Gaze-Contingent Automatic Speech Recognition

A thesis submitted for the degree of

Doctor of Philosophy

Neil James Cooke

Department of Electronic, Electrical and Computer Engineering
University of Birmingham
December 2006

Neil James Cooke

Department of Electronic, Electrical and Computer Engineering
School of Engineering
University of Birmingham
Birmingham
B15 2TT
United Kingdom
http://www.eece.bham.ac.uk/

First draft: 24th March 2006
Second draft: 23rd August 2006
Submitted: 10th October 2006
Published: 15th December 2006

© 2006 All rights reserved.

Abstract

This study investigated recognition systems that combine loosely coupled modalities, integrating eye movements in an Automatic Speech Recognition (ASR) system as an exemplar. A probabilistic framework for combining modalities was formalised and applied to the specific case of integrating eye movement and speech. A corpus of a matched eye movement and related spontaneous conversational British English speech for a visual-based, goal-driven task was collected. This corpus enabled the relationship between the modalities to be verified. Robust extraction of visual attention from eye movement data was investigated using Hidden Markov Models and Hidden Semi-Markov Models. Gaze-contingent ASR systems were developed from a research-grade baseline ASR system by redistributing language model probability mass according to the visual attention. The best performing systems maintained the Word Error Rates but showed an increase in the Figure of Merit - a measure of the keyword spotting accuracy and integration success. The core values of this work may be useful for developing robust multimodal decoding system functions.

Acknowledgements

I want to thank the following people: Professor Martin Russell, for supervising this study and giving me the opportunity to persue an academic career; Professor Antje Meyer for allowing the use of facilities in the school of Psychology; Dr. Sridhar Pammu for technical guidance; Mrs Mary Winkles for her counsel; Colleagues for their help, comments and advice.

Finally I wish to thank my wife Helen, for her steadfast support and love.

Dedicated to Kay Cooke *née* Grant (1947-1998) and my son, Matthew James Cooke (2006-).

Contents

List of Figures

List of Tables

List of Abbreviations

ASR	Automatic Speech Recognition
BEEP	British English Example Pronunciation
BNC	British National Corpus
CDIF	Corpus Document Interchange Format
CMN	Cepstral Mean Normalisation
CSR	Continuous Speech Recognition
CUED	Cambridge University Engineering Department
DAG	Direct Acyclic Graph
DARPA	Defense Advanced Research Projects Agency
DBN	Dynamic Bayesian Network
DCT	Discrete Cosine Transform
DMI	Direct Manipulation Interface
DRM	Distributed Source Management software
EM	Expectation-Maximisation
ETRA	Eye Tracking Research Applications symposium
FA	False Alarms
FOM	Figure of Merit
FOVA	Focus Of Visual Attention
GMM	Gaussian Mixture Model
GUI	Graphical User Interface
HCI	Human Computer Interface
HMM	Hidden Markov Model
HSMM	Hidden Semi-Markov Model
HTK	Hidden markov model ToolKit
ICC	Intel C++ Compiler
IEEE	Institute of Electrical and Electronics Engineers
IPA	International Phonetic Alphabet
LDC	Lynguistic Data Consortium
MAP	Maximum-APosteriori
MFCC	Mel-Frequency Cepstral Coefficients
MLLR	Maximum Likelihood Linear Regression
MP3	Moving Picture experts group 3 standard
MPEG-4	Moving Picture Experts Group 4 standard
MRI	Magnetic Resonance Imaging

NIST	National Institute for Standards and Testing
PDF	Probability Density Function
PER	Perplexity
RP	Received Pronunciation
SGML	Standard Generalised Mark-up Language
SPHERE	SPeech HEader REsources
TP	True Positives
WER	Word Error Rate
WSJ0	Wall Street Journal '0' corpus
WSJCAM0	Wall Street Journal CAMbridge '0' corpus
XML	eXtensible Mark-up Language

1 Introduction

1.1 Multimodal interaction

Within the context of this thesis and multimodal interaction research, the following terms and definitions apply:

- *Modality*: A physiological sense used for communication. (e.g. voice, vision, eye movement or gesture)
- *Multimodal*: The consideration of two or more physiological modalities.
- *Integration*: The process by which modalities are combined.
- *Decode* and *Recognition*: The information extracted from modalities. (e.g. speech from voice; visual attention from vision;)

Multimodal research involves the joint consideration of multiple modalities in human-to-human and human-to-machine interaction. In computer engineering and science, the research focusses on realising systems and interfaces that understand and/or participate in, multi-party interaction. The systems and technologies being developed may go some way to realising Weiser's vision of 21st century 'ubiquitous computing' in which computers blend into the environment [Wei99].

From an information-theoretic standpoint, each physiological modality can be seen as transmitting or receiving information, which may or may not be duplicated in another [Ovi99a]. Conversely, system functions which do the same thing must be realised- e.g. for speaking and hearing there is Automatic Speech Recognition (ASR) and speech synthesis respectively; for vision there is image processing. Body movement and gestures (of which eye movement or gaze is an example) are recognised by machines (e.g. eye trackers for eye movement) and replicated using robotics and/or the dynamics of the display.

Extracting the meaning/semantics of multimodal communication is desirable. If the interaction is between a human and machine, then the information required by the machine is that needed to understand the intention of the user - be it the commands the person issues or their current attentive state [HKPH03]. If the interaction is between humans then the machine may be required to store and process information (e.g. multimodal meeting annotations [GBY+00]) or assist in the communication where necessary [Sat01].

In the context of multimodal research, this study considers the technologies used to recognised modalities. Since information from all modalities is available to a multimodal system, it contributes by exploring how the use of one modality (eye movement) can be used to improve recognition of another (speech) due to the information overlap.

1.2 Automatic Speech Recognition (ASR)

Automatic Speech Recognition (ASR) is the system function that extracts a word sequence from human speech. Developments in the last 30 years have resulted in systems which recognise human speech with sufficient accuracy for commercial use[1]. ASR development has drawn on, and contributed to, the field of pattern recognition which involves the classification of raw data by detecting patterns of interest within the data[2].

Currently (2006), the prevalent pattern recognition technique is to periodically extract a set of acoustic features from the speech signal and model the speech as a sequence of hidden Markov models (HMM) [Rab89]. The HMM is a probabilistic model for classifying time series data. HMMs are discussed in general terms in §2.4.3, in relation to ASR in Chapter 4 and to eye movement in Chapter 5.

Being probabilistic, HMM-based ASR cannot recognise speech with 100% accuracy. The HMM formalism assumes a discretised stochastic process whereas speech is a continuous deterministic[3] process. The speech signal is subject to channel noise and consequently may not contain enough information to enable an ASR system to decode speech. Research into systems that consider lip shape in addition to voice into ASR in the last decade had some success [SB96] [Tom96].

[1] For reviews on ASR see [You95] and [Pad02]

[2] For a review on pattern recognition see [Jai00]

[3] Since all the causes of speech and variability are not known, the stochastic assumption is desirable.

Lip movement and voice can be considered as tightly coupled modalities; the information each contains to enable speech recognition temporally co-occurs. In addition, the information contained in each is equivalent - i.e. both can be decoded separately to determine the word sequence. In contrast, loosely coupled modalities contain information that does not temporally co-occur but which would benefit joint consideration in the recognition process. The information contained in each modality is different - e.g. looking at a visual object alone does not enable a recogniser to infer a word sequence.

There is a gap in literature of examining the techniques for integrating loosely coupled modalities into ASR systems. This thesis contributes to this area by considering eye movement and speech as the loosely coupled modalities.

1.3 Eye tracking

Eye tracking is a technology which has many applications in various research fields, notably psychology, medicine and computer science/engineering. Eye tracking enables system awareness of a person's gaze in relation to a shared environment - giving an indication of the person's focus of attention. The environment may be real (i.e. looking at objects in a room) or virtual (i.e. looking at artefacts on a computer screen).

Eye tracking systems broadly fall into two categories - head mounted and remote. The head mounted eye trackers mostly use reflected light to track the eyes [GEN95]. Camera(s) suspended from a contraption mounted on the head capture video of the eye(s). The eye position is determined by shining a light source at the eyeball and measuring the distance between the light reflection and a feature of the eye (e.g. pupil). Head mounted trackers are accurate but can be intrusive and feel unnatural to wear, although component miniaturisation has extended their utility.

Remote eye trackers use multiple fixed cameras in the environment fixating on the face. The remote cameras are sensitive enough to capture images of the eyes and head position [SYW97]. Advances in image processing and computational power have made remote tracking more popular, although the quality of the tracking is inferior to head-mounted tracking.

Eye tracking is used in HCI fields for designing eye movement into interfaces and evaluating interface usability. Cognitive scientists use eye movement analysis to understanding brain processes [Ray98]. Medical researchers have used eye tracking

for characterising people's surgical skills [LAK+04]. Employing the eye as a direct means of control (e.g. Using the eye as a replacement for mouse) is cursed with the 'Midas touch' - the eyes are not under full conscious control and involuntary movements can lead to unsatisfactory results [Jac03]. Despite problems with the technology, people with disabilities have successfully made use of eye movement for keyboard replacement [MKJ02].

The primary information extracted from eye trackers is a person's eye position in relation their field of view. Given the difficulties in reliably tracking eye movements this study considers methods for robust extraction of information from eye data using HMMs, building on the small amount of work in this field [Sal99] [YBZ02] [CAR04], and serving as a precursor to building a Gaze-contingent ASR system.

1.4 Research questions and methodology

For the general case of Multimodal decoding, this research addresses the following questions:

- How do the modalities semantically relate?
- How can multimodal decoding be realised?
- What are the benefits?

This thesis will study these questions in the context of integrating information from eye movement in ASR systems. In answering these questions, the following objectives will be fulfilled:

- Development of a theoretical framework for combining modalities and applying it to the specific case of integrating eye movement and speech.
- Collection of a matched corpus of eye movement and related conversational speech for a visual-based, goal-driven task.
- Robust extraction of information from eye movement data.
- Establishing a research-grade baseline ASR with acoustic and task-specific language syntax adaptation.
- Realisable approaches for integrating loosely coupled modalities, applied to integrating eye movement information into ASR systems for decoding spontaneous, conversational speech.

- Analysis of the Gaze-contingent ASR systems and their benefits.

1.5 Presupposition

The decoding of modalities obtains information from which meaning, or semantics, can be inferred. In a gaze-contingent ASR system the focus of visual attention is decoded from eye movement and speech is extracted from voice.

This thesis makes the presupposition that the focus of an individual's visual attention indicates an increased chance of that person's speech being related to the visual focus. That is to say, the acoustics of speech are related to eye movements via a semantic relationship between what is being looked at, and what is being talked about. If this semantic relationship is strong, then one may reason that an underlying, abstract semantic concept exists which unites both.

An abstract semantic concept is also referred to in this thesis as a 'modality-independent semantic class'. It represents a user's intent. The semantic relationship may be subtle - e.g. saying 'that green apple' while looking at a bowl of fruit would indicate a semantic relationship between words and visual focus that is intuitively correct with an (e.g.) intent of selecting a piece of fruit. The relationship, however, may be more explicit - such as saying 'pass me that green apple in the fruit bowl' while looking at it. In both instances, an ASR could be biased towards the acoustics associated with phrases that have a semantic relationship with the apple and fruit bowl objects.

Furthermore, the semantic relationship between what is said and what is looked at exists regardless of whether the speech turns out, in practice, to be related to the visual focus. Clearly, one can look at the fruit bowl and say something completely unrelated. There may be many instances where speech is observed as not directly related to the visual focus. In fact, in most conversations this would be the case. However, the semantic relationship still exists, and in probabilistic terms indicates that, regardless of the actual speech, an increased chance, or potential, for speech to relate to visual focus exists due to the existence of a semantic relationship that represents the user's intent.

1.6 Thesis structure

This thesis unfolds as follows:

- Chapter 1 gives a broad introduction to the study.
- Chapter 2 develops a multimodal decoding framework and applies it to eye movement and speech.
- Chapter 3 describes the corpus of eye movement and speech collected.
- Chapter 4 documents the development of the Baseline ASR system.
- Chapter 5 explores techniques for robust extraction of information from eye movement data.
- Chapter 6 implements Gaze-contingent ASR systems and evaluates their performance.
- Chapter 7 concludes this thesis detailing the main contributions and future research directions.

The Appendices detail the extensive software developed for this study and publications related to work in this thesis.

Section references are preceded with the symbol §, e.g. §1.6 for this section. Equations are referenced by bracketed number: e.g. (2.1). Footnotes[4] are used for clarifications and references to other parts of the thesis.

1.7 Summary

This chapter introduced this study by describing the background of this research, stating the research questions and outlining the thesis methodology and structure.

[4]Expression (2.1) can be found in Chapter 2 §2.4 on page 17.

2 Integration Theory

This chapter develops a theoretical foundation for multimodal decoding and applies it to the specific case of integrating eye movements into ASR. §2.1 considers multimodal integration from a neurological, data fusion and human-computer interface perspective. §2.2 discusses previous research into eye movement and speech, in order to characterise these modalities and their relationship. The discussions in §2.1 and §2.2 are summarised in §2.3, after which §2.4 outlines a statistical model for multimodal integration and derives a scheme for integrating eye movement and speech into ASR.

2.1 Perspectives

2.1.1 Neurology

The neurological understanding of how humans combine modalities comes from research into two regions of the brain - the cerebral cortex and the superior colliculus.

The cerebral cortex receives, processes and comprehends all senses. Sensory information is combined, together with memory, to produce perceptions, emotion and thought. In addition, the cerebral cortex coordinates voluntary motor function - including speech.

The superior colliculus has been extensively studied in terms of sensor integration, as it contains a high proportion of multi sensory neurons. This region influences the reflexive actions caused by attentive and orientation processes - e.g. saccadic eye movements, head rotation, and hand-eye coordination.

