with processing semantic information. Human brain mapping **30**(11), 3509 (2009).

[7] R. Bolt. Put-that-there: Voice and gesture at the graphics interface. pp. 262–270 (ACM New York, NY, USA, 1980).

[8] F. Qiao, J. Sherwani, and R. Rosenfeld. Small-vocabulary speech recognition for resource-scarce languages. In *Proceedings of the First ACM Symposium on Computing for Development*, p. 3 (ACM, 2010).

[9] J. Liu and M. Kavakli. Hand gesture recognition based on segmented singular value decomposition. Knowledge-Based and Intelligent Information and Engineering Systems pp. 214–223 (2010).

[10] A. Boyali, M. Kavakli, and J. Twamley. Real Time Six Degree of Freedom Pose Estimation Using Infrared Light Sources and Wiimote IR Camera with 3D TV Demonstration. In *Mobile and Ubiquitous Systems: Computing, Networking, and Services*, pp. 137–148 (Springer, 2012).

[11] J. Eisenstein and C. M. Christoudias. A salience-based approach to gesture-speech alignment. In *proceedings of HLT-NAACL*, vol. 4, pp. 25–32 (2004).

[12] S. Carbini, J. E. Viallet, and L. Delphin-Poulat. Context dependent interpretation of multimodal speech-pointing gesture interface. In *Proceedings of the international conference on multimodal interfaces, Trento, Italy* (2005).

[13] S. Oviatt. Ten myths of multimodal interaction. Communications of the ACM **42**(11), 74 (1999).

[14] S. Oviatt, R. Lunsford, and R. Coulston. Individual differences in multimodal integration patterns: What are they and why do they exist? In *Conference on Human Factors in Computing Systems: Proceedings of the SIGCHI conference on Human factors in computing systems*, vol. 2, pp. 241–249 (2005).

[15] P. B. Kricos and S. A. Lesner. Differences in visual intelligibility across talkers. The Volta Review (1982).

[16] K. Sekiyama and Y. Tohkura. McGurk effect in non-English listeners: Few visual effects for Japanese subjects hearing Japanese syllables of high auditory intelligibility. The Journal of the Acoustical Society of America **90**, 1797 (1991).

[17] M. Kavakli and K. Nasser. Impacts of culture on gesture based interfaces: a case study on Anglo-Celtics and Latin Americans (2012).

[18] A. Mulac and T. Lundell. Linguistic contributors to the gender-linked language effect. Journal of Language and Social Psychology **5**(2), 81 (1986).

[19] D. Biber, S. Conrad, R. Reppen, and G. LEECH. Corpus linguistics: Investigating language structure and use. International journal of corpus linguistics **4**(1), 185 (1999).

[20] A. Mulac, J. Bradac, and P. Gibbons. Empirical support for the gender-as-culture hypothesis. Human Communication Research **27**(1), 121 (2001).

[21] M. Mehl and J. Pennebaker. The sounds of social life: a psychometric analysis of students' daily social environments and natural conversations. Journal of personality and social psychology **84**(4), 857 (2003).

[22] M. Newman, C. Groom, L. Handelman, and J. Pennebaker. Gender differences in language use: An analysis of 14,000 text samples. Discourse Processes **45**(3), 211 (2008).

[23] P. Bernardis and M. Gentilucci. Speech and gesture share the same communication system. Neuropsychologia **44**(2), 178 (2006).

[24] J. A. Rodger and P. C. Pendharkar. A field study of the impact of gender and user's technical experience on the performance of voice-activated medical tracking application. International Journal of Human-Computer Studies **60**(5), 529 (2004).

[25] M. Benzeghiba, R. De Mori, O. Deroo, S. Dupont, T. Erbes, D. Jouvet, L. Fissore, P. Laface, A. Mertins, C. Ris, *et al.* Automatic speech recognition and speech variability: A review. Speech Communication **49**(10), 763 (2007).

[26] L. Måhl. Speech recognition and adaptation experiments on childrens speech. KTH, Stockholm, Sweden (2003).

[27] W. D. Andrews, M. A. Kohler, J. P. Campbell, J. J. Godfrey, and J. Hernández-Cordero. Gender-dependent phonetic refraction for speaker recognition. In *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, vol. 1, pp. I–149 (IEEE, 2002).

[28] S. HOMMA and A. KOBAYASHI. Online speech detection and dual-gender speech recognition for captioning broadcast news. IEICE TRANSACTIONS on Information and Systems **90**(8), 1286 (2007).

[29] W. Abdulla, N. Kasabov, and D.-N. Zealand. Improving speech recognition performance through gender separation. Changes **9**, 10 (2001).

[30] K. Wu and D. G. Childers. Gender recognition from speech. Part I: Coarse analysis. The journal of the Acoustical society of America **90**, 1828 (1991).

[31] D. G. Childers and K. Wu. Gender recognition from speech. Part II: Fine analysis. The Journal of the Acoustical society of America **90**, 1841 (1991).

[32] S. Slomka and S. Sridharan. Automatic gender identification optimised for language independence. In *TENCON'97. IEEE Region 10 Annual Conference. Speech and Image Technologies for Computing and Telecommunications., Proceedings of IEEE*, vol. 1, pp. 145–148 (IEEE, 1997).

[33] M. H. Bahari *et al.* Age and Gender Recognition from Speech Patterns Based on Supervised Non-Negative Matrix Factorization. status: published pp. 3–5 (2011).

[34] T. Bocklet, A. Maier, J. G. Bauer, F. Burkhardt, and E. Noth. Age and gender recognition for telephone applications based on gmm supervectors and support vector machines. In *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pp. 1605–1608 (IEEE, 2008).

[35] K. Malhotra and A. Khosla. Automatic identification of gender & accent in spoken Hindi utterances with regional Indian accents. In *Spoken Language Technology Workshop, 2008. SLT 2008. IEEE*, pp. 309–312 (IEEE, 2008).