In the human nervous system, individual senses are activated when a sensing event occurs. There are neural pathways from these senses, which enter the superior colliculus, where the pathways converge. Exiting multi sensory neural pathways terminate in the nervous system and trigger the motor response. The method by which

the superior colliculus combines pathways is regulated (or modulated) by additional neural pathways from the cerebral cortex.

The additional pathways from the cerebral cortex modify the integration method used by the superior colliculus. The integration is modified based on a context, or percept, derived from higher-level process. These higher-level cognitive processes are complex, and not well understood. The integration method is accrual across senses; multiple weak events from multiple senses produce stronger activations, compared to strong events from single senses. The similarity of the individual sensing event cues determine the strength of an activation of a multi-sensor neural pathway, in terms of their temporal and spatial alignment [SWS99].

The cerebral cortex's exact influence on the integration process is not known. However, in cognitive psychology, Bowers proposed a taxonomy of fusion models [Bow74] to describe the different methods of integration. The taxonomy considered the contention (or discordance) between sensors when they are combined. If contention is detected, it may be ignored or feedback initiated to modify fusion process to reduce the contention. The feedback recalibrates sensors, suppresses them, or stops the integration. This taxonomy has been considered in relation to robotics, to justify closed-loop feedback control in multi-sensor integration [Mur96]. It has also been proposed as a design heuristic in HCI [SPH98]. Bower's taxonomy does not mean that closed feedback is required to deal with contention in multimodal systems, it only suggests that feedback is one method to deal with controlling the integration process - for example, integration could be controlled by a separate process altogether.

2.1.2 Data fusion

The literature on data fusion provides a taxonomy for describing at which stage combining signals can occur in the pattern recognition process [AH91]. Signals may be combined early, at sensor level (sensor fusion); midway, at the feature extraction level (feature fusion); or later on, after classification (decision fusion). The more temporally and semantically disparate the signals are, the later in the process they can be realistically combined. In the case of integrating modalities, sensor fusion may be ruled out as multiple modalities are, by definition, physically different. This leaves feature and decision level fusion as integration options. Figure 2.1 shows the options for data fusion to combine modalities.

The feature extraction process in pattern recognition typically results in an N-dimensional vector representing a sample of the signal. Feature vectors from parallel,

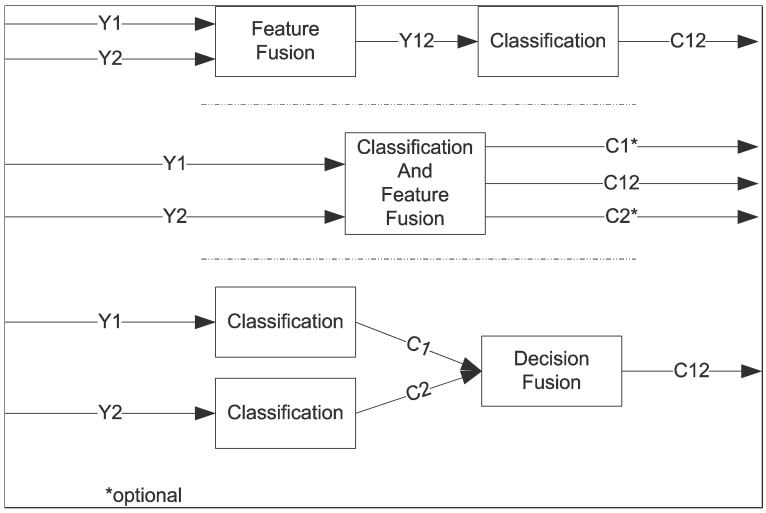


Figure 2.1: Three data fusion options for multimodal integration, considering two modalities. Top, middle and bottom pictures show fusion before, during, and after classification respectively. Y1 and Y2 represent feature vectors for the two modalities. Y12 is the combined feature vector and C12 the classification resulting from both modalities. Likewise, C1 and C2 represent classifications for Y1 and Y2 respectively. Note that fusion during classification (mid diagram) affords combined classifications, and optionally, separate classifications. This thesis considers fusion during classification for Gaze-contingent ASR.

matched signals may be combined using feature fusion, producing a combined feature vector for the signals. This vector provides a single input into the classification process. Alternatively, the feature vectors from multiple senses can remain separate, each as an input into the classification process. The classification process outputs a single classification for both signals, and optionally, individual classifications for each modality. Both these methods may be considered as 'early integration'.

In decision level fusion, signals are classified by separate processes, with the classifications combined to produce an overall classification. This approach has the potential to perform well when the original signals are uncorrelated [KHDM98]. Decision level fusion can be considered as 'late integration', and is prone to non-recoverable error propagation [Hun87]. Another application for decision level fusion is using several classifiers for the same signal, each using a different technique. The output of these multiple classifiers may be fused in a weighted fashion, or switched, depending on signal value [Kun02].

For eye movement and speech, this thesis considers fusion during classification (i.e. the middle diagram in Figure 2.1). $Y1$ and $Y2$ represent the speech and eye

movement feature vectors respectively, and $C1*$ and $C2*$ the word sequence and visual attention respectively. $C12$ represents the semantic relationship between the eye movement and speech[1].

2.1.3 Human Computer Interfaces

The Human Computer Interface (HCI) perspective on multimodal integration considers the use of input and output modalities in the user interface[2]. Input modalities are concerned with the communication between human and machine, requiring machine recognition and interpretation of human actions to classify user intent[3] (e.g. Speech recognition and eye tracking.). Output modalities are concerned with the communication between machine and human, requiring synthesis and presentation of information (e.g. speech synthesis and visual displays). The interaction between input and output modalities is guided by the system application and HCI design principles, often informed by models of interaction [BM01].

The integration of input modalities provides an opportunity to use multiple modalities to interpret user intent and reduce recognition errors. E.g. Decision-level fusion architectures have been implemented to recover errors from gesture and speech. This so-called 'mutual disambiguation' of errors can also be utilised as an evaluation measure for assessing integration worthiness [Ovi99b]. Patterns of multimodal interaction [Ovi99a] can inform interface design.

To explore the effectiveness of eye tracking interfaces in HCI research, wearable eye trackers have been used to observe human behaviour in respect to preparing food and other natural tasks, in order to recognise sequences of actions based on visual attention [YBZ02].

Beyond the interaction between modalities for determining the user intent, many modalities provide partial information about a user's attentive processes. To this end, Horvitz et al see attention as a 'central construct and organising principle' for HCI design for ensuring an interface provides relevant information and timely interruptions. Each input modality is seen as providing specific cues, or more formally, probabilistic evidence, about the users available attentive resource and attentive state. Eye movements provide an important cue - the current focus of visual attention [HKPH03].

[1] §1.5

[2] See [BJP+00] for a substantial review of audio-visual multimodal systems.

[3] For brevity, outside this section of the thesis 'modality' means to an input modality.

2.1.3.1 Previous eye/speech HCI

Useful design heuristics and interesting anecdotes can be gained from looking at existing multimodal systems. In this section, notable implementation types that were concerned primarily with the integration of eye movement and speech as multimodal inputs are described. Review papers on multimodal system taxonomies and technologies cite numerous systems [BJP+00] [Ovi02].

Bolt's seminal proposal for a direct manipulation interface (DMI) [Bol84] 'Put That There' was realised by computer technology by the early 1990s, where together with Thorisson and Koons he presented one of the first working eye/speech and gesture recognition systems [TKB92]. Other similar systems followed (i.e. VisualMan [Wan95]). More recently, the interaction between eye movement and speech and its variation amongst different users has been characterised in the 'PICCS' system [Kau03].

Eye movements have been used to resolve speech recognition errors while using conversational dialogue systems, by boosting word scores [SH97] and resolving the differences between similar sounding words by looking at the focus of visual attention [ZIGM04]. Knowledge of the visual field (rather than eye movements) has been used to guide ASR towards words in its vocabulary [RM05].

2.2 Relationship between eye movement and speech

2.2.1 Exploring cognition from observing eye movements

Eye movements have long been of interest in cognitive psychology and science as they offer an insight into cognitive processes during activities involving visually oriented information processing. They also provide an indication of attentive processes through the analysis of eye fixation durations and direction. Cognitive psychology experiments chiefly involve the controlled elicitation of selected brain processes by giving the subject an explicit goal or task. Studies investigating eye movements during scene perception and language production are therefore relevant and worth consideration.

2.2.2 Attentive processes and scene perception

In essence, scene perception concerns looking at a visual scene to extract information from it in order to form a percept. This involves attending foci of interest in the scene by fixating on one of the foci. Fixating involves moving the eye so that the light from a focus falls within the foveal centralis region of the retina (the centre).

The retina consists of photoreceptors called rods and cones. The rods are more sensitive to light and the cones to colour. The foveal centralis consists only of dense cones (no rods), and thus has the highest visual acuity and colour sensitivity. Outside of the fovea, visual acuity decreases as cone density decreases. The increasing proportion of rods makes peripheral vision sensitive to movement (i.e. changes in light level) at the expense of acuity.

The foveal centralis is 1.5 degrees in diameter from the pupil centre. The average duration for a fixation during scene perception is 330ms, and the first fixations in scene perception are concerned with acquiring the gist (outline) of the scene, with later fixations filling in areas of interest [Ray98].

Jenkin and Harris [HJ98] [PP90] offered a concise definition for 'Attending visual foci of interest', or more succinctly, 'Visual attention':

> Attention implies allocating resources, perceptual or cognitive, to some things at the expense of not allocating them to something else.

As a concept, attention was described by William James, in his influential work 'The Principles of Psychology' [Jam90]. James hypothesised that the things we attend to may be directed towards sensory objects (e.g. visual objects) or ideas (e.g. thought and memories), and that attention can be either a voluntary or involuntary reflex.

Likewise, the prevalent view of attention in neuroscience is of a distinct neural system with limited capacity, which results in a shifting 'Attentional spotlight' [Pos80][4]. Posner identified three states in the visual attention system [PP90]: Shifting attention to a focus; Attending the focus; and maintaining an alert state.

Shifting attention to a focus normally involves moving the eyes, and optionally the head, so that light from the focus falls on the foveal centralis. It is, however, possible to shift attention and attend a focus outside of the fovea [Jam90]. Naked-eye

[4]Research of human brain scans using magnetic resonance imaging (MRI) [BD99] confirms Posner's spotlight metaphor. Recent MRI findings suggest multiple spotlights with a degree of parallel processing [MS04].

astronomers, for example, attend focus off the centre of their vision (the para-fovea) to improve detection of faint objects[5].

Attending the focus involves processing the visual information contained at it. Within 150ms of attending a focus, neural processing and response to events at the focus are faster [PP90]. Maintaining an alert state involves not attending the current focus while anticipating new foci. The time to attend a focus is finite. This is due to a cognitive suppression in noticing the stimuli the longer a fixation is held, leading to a natural reflex to refixate on a focus to maintain stimuli. This effect is called Troxler fading [HS04], which is the earliest known type of 'disappearance phenomenon', where visual information is ignored by the visual system although detected by retinal photoreceptors.

In a recent review of the shifts in the focus of visual attention during scene perception [Hen03], eye movement control was considered from both a knowledge and stimulus-driven perspective. Both perspectives simultaneously exert influence on what determines the focus of visual attention. The knowledge-driven control of visual attention is based on memory, semantic context and the user goals. The stimulus driven control perspective is based on visual characteristics of the scene [RZHB97]. The sequence of focus shifts during scene perception has been shown to be highly variable and is not related to representations of the visual scene in memory[6]. Fixation sequence dependency on knowledge and/or stimulus driven control is a current research topic.

2.2.3 Definition of visual attention

Using Posner and James' understanding of attention from neurological and cognitive viewpoints, and Henderson's understanding on the control of visual attention, a definition of visual attention can be made which serves this thesis:

> Visual attention involves attending a visual sensory object (focus) and may be either a voluntary or an involuntary reflex. The eye will rotate so that the foveal centralis region of the retina receives light from the object. Once attending the focus is complete, an alert state ensues until a new focus is identified. Identification of foci is guided both by low-level

[5] As noted by [Duc03], all eye tracking work tacitly assumes attention is linked to gaze direction but acknowledges this is not always so.

[6] This was known as scan-path theory [NS70].

stimulation from sensory modalities and higher-level memory, goals and context.

2.2.4 Language production

Due to the cognitive processes involved in language production, when a person makes verbal reference to objects in their visual field, this potentially provides constraints on shifts in the focus of visual attention.

'The theory of lexical access in speech production' was proposed by psycholinguistic researchers Levelt et al [LRM99]. In brief, the theory proposed that lexical access (word production) consisted of six sequential stages of cognitive processing, from concept to sound wave. The stages were conceptual preparation; Lexical selection; Morphological encoding; Phonological encoding; Phonetic encoding; and Articulation. By controlling the duration that labelled objects and words were shown to a person, these stages and their respective durations were inferred from fixation and saccadic eye movement. The theory was tested against the finding that, when viewing and naming visual objects, they are fixated on long enough to complete lexical access [MSL98].

The effect of eye movement during lexical access has been heavily researched in relation to reading. Rayner provides a thorough review [Ray98]. The person's goal while reading is information comprehension, with lexical access the underlying cognitive process. In reading, the foci of visual attention are the individual words on a page. Word fixation durations during reading are between 200-350ms with a saccadic duration of 15-40ms covering 7-9 letter spaces. During a fixation, cognitive processing primarily consists of lexical access of the word being fixated upon, and programming the next saccade to the next fixation point, which is determined partially by parafoveal preview. The eye only makes a saccade once lexical access is complete. Oral reading is, on average, half the speed of silent reading with more corrective movements (refixations), shorter saccades and longer fixations. Eyes are typically ahead of the voice, which is commonly known as the 'eye-voice span', by two words. Oral reading introduces more errors than silent reading due to the additional cognitive processing required for articulation. Errors are measured by counting the number of corrective eye movements.