[36] K. Rakesh, S. Dutta, and K. Shama. Gender recognition using speech processing techniques in LABVIEW. International Journal of Advances in Engineering & Technology **1**(2), 51 (2011).

[37] C. Kramer. Perceptions of female and male speech. Language and Speech **20**(2), 151 (1977).

[38] J. Freeman. *Women: A feminist perspective* (ERIC, 1984).

[39] P. Morrel-Samuels and R. Krauss. Word familiarity predicts temporal asynchrony of hand gestures and speech. Journal of Experimental Psychology: Learning, Memory, and Cognition **18**(3), 615 (1992).

[40] D. McNeill. *Hand and Mind: what Gestures Reveal about Thought* (University of Chicago Press, 1992).

[41] R. Krauss, Y. Chen, and P. Chawla. Nonverbal behavior and nonverbal communication: What do conversational hand gestures tell us? Advances in experimental social psychology **28**, 389 (1996).

[42] R. Krauss, Y. Chen, and R. Gotfexnum. Lexical gestures and lexical access: a process model. Language and gesture **2**, 261 (2000).

[43] B. Butterworth and G. Beattie. Gestures and silence as indicators of planning in speech. in Recent Advances in the Psychology of Language: Formal and Experimental Approaches, ed. (1978).

[44] K. Chui. Temporal patterning of speech and iconic gestures in conversational discourse. Journal of pragmatics **37**(6), 871 (2005).

[45] J. Blake, D. Myszczyszyn, A. Jokel, and N. Bebiroglu. Gestures accompanying speech in specifically language-impaired children and their timing with speech. First Language **28**(2), 237 (2008).

[46] G. Ferré. Timing relationships between speech and co-verbal gestures in spontaneous French. In *Proceedings of LREC: Workshop on Multimodal Corpora*, vol. 6, pp. 86–91 (2010).

[47] A. D. Baddeley. *Essentials of human memory*, vol. 1 (Psychology Press Hove, 1999).

[48] P. C. John Sweller, Sharon K. Tindall-ford. When Two Sensory Modes are Better Than One. Journal of Experimental Psychology: Applied **3**(4) (1997).

[49] C. D. Wickens. Multiple resources and performance prediction. Theoretical issues in ergonomics science **3**(2), 159 (2002).

[50] G. Miller. The magical number seven, plus or minus two: Some limits on our capacity for processing information. The psychological review **63**, 81 (1956).

[51] J.-P. Thiran, H. Bourlard, and F. Marqués. *Multi-Modal Signal Processing: Methods and Techniques to Build Multimodal Interactive Systems* (Academic Press, 2009).

[52] L. Mulder. Measurement and analysis methods of heart rate and respiration for use in applied environments. Biological psychology **34**(2), 205 (1992).

[53] R. Ping and S. Goldin-Meadow. Gesturing saves cognitive resources when talking about nonpresent objects. Cognitive Science **34**(4), 602 (2010).

[54] S. Oviatt, R. Coulston, S. Tomko, B. Xiao, R. Lunsford, M. Wesson, and L. Carmichael. Toward a theory of organized multimodal integration patterns during human-computer interaction. In *Proceedings of the 5th international conference on Multimodal interfaces*, pp. 44–51 (ACM, 2003).

[55] S. Oviatt. Human-centered design meets cognitive load theory: designing interfaces that help people think. In *Proceedings of the 14th annual ACM international conference on Multimedia*, pp. 871–880 (ACM, 2006).

[56] L. Marstaller and H. Burianová. Individual differences in the gesture effect on working memory. Psychonomic bulletin & review pp. 1–5 (2013).

[57] A. Moir and D. Jessel. *Brain sex : the real difference between men and women* (Laurel, New York :, 1992).

[58] C. T. Gauthier, M. Duyme, M. Zanca, C. Capron, *et al.* Sex and performance level effects on brain activation during a verbal fluency task: A functional magnetic resonance imaging study. Cortex **45**(2), 164 (2009).

[59] I. E. Sommer, A. Aleman, A. Bouma, and R. S. Kahn. Do women really have more bilateral language representation than men? A meta-analysis of functional imaging studies. Brain **127**(8), 1845 (2004).

[60] M. Suwa, J. Gero, and T. Purcell. Analysis of cognitive processes of a designer as the foundation for support tools. In *Artificial Intelligence in Design*, vol. 98, pp. 229–248 (1998).

[61] S. Oviatt. Advances in robust multimodal interface design. IEEE Computer Graphics and Applications **23**(5), 62 (2003).

[62] A. Erol, G. Bebis, M. Nicolescu, R. Boyle, and X. Twombly. Vision-based hand pose estimation: A review. Computer Vision and Image Understanding **108**(1-2), 52 (2007).

[63] S. Oviatt. Mulitmodal interactive maps: Designing for human performance. Human-Computer Interaction **12**(1), 93 (1997).

[64] M. Vo and A. Waibel. Multi-modal HCI: combination of gesture and speech recognition. In *INTERACT'93 and CHI'93 conference companion on Human factors in computing systems*, pp. 69–70 (ACM, 1993).

[65] A. Salah. Perceptual Information Fusion in Humans and Machines. In *appears in Cognitive Neuroscience Forum* (2007).

[66] Y. Cao. *Multimodal information presentation for high-load human computer interaction* (University of Twente, 2011).

[67] Z. Obrenovic and D. Starcevic. Modeling multimodal human-computer interaction. Computer **37**(9), 65 (2004).

[68] S. Seneff, D. Goddeau, C. Pao, and J. Polifroni. Multimodal discourse modelling in a multi-user multi-domain environment. In *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, vol. 1, pp. 192–195 (IEEE, 1996).