Reading provides a good insight into the eye behaviour expected when viewing objects and comprehending them, as recognition of the name of an object from its visual form in order to verbalise it explicitly requires lexical access. An eye-voice

span for reading of 'approximately two words ahead' [Ray98] corresponds to a time of between 430ms and 780ms. This is similar to that of mean eye-voice spans for naming pictures of between 740ms and 805ms [MSL98]. Mean eye-voice spans of 902ms and 910ms for subject and object nouns respectively have been reported for describing visual scenes using spontaneous speech [GB00]. The eye-voice span decreases on subsequent viewing of an object due to the use of recent memory, pronouns or common nouns [MML01].

Beyond considering the eye-voice span for naming words, as a constraint on shifting the focus of visual attention, the syntactic structure of an utterance containing multiple objects has been shown to reflect indirectly in eye movement. In particular, indecision in how to structure an utterance results in multiple fixations on candidate objects [GS04].

2.3 Discussion summary

Considering multimodal integration perspectives in general (§2.1), and the relationship between eye movement and speech (§2.2), the questions posed in Chapter 1 may now be revisited:

- How do the modalities semantically relate?
- How can multimodal decoding be realised?
- What are the benefits?

The neurological perspective of multimodal integration in §2.1.1 gave some background on understanding how decoding is realised. It offered the idea of selective use of unimodal events from modalities for integration - i.e. not all of the neural pathways from a sense converge with other senses. Mechanisms in the superior colliculus indicate the weakening of integration as a function of temporal asynchrony. The contextual influence in the integration process from the cerebral cortex points to the idea of integration being adaptive, with mechanisms for dealing with contention across modalities.

The data fusion perspective in §2.1.2 further informed the realisation question, outlining a framework for deciding when modalities should be combined in the pattern recognition process, and the requirement for defining appropriate classification spaces for the integrated modalities - i.e. separate and/or combined.

The HCI perspective in §2.1.3 pointed to previous results of integrating modalities using decision-level fusion architectures and measuring the benefits of integration.

For deciding how the modalities are related, and considering the specific case of eye movement and speech, the research summarised in §2.2 suggests that speech should be decoded as a word sequence and eye movement should be decoded as a focus of visual attention (FOVA) sequence. Speech should be related to the FOVA via words that are associated with the potential visual foci. Decoding should primarily account for the fact that eye movement precedes related speech by the 'eye-voice span'. Considering various studies, the 'eye-voice span' ranges from 430ms and 910ms [Ray98] [MSL98] [GB00]. The randomness in eye movement while reading indicates that not all fixations have a cognitive purpose. The randomness in the sequence of FOVA during scene perception indicates a limited utility in analysis of fixation sequences of scene perception related tasks.

Developing a framework for multimodal decoding that incorporates eye movement and speech should therefore focus on the following aspects, which are also applicable, and considered for, the general case of multimodal decoding:

1. Temporal asynchrony: Signals from modality sensors may differ in terms of data sampling rate. The information in the modalities that can be integrated may not temporally co-occur. This is apparent from analysing eye movements and speech where people may look at an object before they explicitly refer to it, and the use of different physical sensors.

2. Semantic asymmetry: Modalities contain information from which meaning, at various levels, can be inferred. The overlap in semantic representations for modalities facilitates their integration. The overlap may be dynamic, which means the appropriateness and method of integration depends on when, and how, the modalities relate. There must be a way of relating what is being looked at to what is being said. Importantly, integration must consider whether or not there is a relationship and deal with any contentions.

3. Early vs. late integration: Temporal asynchrony and semantic asymmetry promote integrating modalities at a later stage in decoding, however early integration is desirable as it prevents error propagation through the decoding system by delaying decisions.

4. Implementation: The realisations of formalisms for integration must be computationally efficient by the use of appropriate simplifications.

This chapter develops this framework.

2.4 A probabilistic framework for combining modalities

This section formalises the problem of recognition of multiple modalities using probability calculus and Bayesian inference. Bayesian inference is the prevalent framework for reasoning under uncertainty. It begins by deriving a generalised expression for decoding, which may be applied to a specific multimodal decoding problem by making conditional independence assumptions.

The evaluation of the terms in the formalised expression is discussed with relation to statistical models for classifying time-series data and a graphical model formalism. A scheme is proposed to characterise the coupling between modalities using mutual information measures to inform the coupling characterisation. Throughout, the application to decoding eye movement in speech is demonstrated, resulting in a realisable eye movement and speech decoding scheme. This decoding scheme forms the basis of experimental design in Chapter 3 and the integration experiments in Chapter 6.

2.4.1 Generalised expression for multimodal recognition

Let M be the set of N modalities $M = \{M_1, \ldots, M_N\}$ and C be the set of class types $C = \{C_1, \ldots, C_S\}$ which will be decoded. Let m be a set of measurements taken from M where $m = \{m_1, \ldots, m_N\}$ and m_i represents a feature vector for modality M_i. Let c be a set of classes from C where $c = \{c_1, \ldots, c_S\}$ and c_i represent a classification for class type C_i. For example, for this study, M represents the eye movement and speech, C represents the set of possible words and foci of visual attention, and c represents either the individual word or the focus of visual attention.

The aim of multimodal recognition is to maximise the probability of correct classification $\widehat{c}$ conditioned on the modality measurements:

$$p(\widehat{c}|m) = \max_{c} p(c|m)p(c) \tag{2.1}$$

By applying Bayes' inversion formula the conditional probability $p(c|m)$ may be expressed in terms of the likelihood function $p(m|c)$ and prior $p(c)$:

$$p(c|m) \propto p(m|c)p(c) \tag{2.2}$$

The likelihood function $p(m|c)$ represents all modalities used for recognition. This expression may be expanded to obtain an expression in terms of the likelihood function for each modality:

$$\begin{aligned} p(m|c) &= p(\{m_1, \ldots, m_N\}|c) && (2.3) \\ &= p(\{m_i\}|c, m - \{m_i\})p(m - \{m_i\}|c) && (2.4) \end{aligned}$$

Since m_i represents a measurement of any modality, (2.3) can be restated to consider the likelihood functions for each modality:

$$p(m|c) = \frac{\sum_{i=1}^{N} p(\{m_i\}|c, m - \{m_i\})p(m - \{m_i\}|c)}{N} \tag{2.5}$$

Substituting (2.5) into (2.2) gives a generalized probabilistic expression for multimodal pattern recognition in terms of all modalities and classes:

$$p(c|m) \propto p(c) \sum_{i=1}^{N} p(\{m_i\}|c, m - \{m_i\})p(m - \{m_i\}|c) \tag{2.6}$$

Expression 2.6 factors out decoding of individual modalities, conditioning them on all other modalities and classifications. This generalised expression may be applied to any multimodal decoding problem by making conditional independence assumptions between modalities and classes[7].

2.4.2 Application to speech and eye movement

The generalised expression for multimodal decoding in (2.6) can be applied to the specific problem of decoding eye movements and speech by defining class types, modalities, and making conditional independence assumptions.

Let e and y be the set of modality measurements m (2.7), representing the sequence of feature vectors for eye movement (2.8), of length et, and speech (2.9) of length yt, respectively:

[7] A less generalised expression for the two modality case (speech and gesture) was proposed in [WOC00].

$$m = \{e, y\} \tag{2.7}$$
$$e = (e_1, \ldots, e_{et}) \tag{2.8}$$
$$y = (y_1, \ldots, y_{yt}) \tag{2.9}$$

Let v and w be the sequence of class types, c, (2.10) that represent the FOVA sequence (2.11) and the word sequence. (2.12):

$$c = \{v, w\} \tag{2.10}$$
$$v = (v_1, \ldots, v_{vt}) \tag{2.11}$$
$$w = (w_1, \ldots, w_{wt}) \tag{2.12}$$

Substituting (2.10) into (2.6) yields:

$$p(v, w|e, y) \propto p(v, w)p(e|v, w, y)p(y|v, w, e) \tag{2.13}$$

Expression 2.13 may be simplified by making more conditional independence assumptions. A first assumption is that the effect of one modality on another is only via the other's classification, and not the measurement of the other modality itself. Consider, for example the FOVA v (i.e. focus being viewed) and the actual eye position e (i.e. spatial coordinates on the visual scene). In expression 2.13 the likelihood function for speech y, $p(y|v, w, e)$, is conditioned on both v and e. A conditional independence assumption between y and e, given v, can be made:

$$p(e|v, w, y) \approx p(e|v, w) \tag{2.14}$$

Likewise, an assumption of conditional independence can be made between e and y given w:

$$p(y|v, w, y) \approx p(y|v, w) \tag{2.15}$$

Such independence assumptions may be supported empirically by looking at the lack of correlation between the modality feature spaces for a given classification. In the case of eye movement and speech, this lack of correlation may be assumed

intuitively since the eye position vectors (i.e. spatial coordinates) and acoustic vectors (i.e. spectral features) are each concerned with separate classification types (i.e. the effect of the eye position vectors on a word sequence is entirely due to the effect of the FOVA). This independence assumption yields:

$$p(v, w|e, y) \quad \propto \quad p(v, w)p(e|v, w)p(y|v, w) \tag{2.16}$$

A further simplification is to assume that a modality's effect on the other modality is via the joint probability $p(v, w)$, and not the posterior probabilities for each modality:

$$p(v, w|e, y) \quad \propto \quad p(v, w)p(e|v)p(y|w) \tag{2.17}$$

Thus expression 2.1 for the general case of maximisation is simplified to:

$$p(\widehat{v}, \widehat{w}|e, y) \quad \propto \quad \max_{v,w} p(w, v)p(e|v)p(y|w) \tag{2.18}$$

This expression for decoding requires the maximisation of the joint probability $p(w, v)$. $p(w, v)$ may be expressed in terms of the likelihood function of one modality, and the prior for the other:

$$\begin{aligned} p(w, v) &= p(v|w)p(w) && (2.19) \\ &= p(w|v)p(v) && (2.20) \end{aligned}$$

Substituting 2.19 into 2.16 yields two possible expressions for $p(\widehat{v}, \widehat{w}|e, y)$:

$$\begin{aligned} p(\widehat{v}, \widehat{w}|e, y) &\propto \max_{v,w} p(v|w)p(w)p(e|v)p(y|w) && (2.21) \\ &\propto \max_{v,w} p(w|v)p(v)p(e|v)p(y|w) && (2.22) \end{aligned}$$

Where the first two terms in (2.21) and (2.22) encode the conditioning between the two modalities and the final two terms are the class conditional probabilities of the measurements. Thus, the multimodal decoding of eye movement and speech may be expressed as a joint optimisation problem of the probability distribution of two hidden discrete-time stochastic variables: the word sequence, w; and the FOVA, v.

2.4.3 The Hidden Markov Model

The posterior probabilities $p(e|v)$ and $p(y|w)$ in (2.21) and (2.22) may be estimated using Hidden Markov Models (HMM). The HMM is a probabilistic model for classifying time series data. It has two stochastic processes; a hidden first order Markov process consisting of a discrete state-space, and an observation set which includes the known time-series data requiring classification.

The HMM emits an observation at time t conditioned on the state in the hidden first order Markov process at time t. For simple classification problems, the classes correspond to states in the Markov process. In more advanced classification (e.g. speech recognition), each class (e.g. unit of speech) corresponds to a multi-state model.

HMMs are popular for extracting information from modalities since their training affords an optimised model with probabilistic (degree of belief) classification. Any falsity in the 1st order Markov assumption for the hidden stochastic process can be offset by the number of states in the model since any n_{th} order Markov process may be represented as a 1st order Markov process given a sufficient number of states [Bil02], and the statistical rigour of the training and inference algorithms.

2.4.3.1 HMM Structure

In an ergodic HMM each state may transition to every other. For an ergodic HMM with N hidden states, $\omega_i (i = 1, \ldots, N)$, the observation symbol set defines the set of possible observations. The observation sequence, $\mathbf{O}$, represents the time series data to be classified (2.23). The state sequence, $\mathbf{s}$, describes the state sequence over time t (2.24). The state transition probability matrix, $\mathbf{A}$ (2.25), describes the probability of transitioning from one state to another, a_{ij} (2.26). $\mathbf{A}$ is row stochastic (2.25). The choice of initial states is governed by an initial state probability matrix, π (2.28) (2.29). The state output PDF b_i for each state i defines the set of possible observations from the observation symbol set with $b_i(o_t)$ being the probability of the observation o_t (2.30):

$$\mathbf{O} = (o_1, \dots , o_t, \dots, o_T) \tag{2.23}$$
$$\mathbf{s} = (s_i, \dots, s_t, \dots, s_T) \tag{2.24}$$
$$\mathbf{A} = [a_{ij}]_{i,j=1,\dots,N} \tag{2.25}$$
$$a_{ij} = P(s_t = \omega_j | s_{t-1} = \omega_i) \tag{2.26}$$
$$\sum_{j=1}^{N} a_{ij} = 1 \; i = 1, \dots, N \tag{2.27}$$
$$\pi = [\pi_1, \dots, \pi_N] \tag{2.28}$$
$$\pi_i = P(s_1 = w_i) \tag{2.29}$$
$$b_i(o_t) = P(o_t | s_i) i = 1, \dots, N \tag{2.30}$$

2.4.3.2 Viterbi Decoding

Viterbi Decoding recovers the optimum state sequence $\widehat{s}$ through which the hidden Markov process generates the observation sequence o, relative to model M:

$$\widehat{s} = \arg\max_{s} p(o, s|M) \tag{2.31}$$

A state(i)-time(t) trellis is constructed to represent hidden state changes over time. At each node (i, t) a record is kept, j_{max}, of the state, j, at time $t-1$ that gives the maximum calculated value of the forward probability, $\hat{\alpha}_t(i)$:

$$j_{max}(i, t) = \arg\max_{j}(\hat{\alpha_{t-1}}(j) a_{ji} b_i(o_t)) \tag{2.32}$$
$$\hat{\alpha}_t(i) = \max_{j}(\hat{\alpha_{t-1}}(j) a_{ji} b_i(o_t)) \tag{2.33}$$

Where o_t is the observation at time t and j is the hypothesized state at time $t-1$. The optimum hidden state sequence, $\widehat{s}$, is thus recovered by tracing back from the final node in the trellis using j_{max} to identify the previous state. Figure 2.2 shows a single path through the trellis.

Viterbi Reestimation is covered in Appendix C.

2.4.4 Dynamic Bayesian Networks

The definition of HMMs given in §2.4.3 and frequently cited from Levison's [LRS82] and Rabiner's [Rab89] seminal papers in speech recognition, can be described in terms of Dynamic Bayesian Networks [Bil02].

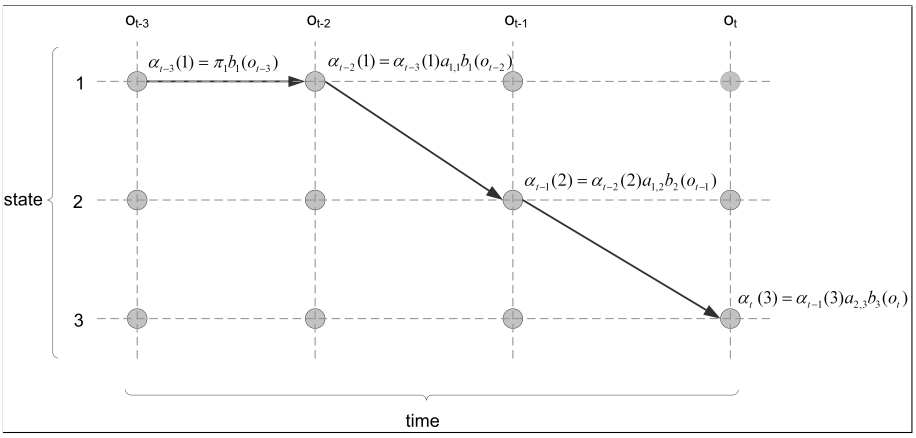


Figure 2.2: The state-time trellis for a HMM showing a single path through the trellis, which is denoted by the arrowed line.

A Dynamic Bayesian Network (DBN) is a directed acyclic graph (DAG) consisting of nodes and directed arcs, with the arcs directed in time. Random variables are represented as nodes, and conditional dependence assumptions between the random variables as directed arcs between nodes. The random variables generate discrete or continuous values according to a distribution of probabilities about their possible range of values [Mur98].

The DBN formalism enables statistical models, including the HMM, to be represented in DAG form[8], with universal algorithms for inference and estimation. The Viterbi algorithm, for example, is a specific form of the DBN Dawid algorithm [Daw92]. The DBN DAG representation for a HMM shows the hidden discrete-time stochastic process as a 1st-order Markov chain of hidden random variables ordered in time. Figure 2.3 shows a DBN representation of a HMM.

Inference algorithms typically consist of transformation of the DAG using graph theory, followed by belief propagation in the transformed DAG to find the values of hidden random variables, given the known (observed) random variables.

The use of DAG representation for HMM variants preceded the use of the DBN algorithms to provide corresponding inference and parameter estimation. Producing tractable versions of such algorithms, however, so that DBN provide more than just a graphical viewpoint, is a current research challenge. For this reason, this discussion restricts the DBN use to model visualisation.

[8]DBN DAG's do not show the random variable observation form, conditional dependency implementation and time-homogeneity assumptions, which must be stated separately [Bil02].

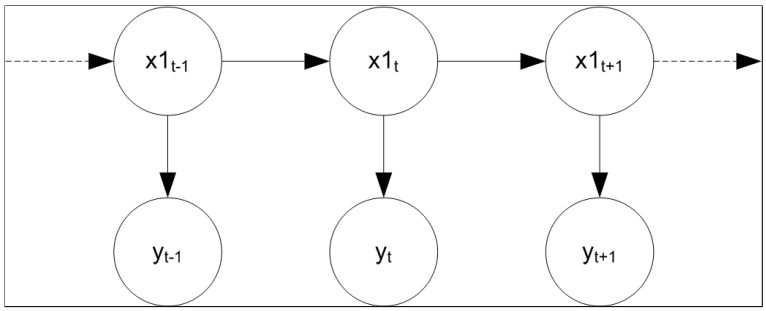


Figure 2.3: DBN representation of a HMM. For HMMs, random variables x are discrete scalars and form a 1st-order Markov chain. Random variables y and continuous vectors conditioned on the hidden state random variable x.

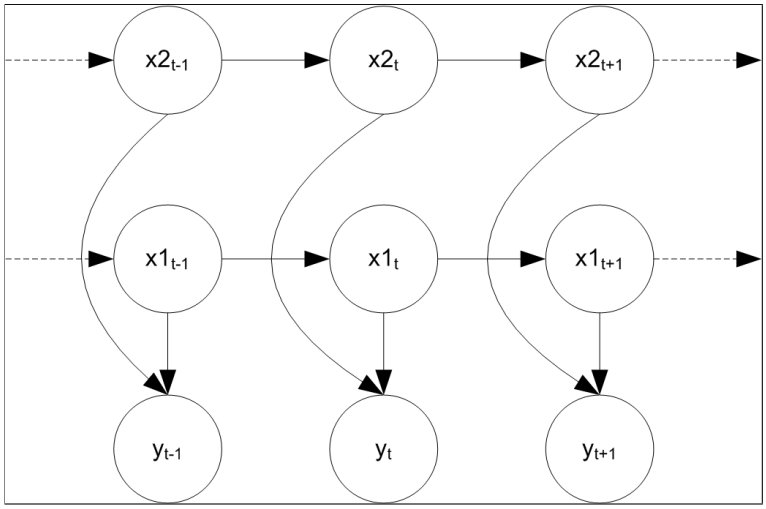


Figure 2.4: DBN representation for a Factorial HMM

2.4.5 Factorial HMM and variants

A factorial HMM [GJ97] [GJ95] describes the family of HMM variants which assume that there are multiple hidden Markov chains in the hidden stochastic process - that is, the hidden state may be decomposed into more than one random variable [SJ99]. These are potentially useful models to consider when decoding multiple modalities, as each modality can be represented as a separate time-series of random variables. Figure 2.4 shows an example of a FHMM.

In the standard FHMM, the Markov chains are marginally independent, but conditionally dependent, given the observation. Inference involves exploring the state space S_t, at time t, which is formed from the Cartesian product of all individual hidden Markov chains' state spaces, s_t^c:

$$S_t = s_t^1 \times s_t^2 \times \ldots \times s_t^{Nc} \tag{2.34}$$

Where N_c is the number of hidden chains. The optimum state sequence, $\widehat{S}$ is determined from all possible state sequences, $S = \{S_1, \ldots, S_t, \ldots, S_T\}$:

$$\widehat{S} = \arg\max_S p(o, S|M) \tag{2.35}$$

The factorisation of the state space in 2.34 makes exact inference and estimation intractable, due to a combinatorial explosion of state transition parameters. To overcome this, various approximate inference and estimation schemes exist [GJ97]. These, in essence, simplify the model by relaxing conditional dependence assumptions between states.

There are two simplifications worth noting. The first simplification reduces the number of transition probabilities of the factorised state space, by assuming each Markov chain at time t is independent of all other Markov chains at time t, given the factorised states at $t-1$ [SJ99]:

$$p(S_t|S_{t-1}) = \prod_c p(s_t^c|S_{t-1}) \tag{2.36}$$

The 2nd simplification uses the chain-wise Viterbi algorithm for inference [SJ99] [NY00]. The chain-wise Viterbi algorithm initially performs Viterbi decoding[9] on each hidden chain separately. After this initialisation, each chain is Viterbi decoded in turn, considering the current decoded state of all the other chains.

2.4.6 A FHMM-inspired model for integration

The FHMM described in §2.4.5 goes some way towards the aim of feature-fusion integration of eye movement and speech. The semantic asymmetry between modalities is accounted for, because each modality decode is represented by a separate Markov chain. The factored state space represents the 'coming together' of the individual modalities' state representations.

The FHMM model structure assumes conditional dependency between chains given an observation, leading to a factorised state space. If each Markov chain, however, emits its own observations, then the conditional dependence between modalities

[9] §2.4.3.2

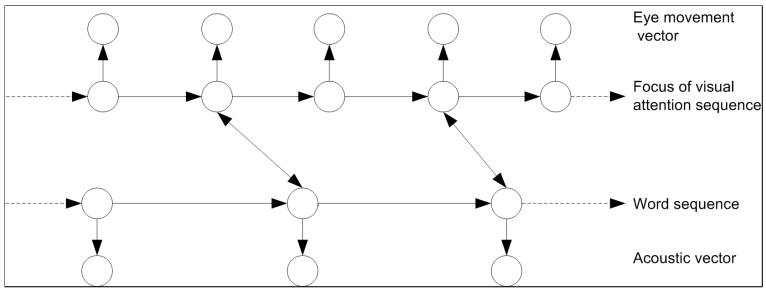


Figure 2.5: DBN representation of integration of eye movement and speech, showing the loose coupling between the two temporally asynchronous streams.

can be either represented as a bidirectional arc between random variables in the separate Markov chains, or via an additional, hidden, random variable.

Figure 2.5 shows a DBN DAG representation of a feature fusion model for decoding eye movement and speech. Each decoded modality is represented as a hidden Markov chain. The bidirectional inference between Markov chains represents the fact that decoding eye movement and speech is a joint optimisation problem, as previously stated in (2.18).

The bidirectional inference in Figure 2.5 represents the coupling between streams. Unfortunately, this makes the DBN intractable in terms of belief propagation during inference. To solve this, inference can be considered in only one direction at any one time - i.e. one step of the Viterbi chain-wise algorithm. Additionally, it is intuitively desirable to assume causal directionality from the random variables temporal order - i.e. the inference should only point forward in time[10]. From discussions in §2.2.2, the 'eye-voice span' would indicate a temporal precedence of eye movement over speech.

This assumption of causal directionality is motivated by the pragmatic consideration of real-time decoding, where future observations are not available. However, if one assumes that a decoding decision can be delayed until all necessary observations of modalities have been made, the confidence in the decode of one stream over another may be used determine the directionality i.e. the stream decoded with the lowest confidence could be dependent on the stream decoded with more confidence, regardless of temporal order.

To decode the FHMM shown in Figure 2.5, factorised state spaces, even with appropriate simplifications, may not be suitable for integrating loosely coupled modalities. The discussion so far has tacitly implied that the values of random variables

[10]See Judea Pearl's writings on causation and temporal precedence [Pea88] [Pea01] and its philosophical justification in probabilistic reasoning.

correspond to unimodal classifications - i.e. words for ASR and visual foci for eye movements. In ASR, however, words correspond to a sequence of random variables, with each random variable representing a sub-word unit[11]. This means that a factorised state space for decoding is not necessarily desirable, regardless of its tractability, since sub-word units and visual foci are not semantically related.

It maybe more reasonable, therefore, to consider the transitions into states that represent the onset of classifications (i.e. onsets of words in ASR) as opportunities for coupling Markov streams, as opposed to factorising the state space. Assuming that visual attention stream influences the speech, the language model probability distribution $p(W)$ at time t is modified to $p^*(W)$ by function f due to the value of random variables i.e. the most probable sequence of visual attention, $\widehat{v}$ in the eye movement stream:

$$p^*(W,t) \quad = \quad f(p(W),\widehat{v},t) \tag{2.37}$$

2.4.7 Learning the coupling between modalities by considering Mutual Information

In §2.4 the joint optimisation problem for decoding eye and speech was given as:

$$p(\widehat{v},\widehat{w}|e,y) \quad \propto \quad \max_{v,w} p(w,v)p(e|v)p(y|w) \tag{2.18}$$

As proposed in §2.4.3, the distribution $p(e|v)$ and $p(y|w)$ may be calculated separately using HMMs. Their joint calculation using an FHMM requires the estimation of $p(w,v)$. It was suggested that temporal precedence or the relative confidence of the stream decodes could be used to determine whether to use either $p(w|v)$ or $p(v|w)$ during decoding. To avoid factorising state spaces during decoding, it was proposed that integration could be achieved by modifying language model probabilities by a function $f(p(W),\widehat{v},t)$ (2.37).

The coupling between modalities may be characterised, the corresponding coupling function determined and the direction of conditional dependency between v and w resolved by using an information theoretic approach. This approach considers the mutual information (MI) between modalities. A simple coupling function would be to consider when one modality is shifted by time τ against the other. This time

[11] ASR is described in Chapter 4.

shift potentially accounts for the temporal asynchrony between modalities during integration.