[69] P. Cohen, M. Dalrymple, D. Moran, F. Pereira, J. Sullivan, R. Gargan, J. Schlossberg, and S. Tyler. Synergistic use of direct manipulation and natural language. Readings in intelligent user interfaces pp. 29–35 (1998).

[70] J. Siroux, M. Guyomard, F. Multon, and C. Remondeau. Modeling and processing of oral and tactile activities in the GEORAL system. Multimodal Human-Computer Communication pp. 101–110 (1998).

[71] W. Wahlster *et al.* User and discourse models for multimodal communication. Readings in intelligent user interfaces pp. 359–370 (1998).

[72] M. Vo and C. Wood. Building an application framework for speech and pen input integration in multimodal learning interfaces. In *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*, vol. 6, pp. 3545–3548 (IEEE, 1996).

[73] S. Oviatt, P. Cohen, L. Wu, L. Duncan, B. Suhm, J. Bers, T. Holzman, T. Winograd, J. Landay, J. Larson, *et al.* Designing the user interface for multimodal speech and pen-based gesture applications: state-of-the-art systems and future research directions. Human-computer interaction **15**(4), 263 (2000).

[74] M. Johnston, S. Bangalore, G. Vasireddy, A. Stent, P. Ehlen, M. Walker, S. Whittaker, and P. Maloor. MATCH: An architecture for multimodal dialogue systems. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pp. 376–383 (Association for Computational Linguistics, 2002).

[75] S. Dusan, G. Gadbois, and J. Flanagan. Multimodal interaction on PDAs integrating speech and pen inputs. Proceedings of EUROPSPEECH, Geneva, Switzerland pp. 2225–2228 (2003).

[76] P. Hui and H. Meng. Joint interpretation of input speech and pen gestures for multimodal human-computer interaction. In *Proc. Interspeech*, pp. 1197–1200 (2006).

[77] Y. Watanabe, K. Iwata, R. Nakagawa, K. Shinoda, and S. Furui. Semi-synchronous speech and pen input. In *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, vol. 4, pp. IV–409 (IEEE, 2007).

[78] A. Adler and R. Davis. Speech and sketching for multimodal design. In *ACM SIGGRAPH 2007 courses*, p. 14 (ACM, 2007).

[79] E. Vatikiotis-Bateson, K. Munhall, M. Hirayama, Y. Lee, and D. Terzopoulos. The dynamics of audiovisual behavior in speech. NATO ASI SERIES F COMPUTER AND SYSTEMS SCIENCES **150**, 221 (1996).

[80] D. Stork and M. Hennecke. *Speechreading by humans and machines: models, systems, and applications*, vol. 150 (Springer, 1996).

[81] 3Gear systems. [Online; accessed January-2013], URL http://www.threegear.com/getStarted.html.

[82] L. Hoste, B. Dumas, and B. Signer. SpeeG: a multimodal speech-and gesture-based text input solution. In *Proceedings of the International Working Conference on Advanced Visual Interfaces*, pp. 156–163 (ACM, 2012).

[83] T. Holzman. Computer-human interface solutions for emergency medical care. Interactions **6**(3), 13 (1999).

[84] X. Huang, A. Acero, C. Chelba, L. Deng, D. Duchene, J. Goodman, H. Hon, D. Jacoby, L. Jiang, R. Loynd, *et al.* MiPad: A next generation PDA prototype. In *Proc. ICSLP-2000*, pp. 33–36 (2000).

[85] P. E. Brooke N.M. Seeing speech: Investigations into the synthesis and recognition of visible speech movements using automatic image processing and computer graphics. Proceedings International Conference Speech Input and Output: Techniques and Applications **258**, 104 (1986).

[86] L. E. Bernstein and C. Benoit. For Speech Perception By Humans Or Machines, Three Senses Are Better Than One. Proceedings of the International Conference on Spoken Language Processing, (ICSLP 96) **3**, 1477 (1996).

[87] Q. Summerfield. Lipreading and audio-visual speech perception. Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences **335**(1273), 71 (1992).

[88] F. N. . D. B. M. Taylor. *The Structure of Multimodal Dialogue II.* (Amsterdam: John Benjamins, 2000).

[89] Q. Su and P. Silsbee. Robust audiovisual integration using semicontinuous Hidden Markov Models. In *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, vol. 1, pp. 42–45 (IEEE, 1996).

[90] M. Tomlinson, M. Russell, and N. Brooke. Integrating audio and visual information to provide highly robust speech recognition. In *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*, vol. 2, pp. 821–824 (IEEE, 1996).

[91] R. Frischholz and U. Dieckmann. BiolD: a multimodal biometric identification system. Computer **33**(2), 64 (2000).

[92] T. Wark, S. Sridharan, and V. Chandran. The use of temporal speech and lip information for multi-modal speaker identification via multi-stream HMMs. In *Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on*, vol. 6, pp. 2389–2392 (IEEE, 2000).

[93] S. Tamura, K. Iwano, and S. Furui. Multi-modal speech recognition using optical-flow analysis for lip images. The Journal of VLSI Signal Processing **36**(2), 117 (2004).

[94] H. Çetingül, E. Erzin, Y. Yemez, and A. Tekalp. Multimodal speaker/speech recognition using lip motion, lip texture and audio. Signal processing **86**(12), 3549 (2006).

[95] D. McNeill. *Language and gesture*, vol. 2 (Cambridge University Press, 2000).

[96] J. LaViola. A survey of hand posture and gesture recognition techniques and technology. Brown University, Providence, RI (1999).

[97] F. Chen, C. Fu, and C. Huang. Hand gesture recognition using a real-time tracking method and hidden Markov models. Image and Vision Computing **21**(8), 745 (2003).

[98] F. Parvini, D. McLeod, C. Shahabi, B. Navai, B. Zali, and S. Ghandeharizadeh. An approach to glove-based gesture recognition. Human-Computer Interaction. Novel Interaction Methods and Techniques pp. 236–245 (2009).

[99] M. Ganzeboom. How hand gestures are recognized using a dataglove. Human Media Interaction (HMI) (2009).

[100] L. Dipietro, A. Sabatini, and P. Dario. A survey of glove-based systems and their applications. Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on **38**(4), 461 (2008).

[101] A. Corradini, R. Wesson, and P. Cohen. A map-based system using speech and 3d gestures for pervasive computing. In *Proceedings of the 4th IEEE International Conference on Multimodal Interfaces*, p. 191 (IEEE Computer Society, 2002).

[102] I. Rauschert, P. Agrawal, R. Sharma, S. Fuhrmann, I. Brewer, and A. MacEachren. Designing a human-centered, multimodal GIS interface to support emergency management. In *Proceedings of the 10th ACM international symposium on Advances in geographic information systems*, pp. 119–124 (ACM, 2002).

[103] R. Sharma, M. Yeasin, N. Krahnstoever, I. Rauschert, G. Cai, I. Brewer, A. M. MacEachren, and K. Sengupta. Speech-gesture driven multimodal interfaces for crisis management. Proceedings of the IEEE **91**(9), 1327 (2003).

[104] D. Demirdjian, T. Ko, and T. Darrell. Untethered gesture acquisition and recognition for virtual world manipulation. Virtual Reality **8**(4), 222 (2005).

[105] M. Paleari and C. Lisetti. Toward multimodal fusion of affective cues. pp. 99–108 (ACM New York, NY, USA, 2006).

[106] S. Oviatt, P. Cohen, L. Wu, L. Duncan, B. Suhm, J. Bers, T. Holzman, T. Winograd, J. Landay, and J. Larson. Designing the user interface for multimodal speech and pen-based gesture applications: State-of-the-art systems and future research directions. Human-Computer Interaction **15**(4), 263 (2000).

[107] L. Yong-Hee, L. Dong-Woo, C. Eun-Jung, and P. Jun-Seok. Design of Multimodal Interface Framework. In *Advanced Communication Technology, The 9th International Conference on*, vol. 1, pp. 345–348 (Phoenix Park, Korea, 2007).

[108] I. Wachsmuth. Communicative rhythm in gesture and speech. ADVANCES IN CONSCIOUSNESS RESEARCH **35**, 117 (2002).

[109] L. Nigay and J. Coutaz. A design space for multimodal systems: concurrent processing and data fusion. In *Proceedings of the INTERACT '93 and CHI '93 Conference on Human Factors in Computing Systems*, pp. 172–178 (ACM, 1993).

[110] B. Xiao, C. Girand, and S. Oviatt. Multimodal integration patterns in children. In *Proceedings of 7th International Conference on Spoken Language Processing*, pp. 629–632 (2002).

[111] B. Xiao, R. Lunsford, R. Coulston, M. Wesson, and S. Oviatt. Modeling multimodal integration patterns and performance in seniors: Toward adaptive processing of individual differences. In *Proceedings of the 5th international conference on Multimodal interfaces*, pp. 265–272 (ACM, 2003).

[112] A. K. Gupta and T. Anastasakos. Dynamic time windows for multimodal input fusion. In *Proc. 8th International Conference on Spoken Language Processing (INTERSPEECH 2004-ICSLP), Jeju, Korea*, pp. 1009–1012 (2004).

[113] E. R. Kandel, J. H. Schwartz, T. M. Jessell, *et al. Principles of neural science*, vol. 4 (McGraw-Hill New York, 2000).

[114] S. D. Kelly, A. Özyürek, and E. Maris. Two Sides of the Same Coin Speech and Gesture Mutually Interact to Enhance Comprehension. Psychological Science **21**(2), 260 (2010).

[115] S. Y. Mousavi, R. Low, J. Sweller, *et al.* Reducing cognitive load by mixing auditory and visual presentation modes. Journal of educational psychology **87**(2), 319 (1995).

[116] C. D. Wickens, D. L. Sandry, and M. Vidulich. Compatibility and resource competition between modalities of input, central processing, and output. Human Factors: The Journal of the Human Factors and Ergonomics Society **25**(2), 227 (1983).

[117] S. K. Card, T. P. Moran, and A. Newell. The model human processor: An engineering model of human performance. Handbook of Human Perception **2** (1986).

[118] T. S. Jastrzembski and N. Charness. The model human processor and the older adult: Parameter estimation and validation within a mobile phone task. Journal of Experimental Psychology Applied **13**(4), 224 (2007).

[119] S. Oviatt, R. Coulston, and R. Lunsford. When do we interact multimodally?: cognitive load and multimodal communication patterns. In *Proceedings of the 6th international conference on Multimodal interfaces*, pp. 129–136 (ACM, 2004).

[120] S. Goldin-Meadow, H. Nusbaum, S. D. Kelly, and S. Wagner. Explaining math: Gesturing lightens the load. Psychological Science **12**(6), 516 (2001).

[121] E. Morsella and R. M. Krauss. The role of gestures in spatial working memory and speech. The American Journal of Psychology pp. 411–424 (2004).

[122] S. W. Cook, T. K. Yip, and S. Goldin-Meadow. Gestures, but not meaningless movements, lighten working memory load when explaining math. Language and Cognitive Processes **27**(4), 594 (2012).

[123] A. Herlitz, L.-G. Nilsson, and L. Bäckman. Gender differences in episodic memory. Memory & cognition **25**(6), 801 (1997).

[124] D. C. Geary, S. J. Saults, F. Liu, and M. K. Hoard. Sex differences in spatial cognition, computational fluency, and arithmetical reasoning. Journal of Experimental Child Psychology **77**(4), 337 (2000).