Recalling §2.4.2, each modality may be represented as a sequence of random variables. Let v_t and w_t be the random variables at time t in the sequence of FOVA, v, and words, w, respectively. The MI at time t, $I(v_t; w_t)$, is a measure of the difference in entropy[12] between the joint probability distribution $p(v_t, w_t)$ and the product of the prior distributions $p(v_t)$ and $p(w_t)$. This difference measure gives an indication of the degree of reduction in uncertainty of the random variables when they are considered together, rather than separately:

$$I(v_t; w_t) = p(v_t, w_t) \log \frac{p(v_t, w_t)}{p(v_t)p(w_t)} \tag{2.38}$$

MI does not consider the direction, or conditionality of the relationship between the 2 random variables. The optimal shift-coupling between modalities is determined by maximising the mutual information with respect to time τ:

$$\hat{I}(v_t; w_t) = \max_{\tau} I(v_{t+\tau}, w_t) \tag{2.39}$$

Both of the random variables that represent the modalities at time t are discrete valued - v_t represents potential FOVA and w_t represents the potential words. Let v^σ and w^ς refer to these discrete values (i.e. a single visual focus or word). The individual MI measures between specific visual foci and words can be estimated by looking at the matrix formed when calculating the mutual information $I(v_{t+\tau}; w_t)$:

$$\widehat{I}(v_t; w_t) = \max_{\tau}[i(v_{t+\tau} = v^\sigma; w_t = w^\varsigma)]_{\sigma=1\ldots n_v, \varsigma=1\ldots n_w} \tag{2.40}$$

Where n_v and n_w are the number of possible visual foci and words respectively, and $i(v_{t+\tau} = v^\sigma; w_t = w^\varsigma)$ is the entropy between a visual focus and word:

$$i(v_{t+\tau} = v^\sigma; w_t = w^\varsigma) = P(v_{t+\tau} = v^\sigma, w_t = w^\varsigma) \log \frac{P(v_{t+\tau} = v^\sigma, w_t = w^\varsigma)}{P(v_{t+\tau} = v^\sigma)P(w_t = w^\varsigma)} \tag{2.41}$$

The number of class types can be reduced by assuming that one set of words, or keywords, is related to a specific FOVA with no overlap, i.e. words associated with

[12] Also referred to in the literature as 'relative entropy' or 'information divergence'.

more than one focus are ignored, as are words associated with no foci. This simplification means that words considered for coupling with visual foci form a subset of the total number of words. The class types for each modality are therefore equivalent, representing the same modality-independent semantic class which itself represents a person's intent[13]:

$$v^\sigma \equiv k^\sigma \tag{2.42}$$

Thus v^σ and k^σ represent a visual focus and the associated keyword set. The assumption that one class in a modality is semantically equivalent to one class in another modality implies a diagonal MI matrix:

$$\widehat{I}(v_t; w_t) = \max_\tau diag(i(v_{t+\tau} = v^\sigma; w_t = k^\sigma))_{\sigma=1\ldots n} \tag{2.43}$$

With elements defined as:

$$i(v_{t+\tau} = v^\sigma; w_t = k^\sigma) = P(v_{t+\tau} = v^\sigma, w_t = k^\sigma) \log \frac{P(v_{t+\tau} = v^\sigma, w_t = k^\sigma)}{P(v_{t+\tau} = v^\sigma)P(w_t = k^\sigma)} \tag{2.44}$$

If words were associated with more than one visual focus, or vice versa, the matrix would not be diagonal and the departure from diagonality would imply the degree of semantic asymmetry. Thus, by making the assumption that a word may be associated with one or no visual foci, the semantic asymmetry[14] between the modalities has been accounted for.

Up to now, the direction of the relationship between w and v has not been determined. In optimising the temporal shift τ for v_t, an assumption can be made that negative values of τ mean w is conditioned on v, and that positive values of τ assume w is conditioned on v:

$$p(v, w) = \begin{cases} p(v|w)p(w) & \text{if } \tau < 0 \\ p(w|v)p(v) & \text{if } \tau > 0 \end{cases} \tag{2.45}$$

The temporal precedence of eye movement over speech (the 'eye-voice span') leads to a prior expectation that $\tau > 0$. It is also reasonable to assume that v is easier to

[13] E.g. the words 'frog', 'green' and 'toad' could form a keyword set that is associated with an image of a frog. The person's intent would be talking about the frog. See §1.5 for further explanation of the role of abstract semantic concepts underlying the semantic relationships between modalities.

[14] §2.3

decode than w and consequently more confidence can be stated in v than w, leading to a preference of using $p(w|v)$ over $p(v|w)$.

Expression 2.1 can therefore be simplified for the case of $\tau > 0$ to produce a final decoding expression for integrating eye movement in ASR:

$$p(\widehat{v}, \widehat{w}|e, y) \quad = \quad \max_{w} p(w|\widehat{v})p(y|w) \tag{2.46}$$

Where $\widehat{v}$ is the most probable focus of visual attention sequence:

$$\widehat{v} \quad = \quad \arg\max_{v} p(e|v)p(v) \tag{2.47}$$

The distribution $p(w|v)$ is informed by the MI matrix:

$$P(w_t = w^\varsigma | v_t = v^\sigma) \quad = \quad \begin{cases} P(w_t = w^\varsigma) & \text{if } w^\varsigma \notin k^\sigma \\ g(i(v_t = v^\sigma; w_t = k^\sigma), P(w_t = w^\varsigma)) & \text{if } w^\varsigma \in k^\sigma \end{cases} \tag{2.48}$$

Where the joint probability is a function, g, of the MI matrix and prior. This function is equivalent to $f(p(W), \widehat{v}, t)$.

2.4.7.1 Generalisation

The MI approach for learning the coupling between eye movement and speech may be generalised. For the eye and speech case, a temporal shift was proposed as a 'Transform function' on eye movement data to maximise the MI between modalities, discover the coupling, and determine the direction of dependency. A more generalised version of (2.39) for N modalities would consider transform functions other than just temporal shifts for all modalities:

$$\widehat{I}(m_t^1; \ldots; m_t^N) \quad = \quad \max_{h_1(.),\ldots,h_N(.)} I(h_1(m_t^1), \ldots, h_N(m_t^N)) \tag{2.49}$$

Where h_n represents the transform function of the n_{th} modality m_t^n at time t. Maximising the MI to discover the optimal coupling between modalities requires determining suitable transforms for $h_n(m_t^n)$.

The mutual information $\widehat{I}(m_t^1; \ldots; m_t^N)$ could be used to indicate whether integration is worthwhile, by using a minimum accepted value for $\widehat{I}(m_t^1; \ldots; m_t^N)$ as a decision threshold. The individual measures that make up the MI measure show the degree of similarity between the class types of different modalities - i.e. the strength of their semantic relationship.

This MI approach to discovering coupling between modalities may have further refinements. The number of class types in each modality may be grouped if they possess similar relationships to those in another modality. This was done in the specific case for eye movement and speech where sets of keywords were associated with one visual focus. Such groupings may be learnt by using a standard analysis of the MI matrix, such as clustering.

2.4.7.2 Related work

The MI matrix is related to the concept of multimodal associative maps [WOC00], where the decision level fusion of unimodal classes in a gesture/speech user interface was captured in a 2-dimensional matrix. Each dimension represented a modality, with values indicating the unimodal classes e.g. speech commands and deictic gestures. Matrix elements thus represented factored multimodal classes, with each class interpreted as a user interface command. A multimodal class representing a user interface command is an example of a modality-independent semantic class representing a user intent[15]. The level of association between unimodal classes was based on frequentist estimates of co-occurrence and prior expectation.

The MI-based multimodal matrix outlined here extends this idea to consider the optimal transforms on unimodal events that are required to maximise the overall mutual information. In doing so, it estimates the loose coupling between unimodal event types across modalities.

Maximising MI has been used for feature space dimensionality reduction for data visualisation and classification, by transforming high dimension feature spaces [TC00]. In a similar vein, MI measures from audio and visual feature spaces have been used to select features which best locate a person speaking from a video consisting of multiple persons [Noc02]. Good progress has been made towards formalising the maximisation of mutual information from audio and visual integration by learning the optimal transforms of individual feature spaces [Fis04].

The MI-based discovery of multimodal coupling therefore relates to current audio-visual multimodal research but considers unimodal class types rather than the continuous signal-level feature vector space (i.e. MFCC for speech).

[15] §1.5

2.5 Summary

This chapter has discussed theories behind multimodal integration and characterised the relationship between eye movement and speech. In doing so, a general framework for integration between loosely coupled modalities has been described. A scheme for integrating eye movement and speech has been proposed that resolves the joint optimisation problem by considering each modality decode as a separate hidden Markov process with dependencies between the underlying classes (i.e. words and FOVAs). It was suggested that temporal precedence and the relative confidence in the modality decoding schemes could dictate the direction of dependence - in the case of eye movement and speech both of these criteria support the speech being dependant on the eye movement and not vice-versa. FHMMs have been considered as a method for integration, and the factorisation of the classification space was suggested as a way to account for semantic asymmetry and discover the coupling between modalities, rather than for use during decoding. Maximising the MI between streams by applying transforms, such as shifts in time to the unimodal classifications, may also be used to optimise and learn the coupling between modalities. It was proposed that integration is best implemented by modifying the language model (word) transition probability during decoding. Finally, a formalised integration scheme for a Gaze-contingent ASR system was presented.

The next chapter, Chapter 3, moves on to describing the collecting of a corpus eye movement and speech, in order to perform integration experiments using the proposed integration scheme.

3 A Corpus of Eye Movement and Speech

Although there are many speech corpora available to researchers, the author is not aware of any datasets for eye tracking and speech. This chapter describes the corpus of eye movement and speech collected for this study, which will be referred to hereafter as the 'eye/speech corpus'.

3.1 Motivation

The motivation for collecting the eye/speech corpus was to obtain a set of eye movement direction data and related spontaneous speech for an unscripted, visually oriented, task constrained, human-to-human dialogue. Constraining the dialogue to a visually oriented task ensured that the eye movements of the participants were related to the subject of their speech. The corpus enabled evalaution of the gaze-contingent ASR systems built for this study.

3.2 The HCRC Map Task

The candidate task that participants undertook in the eye/speech corpus was loosely based on the HCRC Map Task corpus [ABB+91]. The HCRC Map Task corpus contains 128 sessions of unscripted human-to-human task-oriented dialogues. Each session involves two participants, both of whom look at a map. Each map has a set of landmarks, and one map describes a route around the landmarks. The objective of the task is for the participant (the 'Instruction Giver') with a map that includes the route, to describe this route to the other participant (the 'Instruction Follower'), who replicates the route onto their map. Although both are based on the same general map, some of the landmark positions and names differ. In the HCRC Map Task, landmark differences were designed to elicit particular linguistic phenomena in the

participants' speech as they attempt to resolve the differences between their maps. The experimental variables include the presence or absence of eye contact and the degree of familiarity between participants. The effectiveness of the communication was measured by how accurately the route was replicated by the Instruction Follower.

Whereas the HCRC Map Task was designed to study dialogue and elicit linguistic phenomena by differing map details, these were not the objectives of the eye/speech corpus. The task, however, involves a participant describing landmarks on a map and a route through them. This constraint on the task was chosen with the expectation that participants' gaze direction would be related to their speech. The main departure from the HCRC Map Task corpus is that the Instructions Giver's gaze direction was tracked relative to the map being described, and that there was no control over the names of the landmarks on the map.

The HCRC Map Task corpus, as well as inspiring the task used in the eye/speech corpus, was also valuable as data for constructing language models for ASR[1]. The terms 'Instruction Giver' and 'Instruction Follower' from the HCRC Map Task are maintained, and may be referred to in shortened forms as 'Giver' and 'Follower' throughout this thesis.

3.3 Method

Figure 3.1 shows the experimental description: apparatus and participants. The following subsections go on to describe the experiment in further detail.

3.3.1 Participants

One participant - an 'Instruction Giver' - described a geographical map, displayed on a VDU, to another participant - an 'Instruction Follower'. Neither participant could see the other, communicating via microphone and headphones only.

The Instruction Giver wore an S.R. Research Ltd. 'EyeLink 1' eye tracker, which is described in §3.3.6.

Nine participants took part in the experiment. All participants were foundation-year students from the University of Birmingham's School of Engineering. There were 8 male students and 1 female, aged between 18 and 19 years. All participants were British nationals and spoke English as their first language. Participants were paid the University standard per-hour rate for subjects in such an experiment.

[1] §4.9

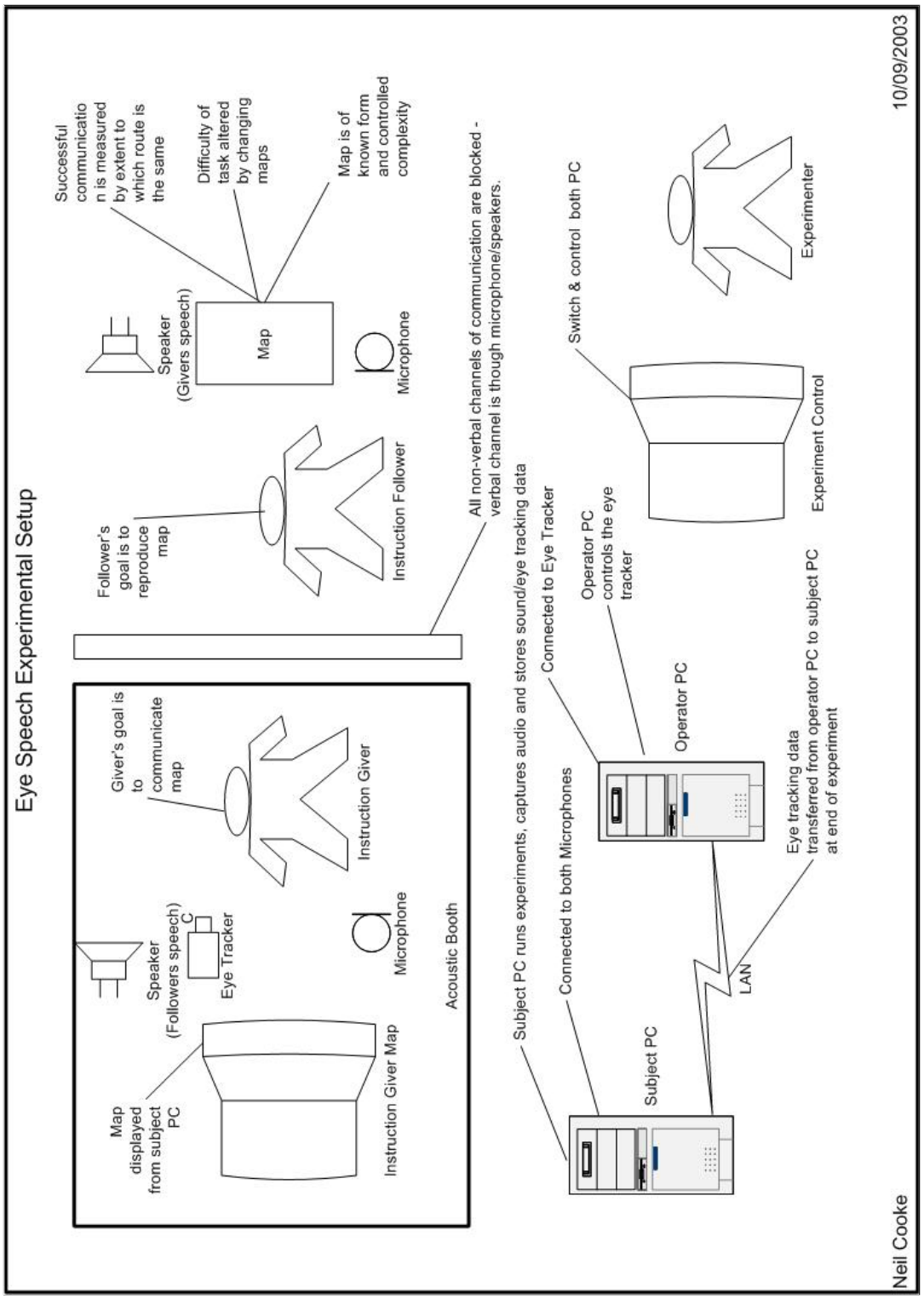


Figure 3.1: Experimental set-up complete with labelled key features of the experiment.