[125] J. S. Hyde. How large are cognitive gender differences? A meta-analysis using! w^2 and d.. American Psychologist **36**(8), 892 (1981).

[126] J. S. Hyde, E. Fennema, and S. J. Lamon. Gender differences in mathematics performance: a meta-analysis. Psychological bulletin **107**(2), 139 (1990).

[127] U. Hadar. Two types of gesture and their role in speech production. Journal of Language and Social Psychology **8**(3-4), 221 (1989).

[128] U. Hadar, D. Wenkert-Olenik, R. Krauss, and N. Soroker. Gesture and the processing of speech: Neuropsychological evidence. Brain and language **62**(1), 107 (1998).

[129] W. Levelt, G. Richardson, and W. La Heij. Pointing and voicing in deictic expressions. Journal of Memory and Language **24**(2), 133 (1985).

[130] R. Krauss. Why do we gesture when we speak? Current Directions in Psychological Science **7**(2), 54 (1998).

[131] R. M. Krauss and U. Hadar. The role of speech-related arm/hand gestures in word retrieval. Gesture, speech, and sign pp. 93–116 (1999).

[132] D. McNeill. So you think gestures are nonverbal. Psychological review **92**(3), 350 (1985).

[133] D. P. Loehr. *Gesture and intonation*. Ph.D. thesis, Georgetown University (2004).

[134] L. Wu, S. Oviatt, and P. Cohen. Multimodal integration-a statistical view. Multimedia, IEEE Transactions on **1**(4), 334 (1999).

[135] T. Reiman. http://www. bodylanguageuniversity. com (2013).

[136] A. Moir and D. Jessel. *Brain sex: The real difference between men and women*, vol. 149 (Mandarin, 1989).

[137] K. Payne. *Different but equal: Communication between the sexes* (Praeger Publishers, 2001).

[138] A. Mulac, L. Studley, and S. Blau. The gender-linked language effect in primary and secondary students' impromptu essays. Sex Roles **23**(9), 439 (1990).

[139] S. Cohen. Gender Differences in Speech Temporal Patterns Detected Using Lagged Co-occurrence Text-analysis of Personal Narratives. Journal of psycholinguistic research **38**(2), 111 (2009).

[140] N. Briton and J. Hall. Beliefs about female and male nonverbal communication. Sex Roles **32**(1), 79 (1995).

[141] E. V. College. GENDER DIFFERENCE IN NONVERBAL COMMUNICATION. http://evc-cit.info/psych018/observation_samples/observation9028.pdf. [Online; accessed August-2012].

[142] D. M. Saucier and L. J. Elias. Lateral and sex differences in manual gesture during conversation. Laterality: Asymmetries of Body, Brain and Cognition **6**(3), 239 (2001).

[143] H. Cochet and J. Vauclair. Hand preferences in human adults: Non-communicative actions versus communicative gestures. cortex **48**(8), 1017 (2012).

[144] L. Beckwith and M. Burnett. Gender: An important factor in end-user programming environments? In *Visual Languages and Human Centric Computing, 2004 IEEE Symposium on*, pp. 107–114 (IEEE, 2004).

[145] T. Busch. Gender differences in self-efficacy and attitudes toward computers. Journal of Educational Computing Research **12**(2), 147 (1995).

[146] M. Czerwinski, D. Tan, and G. Robertson. Women take a wider view. In *Proceedings of the SIGCHI conference on Human factors in computing systems: Changing our world, changing ourselves*, pp. 195–202 (ACM, 2002).

[147] D. Tan, M. Czerwinski, and G. Robertson. Women go with the (optical) flow. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 209–215 (ACM, 2003).

[148] J. Cassell *et al.* Genderizing HCI. The Handbook of Human-Computer Interaction. Mahwah, NJ: Lawrence Erlbaum pp. 402–411 (2002).

[149] J. Brewer and A. Bassoli. Reflections of gender, reflections on gender: Designing ubiquitous computing technologies. In *Gender & Interaction: Real and Virtual Women in a Male World, Workshop at AVI*, pp. 9–12 (2006).

[150] T. L. Perry, R. N. Ohde, and D. H. Ashmead. The acoustic bases for gender identification from childrens voices. The Journal of the Acoustical Society of America **109**, 2988 (2001).

[151] M. Feld, F. Burkhardt, and C. Müller. Automatic speaker age and gender recognition in the car for tailoring dialog and mobile services. Proc. of Interspeech 2010 pp. 2834–2837 (2010).

[152] T. Bocklet, G. Stemmer, V. Zeissler, and E. Nöth. Age and Gender Recognition Based on Multiple Systems Early vs. Latefusion. In *Proceedings of the 11th Annual Conference of the International Speech Communication Association*, pp. 2830–2833 (2010).

[153] M. Li, C.-S. Jung, and K. J. Han. Combining five acoustic level modeling methods for automatic speaker age and gender recognition. In *Proc. Interspeech* (2010).

[154] K. P. Cosgrove, C. M. Mazure, and J. K. Staley. Evolving knowledge of sex differences in brain structure, function, and chemistry. Biological psychiatry **62**(8), 847 (2007).

[155] E. Luders, K. L. Narr, P. M. Thompson, D. E. Rex, L. Jancke, H. Steinmetz, and A. W. Toga. Gender differences in cortical complexity. Nature neuroscience **7**(8), 799 (2004).

[156] J. McGlone. Sex differences in human brain asymmetry: A critical survey. Behavioral and Brain Sciences **3**(2), 215 (1980).

[157] B. A. Shaywitz, S. E. Shaywltz, K. R. Pugh, R. T. Constable, P. Skudlarski, R. K. Fulbright, R. A. Bronen, J. M. Fletcher, D. P. Shankweiler, L. Katz, *et al.* Sex differences in the functional organization of the brain for language (1995).

[158] K. Kansaku, A. Yamaura, and S. Kitazawa. Sex differences in lateralization revealed in the posterior language areas. Cerebral Cortex **10**(9), 866 (2000).

[159] R. C. Gur, D. Alsop, D. Glahn, R. Petty, C. L. Swanson, J. A. Maldjian, B. I. Turetsky, J. A. Detre, J. Gee, and R. E. Gur. An fMRI study of sex differences in regional activation to a verbal and a spatial task. Brain and language **74**(2), 157 (2000).

[160] E. Weiss, C. Siedentopf, A. Hofer, E. Deisenhammer, M. Hoptman, C. Kremser, S. Golaszewski, S. Felber, W. Fleischhacker, and M. Delazer. Sex differences in brain activation pattern during a visuospatial cognitive task: a functional

magnetic resonance imaging study in healthy volunteers. Neuroscience letters **344**(3), 169 (2003).

[161] J. A. Frost, J. R. Binder, J. A. Springer, T. A. Hammeke, P. S. Bellgowan, S. M. Rao, and R. W. Cox. Language processing is strongly left lateralized in both sexes Evidence from functional MRI. Brain **122**(2), 199 (1999).

[162] E. Weiss, C. Siedentopf, A. Hofer, E. Deisenhammer, M. Hoptman, C. Kremser, S. Golaszewski, S. Felber, W. Fleischhacker, and M. Delazer. Brain activation pattern during a verbal fluency test in healthy male and female volunteers: a functional magnetic resonance imaging study. Neuroscience Letters **352**(3), 191 (2003).

[163] M. Anusuya and S. Katti. Speech recognition by machine, A review. arXiv preprint arXiv:1001.2267 (2010).

[164] B.-H. Juang, W. Hou, and C.-H. Lee. Minimum classification error rate methods for speech recognition. Speech and Audio Processing, IEEE Transactions on **5**(3), 257 (1997).

[165] V. I. Pavlovic, R. Sharma, and T. S. Huang. Visual interpretation of hand gestures for human-computer interaction: A review. Pattern Analysis and Machine Intelligence, IEEE Transactions on **19**(7), 677 (1997).

[166] F. Chen, E. H. Choi, and N. Wang. Exploiting speech-gesture correlation in multimodal interaction. In *Human-Computer Interaction. HCI Intelligent Multimodal Interaction Environments*, pp. 23–30 (Springer, 2007).

[167] M. Kipp, M. Neff, and I. Albrecht. An Annotation Scheme for Conversational Gestures: How to economically capture timing and form. Language Resources and Evaluation **41**(3-4), 325 (2007).

[168] M. Kipp, M. Neff, K. H. Kipp, and I. Albrecht. Towards natural gesture synthesis: Evaluating gesture units in a data-driven approach to gesture synthesis. In *Intelligent Virtual Agents*, pp. 15–28 (Springer, 2007).

[169] M. Neff, M. Kipp, I. Albrecht, and H.-P. Seidel. Gesture modeling and animation based on a probabilistic re-creation of speaker style. ACM Transactions on Graphics (TOG) **27**(1), 5 (2008).

[170] D. McNeill. *Gesture and thought* (University of Chicago Press, 2008).

[171] M. Kipp. ANVIL - A Generic Annotation Tool for Multimodal Dialogue (2001).

[172] A. Heloir, M. Neff, and M. Kipp. Exploiting Motion Capture for Virtual Human Animation. In *Proceedings of the Workshop Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality at LREC-2010* (2010).

[173] P. Boersma and D. Weenink. Praat doing phonetics by computer. [Online; accessed March-2013], URL http://www.praat.org/.

[174] L. Campbell and M. Wanderley. The Observation of Movement. IDMIL Report. Disponible en http://www. idmil. org/home (2005).

[175] G. Castellano, I. Leite, A. Pereira, C. Martinho, A. Paiva, and P. McOwan. It's all in the game: Towards an affect sensitive and context aware game companion. In *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on*, pp. 1–8 (IEEE, 2009).

[176] M. H. M. Saad, A. Hussain, L. X. Loong, W. N. A. Baharuddin, and N. M. Tahir. Event description from video stream for anomalous human activity and behaviour detection. In *Signal Processing and its Applications (CSPA), 2011 IEEE 7th International Colloquium on*, pp. 503–506 (IEEE, 2011).

[177] S. Kita, I. Van Gijn, and H. Van der Hulst. Movement phases in signs and co-speech gestures, and their transcription by human coders. Gesture and sign language in human-computer interaction pp. 23–35 (1998).

[178] Y. Yasinnik, M. Renwick, and S. Shattuck-Hufnagel. The timing of speech-accompanying gestures with respect to prosody. Proceedings of Sound to Sense, MIT (2004).

[179] M. E. Sargin, F. Ofli, Y. Yasinnik, O. Aran, A. Karpov, S. Wilson, E. Erzin, Y. Yemez, and A. M. Tekalp. Combined gesture-speech analysis and synthesis. Proc. of the eNTERFACE **5** (2005).

[180] J. Liu and M. Kavakli. Temporal Relation between Speech and Co-verbal Iconic Gestures in Multimodal Interface Design (2011).

[181] R. L. Brennan and D. J. Prediger. Coefficient kappa: Some uses, misuses, and alternatives. Educational and psychological measurement **41**(3), 687 (1981).

[182] M. Kipp. Multimedia annotation, querying and analysis in ANVIL. Multimedia Information Extraction **107** (2010).

[183] K. Ericsson and H. Simon. Protocol analysis: Verbal reports as data (rev. ed.) MIT Press. Cambridge, MA (1993).

[184] Ö. Akin and C. Lin. Design protocol data and novel design decisions. Design Studies **16**(2), 211 (1995).