3.3.2 Session structure

Participants worked in groups of 3, forming 6 pairs of Instruction Givers and Instruction Followers, with each participant taking both the role of Instruction Giver and Instruction Follower twice. Each group used all 3 map sets, so that an Instruction Follower previously participating as an Instruction Giver was not replicating a map he or she had seen before. In total 18 sessions were recorded.

Each session was given a unique reference, based on the map set used and the participant involved. The format of this was:

```
M<map set number>G<participant number>F<participant number>
```

Each participant in a group of 3, was designated a unique participant number between 1 and 3. Likewise, the map sets were numbered between 1 and 3. Table 3.1 show the breakdown of sessions recorded per group.

Group no.	Sessions recorded
1	M1G1F2,M1G1F3,M2G2F1,M2G2F3,M3G3F1,M3G3F2
2	M1G2F1,M1G2F3,M2G3F1,M2G3F2,M3G1F2,M3G1F3
3	M1G3F1,M1G3F2,M2G1F2,M2G1F3,M3G2F1,M3G2F3

Table 3.1: Sessions recorded against group.

3.3.3 Materials

There are 3 map sets used and each map set consists of 3 maps. A map comprises a number of objects (landmarks) and a route through the objects. Each map in a map set had the same objects (landmarks) as other maps in the same set, but each shows a different route. The objects were detailed to encourage participants to describe, rather than simply name them. Figure 3.2 shows an example map. Following the convention in §3.3.2, maps are designated the code:

```
M<map set number>G<participant number>
```

E.g., Map M1G1 was used in sessions M1G1F2 and M1G1F3, and so on.

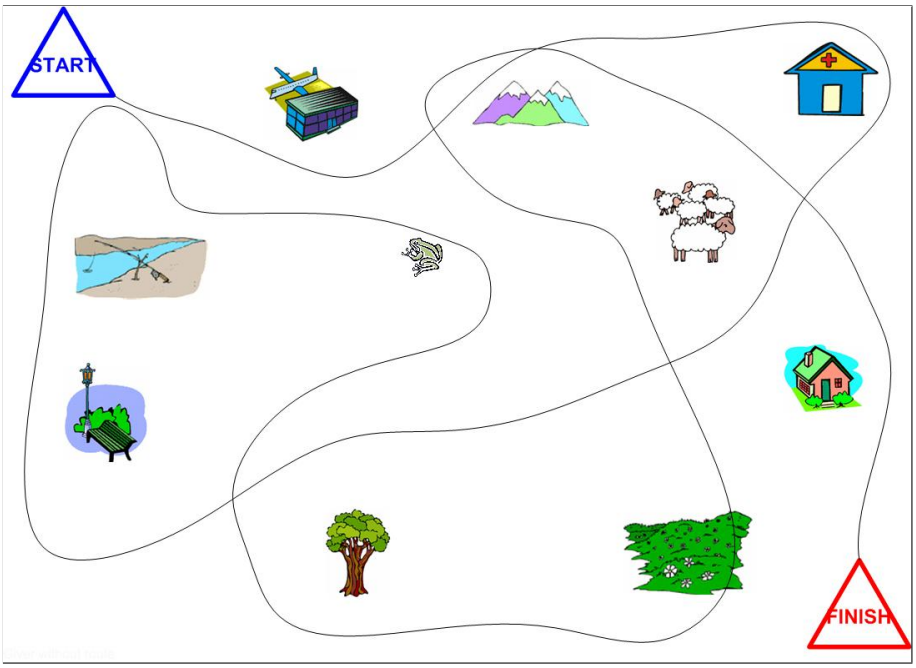


Figure 3.2: Example map from the Eye/Speech corpus.

3.3.3.1 Map landmarks

Landmarks on the maps were devised from the HCRC Map Task, which uses 4 sets of maps, with each map set using approximately 70 landmarks. The majority of landmarks differ between map sets.

Several object equivalence classes were determined from the objects in the HCRC Map Task. These were used to define the objects in the eye/speech corpus maps. Table 3.2 shows the equivalence classes and objects .

Unlike in the HCRC Map Task, object names were not given to the participants (that is, labelled on the map). This was so that the participants would look at the object, rather than read a label underneath it.

3.3.4 Recording

Each session lasted for approximately 30 minutes. Half of the allocated session time was consumed by calibrating the equipment, and the other half was spent capturing the data. The availability of the eye tracker, and of students, allowed 3 hours per week for recording between January and March 2004. Within the 3 hours, up to 6 sessions could be recorded, however in most weeks the number of sessions recorded was less

Equivalence class	Landmarks/ objects
Map	Start triangle; Finish triangle; Route
Land	Meadow; Fields; Hills
Animal	Frog; Sheep
Water	Pond; River; Waterfall; Stream
Park	Park Bench
Building	Train station; House; Factory; Airport; Hospital; Skyscraper
Rock	Cave; Mountain
Vegetation	Tree; Flower

Table 3.2: Landmarks in the eye/speech corpus maps

due to calibration problems and the non-attendance of participants. Consequently, each group of participants attended the lab on at least 2 occasions.

3.3.5 Prototype sessions

Prior to collecting the data, three trial runs were carried out. Two participants without recording apparatus did the first trial run. The participants sat with their backs to one another. The Instruction Follower was given a map that differed from that of the Instruction Giver. The task was for participants to agree the differences between maps, and for the Instruction Follower to correct his or her map. Each participant played both roles. This task bore a closer resemblance to the HCRC Map Task. On both occasions the task was completed in less than 3 minutes, which was not sufficient for capturing a good quantity of spontaneous speech, and nor could it justify the time required to set up the eye tracker.

The aim of the second trial was thus to evaluate a strategy for lengthening the task duration by modifying the task. This was achieved by removing the Instruction Follower's map, and changing the task so that the Instruction Follower had to replicate the Instruction Giver's map, and their route, onto a blank sheet of paper. Task duration doubled to 6 minutes. Consequently, the experiment used the modified task. Audio was recorded during this session, allowing a custom software sound recorder to be tested. Appendix A describes further details of the software.

The third trial was that of the final experiment, using recording apparatus. This allowed some practice in running through the eye tracker's calibration process, and testing the software. Collecting some early data enabled the development of analysis software to commence in-between recording the actual sessions. The third trial also

assisted in planning session durations, in order to allow for calibration and set-up.

3.3.6 Apparatus

3.3.6.1 Eye movement capture

A commercial head-mounted eye tracker, the Sensomotoric Instruments EyeLink System Version 1 [Gmb99], was used for capturing gaze. This system was designed for fixed laboratory use, tracking a subject's gaze direction in relation to a standard 17-inch VDU screen. The EyeLink system consists of a head-mounted camera system and two PCs for processing data and running experiments.

On the head-mounted device, both left and right eye pupil position was tracked using a miniature camera and infrared (IR) LED light source for each eye. A third camera was located on the head-mounted device and pointed in the direction of the participant's field of view. This camera captured head position relative to IR LEDs placed at each corner of the VDU screen, and therefore enabled the eye tracker to compute the position of the head relative to a position on the screen. Combining the position of the head with the pupil movement relative to the screen, enabled recording of the gaze direction. Figure 3.3 shows a participant wearing the head-mounted device.

The PCs used in the experiment were referred to as an 'Operator PC' and a 'Subject PC'. The terms 'Operator PC' and 'Subject PC' originate from the EyeLink user manual [Gmb99]. The Operator PC was connected to the head-mounted device and was responsible for collecting, processing, and storing the eye data during capture. The Subject PC was connected to the Operator PC via an Ethernet data link. The Subject PC was responsible for running the experiment - i.e. displaying images to the user and responding to user input. The eye tracking data was transferred from the Subject PC to the Operator PC at the end of the experiment for long-term storage and subsequent analysis. Figure 3.1 shows the PCs in the experimental set-up.

The Operator PC used a Pentium 3 600MHz CPU, whilst the Subject PC used only Pentium 2 233 MHz CPU. These were obsolete specifications; however, the performance of the machines was adequate for the task.

Eye tracking data for both eyes was captured at a sample rate of 250Hz (4ms samples) with a resolution of 0.005^o visual angle. The VDU screen against which participant's eyes are tracked is capable of displaying VGA-quality[2] images. The

[2]VGA screen resolution is 640x480 pixels.

Figure 3.3: Participant wearing the EyeLink eye tracker plus headphones sitting in front of VDU (off picture). Two cameras were situated in front of his eyes to capture left and right pupil position. A third camera was situated on his forehead to capture head position in relation to the screen.

distance between the user and monitor was approximately twice that of the screen width, as per manual guidelines [Gmb99], suggesting a visual angle of 0.043^o per pixel (to 2 d.p.). The effective resolution, however, was limited by the calibration process which is specified as providing an accuracy of typically between 0.5^o and 1.0^o visual angle. Given this limitation, the eye tracker can be considered to measure gaze direction within 23 pixels in the horizontal and vertical plane.

Each session commenced with calibrating the eye tracker. The calibration scheme consisted of the user viewing a series of regularly spaced markers shown on the VDU. The pupil positions detected by the head mounted cameras were then mapped to the corresponding position of the cross on the VDU. After viewing the series of crosses, the eye tracker estimated a linear transform to map the pupil position to the screen position for each eye. The transforms were then verified by showing the user the same series of markers. The calibration was repeated if it failed for both eyes. The verification scheme reported the calibration accuracy level for each eye. This information was recorded in the eye tracking data so that the eye with the best calibration could be selected for analysis.

In addition to recording gaze direction, the eye tracker also captured pupil size and detected eye movement types such as saccades, fixations and blinking.

3.3.6.2 Speech capture

The Operator PC used the Microsoft Windows 2000 operating system. Audio capture was via a software sound recorder that ran on the Operator PC. This PC was equipped with a specialist external sound card (model 'Edirol UA-25') that provided the facility for monitoring sound channels and ensuring high quality digitisation of audio. The sound was recorded at 22.05 kHz per channel 24-bit stereo in uncompressed WAV (PCM) format. The participants' voices were recorded on separate audio channels. The Instruction Giver's microphone was studio quality and desk mounted (model 'Shure SM48'). The Instruction Follower's microphone was a standard headset.

3.3.7 Synchronisation

Recording multimodal data streams in general requires some form of tagging (in the data) to ensure synchronisation between streams during reconstruction and analysis. For this experiment, speech capture software and eye experiment software were run in parallel on the same PC. The operating system's messaging handling enabled the speech capture software to instruct the eye experiment software to insert the current audio sample count into the eye tracking data stream at periodic intervals. This involved sending audio sample counts from speech capture software to the eye capture software. Latencies in message passing were accounted for by handshaking between the audio and eye tracking software resulting in each audio sample count being inserted twice into the eye tracking data stream. The delay between the two sample count insertions indicated the latency between eye and speech data capture in the system. Figure 3.4 outlines the synchronisation scheme using a UML sequence diagram.

3.4 Quality evaluation

Previous studies indicate that employing eye trackers for usability research is expensive and prone to error. E.g., 10%-20% of participants could not be tracked reliably [Jac03] ; a previous study managed only a 37.5% success rate [ST00]. The eye tracking data was therefore expected to be of variable quality. During the session recordings, an indication of the eye movement was available to the experimenter through the

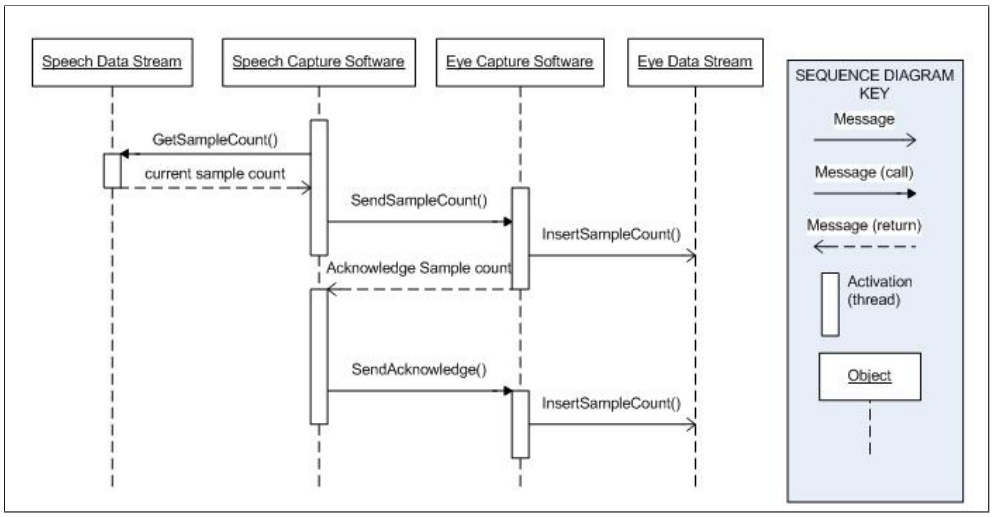


Figure 3.4: UML sequence diagram showing the synchronisation scheme.

eye tracker system. However, it was not easy to notice degradation in calibration until later analysis. Therefore, the data had to be subject to the quality evaluations described below.