[185] J. S. Gero and T. Mc Neill. An approach to the analysis of design protocols. Design studies **19**(1), 21 (1998).

[186] M. Kavakli and J. S. Gero. The structure of concurrent cognitive actions: A case study on novice and expert designers. Design Studies **23**(1), 25 (2002).

[187] N. Warner, M. Letsky, and M. Cowen. Cognitive model of team collaboration: Macro-cognitive focus. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 49, pp. 269–273 (SAGE Publications, 2005).

[188] K. A. Ericsson. Protocol analysis and expert thought: Concurrent verbalizations of thinking during experts performance on representative tasks. The Cambridge handbook of expertise and expert performance pp. 223–241 (2006).

[189] H. Lundgrén-Laine and S. Salanterä. Think-aloud technique and protocol analysis in clinical decision-making research. Qualitative Health Research **20**(4), 565 (2010).

[190] J. S. Gero and H.-H. Tang. The differences between retrospective and concurrent protocols in revealing the process-oriented aspects of the design process. Design Studies **22**(3), 283 (2001).

[191] M. T. Chi. Quantifying qualitative analyses of verbal data: A practical guide. The journal of the learning sciences **6**(3), 271 (1997).

[192] K. Dorst and J. Dijkhuis. Comparing paradigms for describing design activity. Design Studies **16**(2), 261 (1995).

[193] M. Tovey, S. Porter, and R. Newman. Sketching, concept development and automotive design. Design studies **24**(2), 135 (2003).

[194] Y. Jin and P. Chusilp. Study of mental iteration in different design situations. Design studies **27**(1), 25 (2006).

[195] E. T. Kazmierczak. A semiotic perspective on aesthetic preferences, visual literacy, and information design. Information design journal **10**(2), 176 (2001).

[196] C. Giorgis, N. J. Johnson, A. Bonomo, C. Colbert, A. Conner, G. Kauffman, and D. Kulesza. Visual Literacy. Reading Teacher **53**(2), 146 (1999).

[197] A. Bamford. *The visual literacy white paper* (Adobe Systems, 2003).

[198] K. Stamps and Y. Hamam. Towards Inexpensive BCI Control for Wheelchair Navigation in the Enabled Environment C A Hardware Survey **6334**, 336 (2010).

[199] S. Debener, F. Minow, R. Emkes, K. Gandras, and M. Vos. How about taking a low-cost, small, and wireless EEG for a walk? Psychophysiology **49**(11), 1617 (2012).

[200] R. N. Khushaba, L. Greenacre, S. Kodagoda, J. Louviere, S. Burke, and G. Dissanayake. Choice modeling and the brain: A study on the Electroencephalogram (EEG) of preferences. Expert Systems with Applications (2012).

[201] P. Bobrov, A. Frolov, C. Cantor, I. Fedulova, M. Bakhnyan, and A. Zhavoronkov. Brain-computer interface based on generation of visual images. PloS one **6**(6), e20674 (2011).

[202] R. Lievesley, M. Wozencroft, and D. Ewins. The Emotiv EPOC neuroheadset: an inexpensive method of controlling assistive technologies using facial expressions and thoughts? Journal of Assistive Technologies **5**(2), 67 (2011).

[203] H. Ekanayake. P300 and Emotiv EPOC: Does Emotiv EPOC capture real EEG? [Online; accessed June-2013], URL http://neurofeedback.visaduma.info/emotivresearch.htm.

[204] K. Stytsenko, E. Jablonskis, and C. Prahm. Evaluation of consumer EEG device Emotiv EPOC. In *MEi: CogSci Conference 2011, Ljubljana* (2011).

[205] N. V. Thakor and D. L. Sherman. EEG Signal Processing: Theory and Applications. In *Neural Engineering*, pp. 259–303 (Springer, 2013).

[206] T. Limpiti, B. D. Van Veen, H. T. Attias, and S. S. Nagarajan. A spatiotemporal framework for estimating trial-to-trial amplitude variation in event-related MEG/EEG. Biomedical Engineering, IEEE Transactions on **56**(3), 633 (2009).

[207] A. Shahidi Zandi, R. Tafreshi, M. Javidan, and G. A. Dumont. Predicting temporal lobe epileptic seizures based on zero-crossing interval analysis in scalp EEG. In *Engineering in Medicine and Biology Society (EMBC), 2010 Annual International Conference of the IEEE*, pp. 5537–5540 (IEEE, 2010).

[208] F. Lopes da Silva. EEG analysis: theory and practice. Electroencephalography: Basic Principles, Clinical Applications and Related Fields, 4th edition pp. 1135–1163 (1998).

[209] N. J. Davis, S. P. Tomlinson, and H. M. Morgan. The role of beta-frequency neural oscillations in motor control. The Journal of Neuroscience **32**(2), 403 (2012).

[210] R. N. Khushaba, C. Wise, S. Kodagoda, J. Louviere, B. E. Kahn, and C. Townsend. Consumer neuroscience: Assessing the brain response to marketing stimuli using electroencephalogram (EEG) and eye tracking. Expert Systems with Applications (2013).

[211] Further process to the raw EEG data. [Online; accessed June-2013], URL http://www.emotiv.com/forum/messages/forum10/topic524/message2927/?phrase_id=545428#message2927.

[212] P. Comon. Independent component analysis, a new concept? Signal processing **36**(3), 287 (1994).

[213] A. J. Bell and T. J. Sejnowski. An information-maximization approach to blind separation and blind deconvolution. Neural computation **7**(6), 1129 (1995).

[214] S. Makeig, A. J. Bell, T.-P. Jung, T. J. Sejnowski, *et al.* Independent component analysis of electroencephalographic data. Advances in neural information processing systems pp. 145–151 (1996).