3.4.1 Synchronisation performance

Inspection of the eye tracker data showed that the latency between speech capture and eye movement was consistently less than the eye tracker sample period (4ms). This meant that, given the resolution of the eye tracker data, the latency in the synchronisation scheme[3] would not have to be accounted for during data analysis. Figure 3.5 shows some output of the eye tracker data stream with audio synchronisation records. The 1ms latency for handshaking evident shows that inserting an audio sample time tag into the eye data took approximately 0.5ms.

3.4.2 Video production

The quality of eye recordings was evaluated by combining the raw eye and speech data to make a video file of each session, for playback using standard PC video playing software. Still images were produced by overlaying the eye position and fixation event information extracted from the eye tracker output file onto the session map. A still image was generated at 40ms intervals so that the final video would play at 25 frames per second. A cross hair was superimposed onto the image to assist viewers

[3] §3.3.7

```
TIME    | Left pupil             | Right pupil
        | Position      | Pupil  | Position       | Pupil
        | horiz | vert  | size   | horiz | vert   | size
(ms)    |   (pixels)    |        |   (pixels)     |
-----------------------------------------------------------
1370204   140.7   114.5   465.0    99.1    138.8   353.0
1370208   142.7   112.1   461.0    98.8    138.7   353.0
1370212   142.0   111.6   462.0    98.8    138.0   352.0
MSG 1370214 AUDIOSYNC 2499
MSG 1370215 AUDIOECHO 2499
1370216   141.2   112.2   465.0    99.1   137.6   352.0
1370220   139.6   111.5   465.0   100.1   136.8   352.0
1370224   139.7   108.9   459.0   100.0   136.6   351.0
1370228   140.1   109.5   458.0    99.5   134.4   348.0
1370232   140.4   110.0   458.0    99.5   134.3   348.0
1370236   140.2   111.0   465.0    99.7   134.2   349.0
1370240   138.0   110.6   476.0   100.3   134.1   360.0
EFIX L   1370072 1370240 172   142.0   115.4     470
EFIX R   1370068 1370240 176    99.1   136.8     358
SSACC L  1370244
SSACC R  1370244
1370244   142.6   107.6   476.0   104.5   130.6   360.0
1370248   147.5   105.3   476.0   108.7   129.4   360.0
```

Figure 3.5: Sample data output from the eye tracker. Eye position records (lines starting with a number) indicate time (in ms), horizontal and vertical eye position, and pupil size, for both left and right eye. MSG indicates the audio synchronisation - the time and sample count of the audio stream (i.e. 2499). Handshaking between audio capture and eye tracker software ('AUDIOSYNC' and 'AUDIOECHO' messages), shows a 1ms latency. 'EFIX' and 'SSAC' are fixation and saccadic eye movement events.

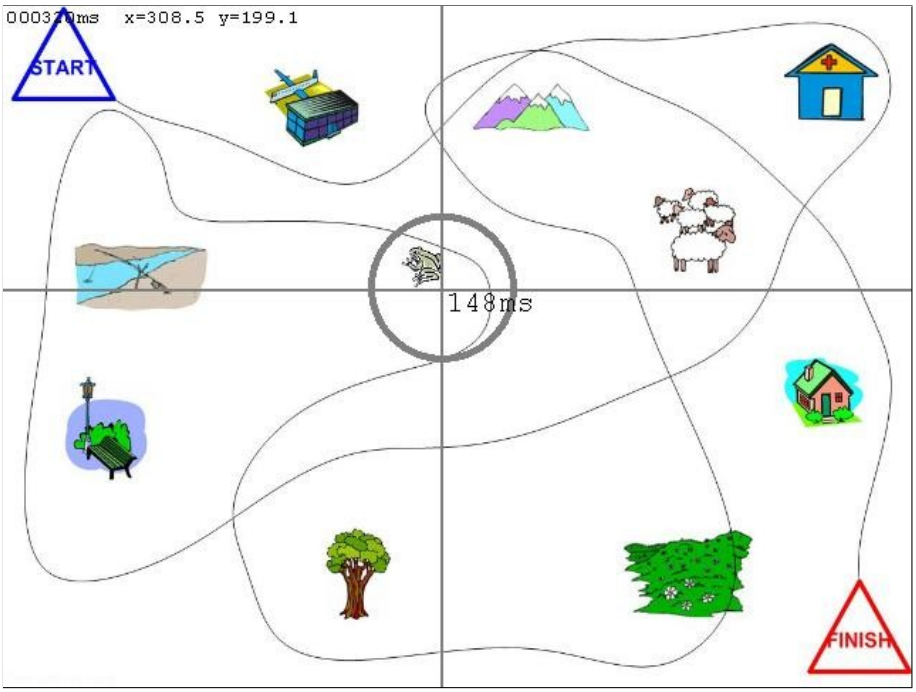


Figure 3.6: Screenshot of a video generated for the quality evaluation. The cross hair shows the eye position together with the duration of the current fixation.

in following the eye movement, in addition to the duration (in ms) of fixation events detected by the eye tracker. Using the analysis software described in Appendix A, still images were generated. These were strung together to create a video sequence using 'JPGVideo' software [JPG], which produced a DIVX 5.1.1 [4] compressed video file. Corresponding audio files were compressed from the original WAV-format audio files (22.05 kHz per channel stereo) to MP3 (at 56 Kbit/s 22 kHz stereo). The sound was added to the DIV-X file using the video capture/processing utility 'VirtualDub' [Vir], an open source software licensed under the GNU General Public License (GPL) [GNU]. Figure 3.6 shows a still image from a video.

Initially, the analysis software was produced to playback videos interactively by combining the video and audio streams at run-time using the synchronisation information embedded in the eye data. This approach was abandoned as it proved memory intensive and less flexible than generating compressed video sequences.

A potential shortfall in generating videos was that synchronisation information was not fully utilised. However, variations in synchronisation were offset by setting the video processing utility to resample the audio recording for a session, so that

[4] DIV-X is an MPEG-4 compression technology

audio recordings were equal in length to that of the video generated from the still images. This was a valid approach, since recordings were initiated and terminated by the EyeLink software using subsequent EyeLink experiment code instructions. Without time stretching, audio recordings were shorter than the video sequences, the likely explanation for this being variations in the nominal eye tracker sample rate and variations in the quartz clocks of the two PCs used in the experiment. The difference in the lengths of the recordings underpins the requirement to synchronise multimodal streams during capture, particularly when using different platforms.

3.4.3 Calibration errors

Each video was viewed from start to finish in order to assess the general quality of eye tracking and audio data. A subjective quality assessment was effective for quickly determining poor recordings. In addition, the times at which the participants were describing the landmarks, as opposed to the route, was noted for use in eye movement analysis in §5.10.

In many sessions, eye movement recordings showed poor calibration. Rejection criteria (as shown in Table 3.4) were horizontal and/or vertical offsets, loss of gaze altogether, and general corruption in the eye signal, making the eye movement unintelligible insofar as the expected pattern of fixations, saccades and microsaccades was non-evident or sporadic.

The calibration offsets were identified from multiple fixations resting in blank regions of the map, or off the screen altogether. These occurred despite the eye tracker reporting good calibration prior to the experiment commencing, as indicated in the headers of EyeLink output files.

There are a number of causes for calibration failures. Overt indications of errors during recording occurred if a participant inadvertently moved the eye tracker when coughing, gesticulating or involuntarily touching his or her face. As the human scalp is elastic, the head mounted device was also liable to moving.

The eye tracker was also sensitive to the ambient light levels in its environment. The bright white background of the maps displayed on the screen altered the ambient light level enough to affect the eye tracker pupil detection. The standard EyeLink calibration screens had dark backgrounds that contrasted greatly with the maps' white backgrounds. This effect was reduced in later sessions by turning down brightness of the screen. In hindsight, using calibration screens with similar contrast to experiment screens would have been beneficial.

Software was produced so that the distribution of eye positions about the image for the entire session could be visualised. This enabled the identification of offsets in eye tracking calibration. In addition to the videos described in §3.4.2, surveying the distribution of eye positions using a frequency histogram provided a broad indication of data quality. If the histogram for a session did not show relatively high frequencies for eye positions around each object, then a calibration failure was suspected and, if verified by viewing the video, the session was rejected. §3.4.8 describes the session selection.Table 3.4 summarises the video evaluations for the sessions and shows the histograms.

Four sessions were identified as potential beneficiaries of correcting calibration offsets. The correction proved problematic as the calibration offset was found to be non-uniform across the visual field and not constant throughout the session duration; sessions commenced only after successful calibration and offsets occurred or amassed during the session.

3.4.4 Instruction follower's maps

The Instruction Followers' recreated versions of the Instruction Givers' maps were electronically scanned for storage. Although it is plausible that the success of the communication between participants could be measured by assessing the similarity of the recreated map to the original, as done in the HCRC Map Task, its validity would be abated as maps created by Instruction Followers who had previously taken the role of Instruction Giver were, without exception, more detailed than those where Instruction Followers had not seen an example map. This was expected. See Figure 3.7 and Figure 3.8 for examples of the maps that Instruction Followers drew.

3.4.5 Audio Post-processing

Speech was transcribed using the 'Transcriber' software [BGWL01], which outputs transcripts in a meta-data (XML) format. The corpus consists of 35920 utterances from a 1116 word vocabulary. Session durations range from 5 to 15 minutes as shown in Table 3.4. Transcriptions were made in 2 passes by 2 different transcribers. Students were employed as transcribers. The transcribed texts were converted into label-format files, the format recognised by HTK speech recognition software [ea02].

The audio was recorded at a sample rate of 22.05 kHz 24 bit levels per channel. Each channel was resampled to 16 kHz in order to ensure compatibility with the

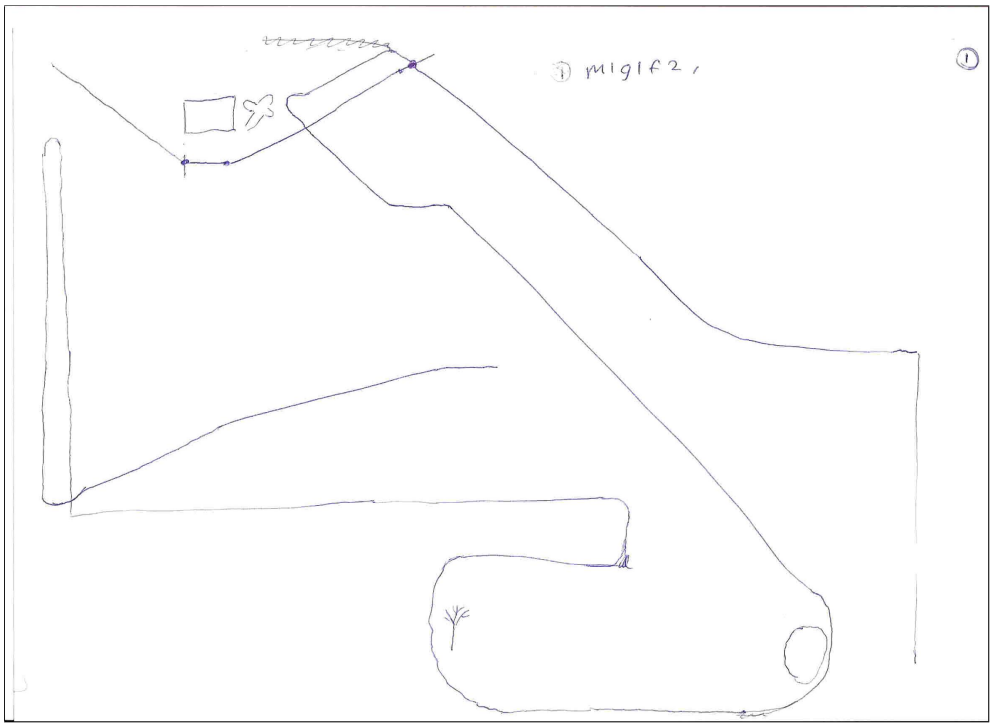


Figure 3.7: Instruction Follower's map before having participated in the experiment as an Instruction Giver.

Baseline ASR System that was trained on 16 kHz data. Chapter 4 describes the baseline ASR system and HTK.

3.4.6 Speech files

The evaluation of the videos of the sessions, described in §3.4.2, showed the audio quality to be consistently good. The Instruction Givers' speech, recorded in an acoustic booth, was high quality. The Instruction Followers' speech, recorded outside the booth, was lower, but acceptable, quality. This was expected due to the poorer microphone used by the Follower and ambient noise in the laboratory. The ambient noise included the Giver's speech from an external speaker, which was rectified in later trials by making the Follower wear headphones. Since these headphones were of the lightweight, open variety typically used with portable music players, wearing the headphones did not noticeably affect the Follower's speech.