[215] T.-P. Jung, S. Makeig, M. J. McKeown, A. J. Bell, T.-W. Lee, and T. J. Sejnowski. Imaging brain dynamics using independent component analysis. Proceedings of the IEEE **89**(7), 1107 (2001).

[216] T.-P. Jung, S. Makeig, C. Humphries, T.-W. Lee, M. J. Mckeown, V. Iragui, and T. J. Sejnowski. Removing electroencephalographic artifacts by blind source separation. Psychophysiology **37**(2), 163 (2000).

[217] S. Makeig, T.-P. Jung, D. Ghahremani, and T. J. Sejnowski. Independent component analysis of simulated ERP data. Institute for Neural Computation, University of California: technical report INC-9606 (1996).

[218] B. Saltzberg, W. Burton Jr, J. Barlow, and N. Burch. Moments of the power spectral density estimated from samples of the autocorrelation function (A robust

procedure for monitoring changes in the statistical properties of lengthy non-stationary time series such as the EEG). Electroencephalography and clinical neurophysiology **61**(1), 89 (1985).

[219] S. R. Safavian and D. Landgrebe. A survey of decision tree classifier methodology. Systems, Man and Cybernetics, IEEE Transactions on **21**(3), 660 (1991).

[220] A. Camurri, B. Mazzarino, M. Ricchetti, R. Timmers, and G. Volpe. Multimodal analysis of expressive gesture in music and dance performances. In *Gesture-based communication in human-computer interaction*, pp. 20–39 (Springer, 2004).

[221] T. Fawcett. An introduction to ROC analysis. Pattern recognition letters **27**(8), 861 (2006).

[222] T. D. Wager, K. L. Phan, I. Liberzon, and S. F. Taylor. Valence, gender, and lateralization of functional brain anatomy in emotion: a meta-analysis of findings from neuroimaging. Neuroimage **19**(3), 513 (2003).

[223] C. Irvine. Intelligence in men and women is a gray and white matter (2014). [Online; accessed June-2014], URL http://archive.today.uci.edu/news/release_detail.asp?key=1261.

Ethics Approval

Approved

Faculty of Science Research Office <sci.ethics@mq.edu.au> Fri, Apr 5, 2013 at 2:28 PM
To: Dr Manolya Kavakli-Thorne <manolya.kavakli@mq.edu.au>, Mr John Porte <john.porte@mq.edu.au>, Ms Jing Liu <jing.liu21@students.mq.edu.au>
Cc: Prof Richie Howitt <richie.howitt@mq.edu.au>, Ms Katherine Wilson <katherine.wilson@mq.edu.au>

Dear Dr Kavakli-Thorne,

RE: Ethics project entitled:"Gender differences in visual cognition for multimodal systems design"
Ref number: 5201300105.

Thank you for your recent correspondence. Your response has addressed the issues raised by the Faculty of Science Human Research Ethics Sub-Committee and you may now commence your research.

This research meets the requirements of the National Statement on Ethical Conduct in Human Research (2007). The National Statement is available at the following web site:

http://www.nhmrc.gov.au/_files_nhmrc/publications/attachments/e72.pdf.

The following personnel are authorised to conduct this research:
Dr Manolya Kavakli-Thorne
Jing Liu
John Porte

NB. STUDENTS: IT IS YOUR RESPONSIBILITY TO KEEP A COPY OF THIS APPROVAL EMAIL TO SUBMIT WITH YOUR THESIS.

Please note the following standard requirements of approval:

1. The approval of this project is conditional upon your continuing compliance with the National Statement on Ethical Conduct in Human Research (2007).

2. Approval will be for a period of five (5) years subject to the provision of annual reports.

Progress Report 1 Due: 5 April 2014
Progress Report 2 Due: 5 April 2015
Progress Report 3 Due: 5 April 2016
Progress Report 4 Due: 5 April 2017
Final Report Due: 5 April 2018

NB. If you complete the work earlier than you had planned you must submit a Final Report as soon as the work is completed. If the project has been discontinued or not commenced for any reason, you are also required to submit a Final Report for the project.

Progress reports and Final Reports are available at the following website:

http://www.research.mq.edu.au/for/researchers/how_to_obtain_ethics_approval/human_research_ethics/forms

3. If the project has run for more than five (5) years you cannot renew approval for the project. You will need to complete and submit a Final Report and submit a new application for the project. (The five year limit on renewal of approvals allows the Committee to fully re-review research in an environment where legislation, guidelines and requirements are continually changing, for example, new child protection and privacy laws).

4. All amendments to the project must be reviewed and approved by the Committee before implementation. Please complete and submit a Request for Amendment Form available at the following website:

http://www.research.mq.edu.au/for/researchers/how_to_obtain_ethics_approval/human_research_ethics/forms

5. Please notify the Committee immediately in the event of any adverse effects on participants or of any unforeseen events that affect the continued ethical acceptability of the project.

6. At all times you are responsible for the ethical conduct of your research in accordance with the guidelines established by the University. This information is available at the following websites:
http://www.mq.edu.au/policy/

http://www.research.mq.edu.au/for/researchers/how_to_obtain_ethics_approval/human_research_ethics/policy

If you will be applying for or have applied for internal or external funding for the above project it is your responsibility to provide the Macquarie University's Research Grants Management Assistant with a copy of this email as soon as possible. Internal and External funding agencies will not be informed that you have final approval for your project and funds will not be released until the Research Grants Management Assistant has received a copy of this email.

If you need to provide a hard copy letter of Final Approval to an external organisation as evidence that you have Final Approval, please do not hesitate to contact the Ethics Secretariat at the address below.

Please retain a copy of this email as this is your official notification of final ethics approval.

Yours sincerely,
Richie Howitt, Chair
Faculty of Science Human Research Ethics Sub-Committee
Macquarie University
NSW 2109