The speech was spontaneous and informal, with disfluencies present. Speakers would stutter, talk over one another, and speak quickly. The Giver's speech was, as expected, richer in terms of vocabulary. The Follower's speech mainly comprised of

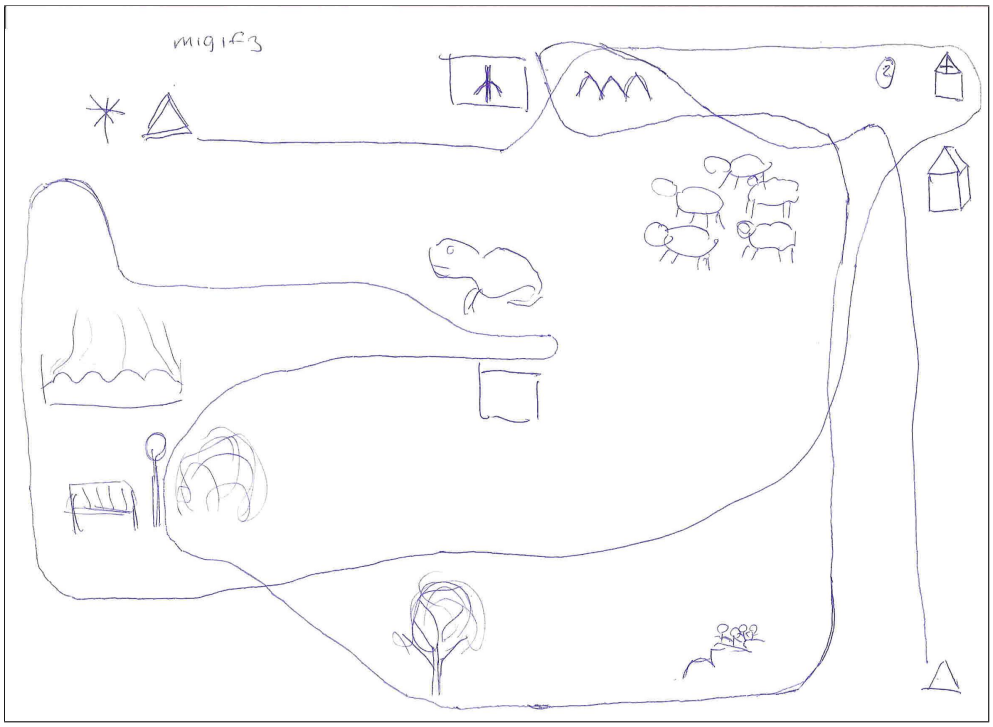


Figure 3.8: Instruction Follower's recreated map after having participated in the experiment as an Instruction Giver.

short affirmations ('ok', 'yes') or sought clarification. Figure 3.9 shows an example of a typical dialogue.

3.4.7 Keywords corresponding to map features

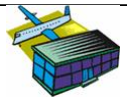
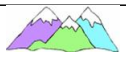
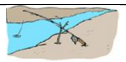
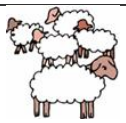
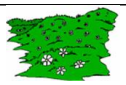
The map landmarks, described in Table 3.2, constitute potential FOVAs. Words referring to each landmark were identified from the transcriptions. These keywords are used in the integration experiments in Chapter 6. Table 3.3 lists the keywords together with their associated landmark.

Table 3.3

Landmark	Keywords	Picture	Used in map set
Start triangle	start		1,2,3

Continues on next page

Table 3.3 – continues from previous page

Landmark	Keywords	Picture	Used in map set
Airport	airport, aeroport, airplane, plane, aeroplane, hangar		1,2,3
Mountains	mountain, mountains		1,2
Hospital	hospital, church, cross		1,2,3
Fishing Rod	river, stream, fishing rod		1,2
Frog	frog, frogs		1,3
Sheep	sheep, sheeps		1,2,3
Park Bench	park, bench, lamp		1,2
House	cottage, house		1,2,3
Tree	tree		1,2
Meadow	meadow, field, grass, flowers		1,2,3
Finish Triangle	finish		1,2,3

Continues on next page

Table 3.3 – continues from previous page

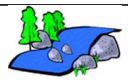
Landmark	Keywords	Picture	Used in map set
Skyscraper	skyscraper, block, flats		2
Hills	hills, rolling		3
Waterfall	river, rivers, stream, waterfall		3
Cave	cave, rock		3
Pond	pond, island		3

Table 3.3: The keywords used to describe map landmarks, together with the landmark picture and the map set(s) that contain the landmark.

3.4.8 Session selection

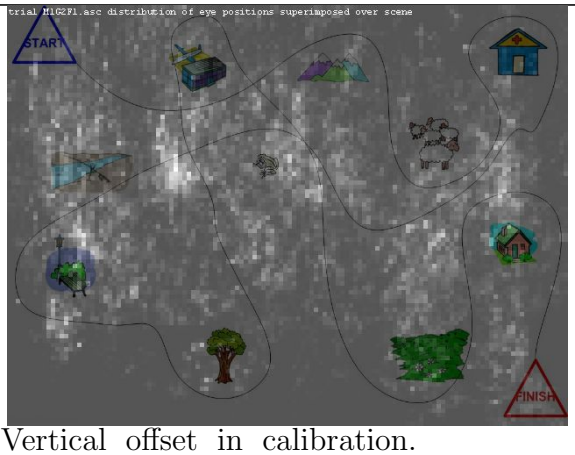
Of eighteen sessions undertaken, seven were deemed to have an acceptable quality of eye tracking data. Given the previous results of eye tracking experiments [ST00], a success rate of 33% is typical and underpins the fact that eye tracking studies are costly. Any research should be designed with this expectation so that the initial aims of the corpus design are sustained.

The sessions rejected are still useful for the speech recognition element of the research in Chapter 4. Table 3.4 shows a summary of the corpus sessions and quality evaluation, including the frequency histograms displaying the distribution of eye positions over the visual field.

```
G: right
G: top left triangle with the word start in it
F: ok
G: right now bottom left triangle with the word finish in it
F: yep
G: right
G: if you go from the start yeah in the middle of the page should be
G: a herd of four or five sheep
G: smack bang in the middle this is
F: right yeah
G: right now in the far right
G: should be a  1d like house 2d house sorry like a
G: red cross on it like a hospital
F: what what up or down is there a middle
G: its  to the right
F: yeah top right
F: of the sheep top right
G: top right
F: i thought you said the sheep was in the middle
G: sheep were in the middle
G: oh the house is top right ok got you
F: yeah go on then
G: right now if you go from the sheep yeah
G: just down to the left there should be a bunch of hills
G: couple of hills
F: yeah this between the finish and the sheep
G: yeah but it shouldn't be it should be just above the middle of the page
G: just above the middle
F: yeah
F: the sheeps in the middle
G: sheep are in the middle at the top
F: middle at the top
F: ok
G: right
G: the hills are probably where you've just drawn the sheep
G: to the left of it
F: oh i've just drawn it from down from the sheep
G: you've just drawn it down from the sheep
G: yeah
F: sort of to the left
F: well to the right of that other side of the sheep
```

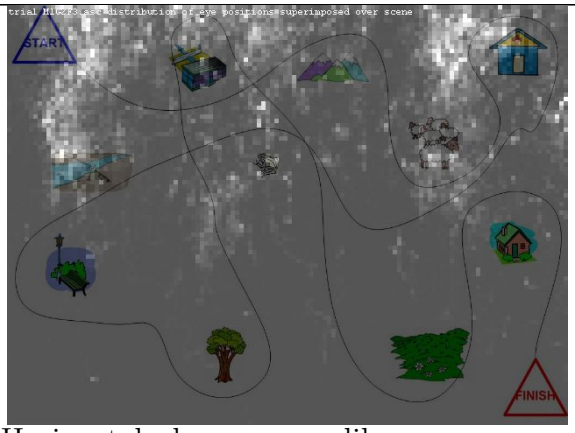
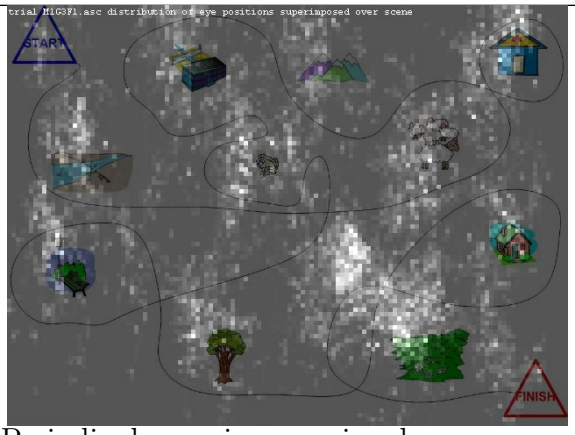
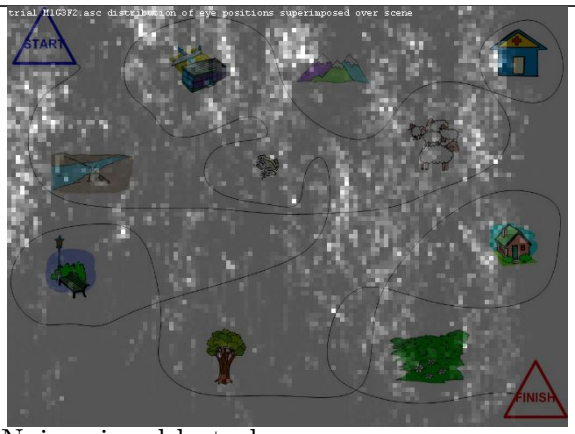
Figure 3.9: Sample transcript from the eye/speech corpus. 'G' indicates the Instruction Giver's speech, 'F' the Instruction Follower's.

Table 3.4

Session	Frequency Histogram	Duration	Eye data
m1g1f2	 Good calibration. Some Signal loss.	13:43	
m1g1f3	 Acceptable calibration.	10.13	✓
m1g2f1	 Vertical offset in calibration. Horizontal plane ok. Eye signal periodically lost.	16.06	

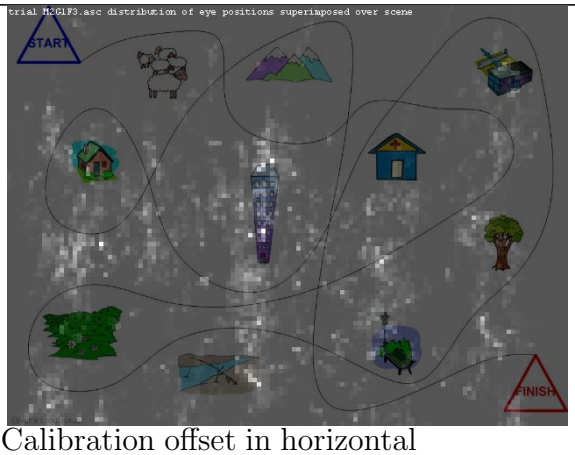
Continues on next page

Table 3.4 – continues from previous page

Session	Frequency Histogram	Duration (mm:hh)	Eye data usable
m1g2f3	 Horizontal plane poor calibration, vertical plane ok.	15.24	
m1g3f1	 Periodic losses in eye signal. Calibration ok.	13.56	✓
m1g3f2	 Noisy signal but ok.	14.43	✓

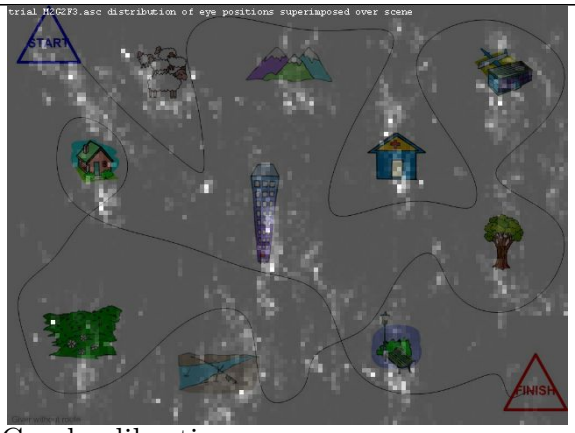
Continues on next page

Table 3.4 – continues from previous page

Session	Frequency Histogram	Duration (mm:hh)	Eye data usable
m2g1f2	 Calibration degrades during experiment in horizontal plane. Vertical plane ok.	14.54	✓
m2g1f3	 Calibration offset in horizontal plane. Vertical plane ok.	9.30	
m2g2f1	 Loss of signal throughout. Reject recording.	6.10	

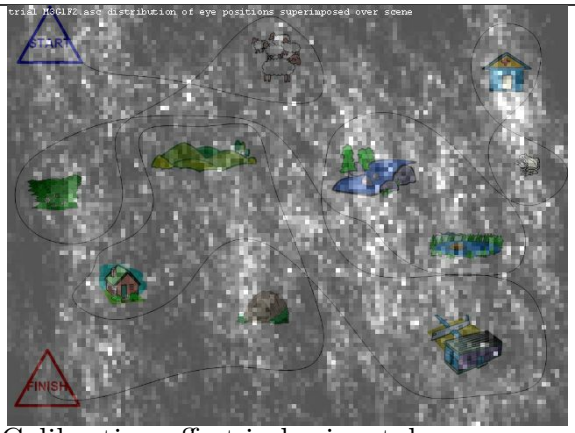
Continues on next page

Table 3.4 – continues from previous page

Session	Frequency Histogram	Duration (mm:hh)	Eye data usable
m2g2f3	 Good calibration.	5.14	✓
m2g3f1	 Eye signal drops out periodically for short ($< 1s$) intervals.	20.22	

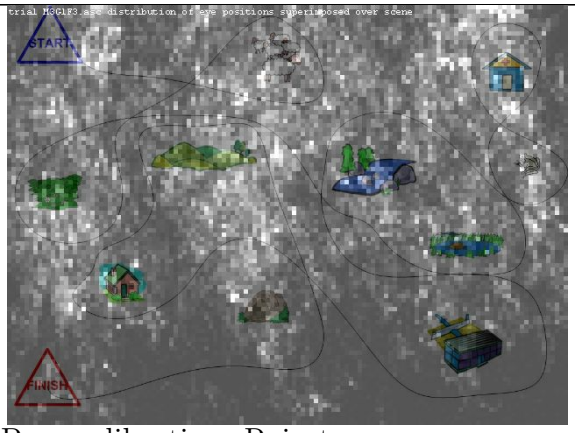
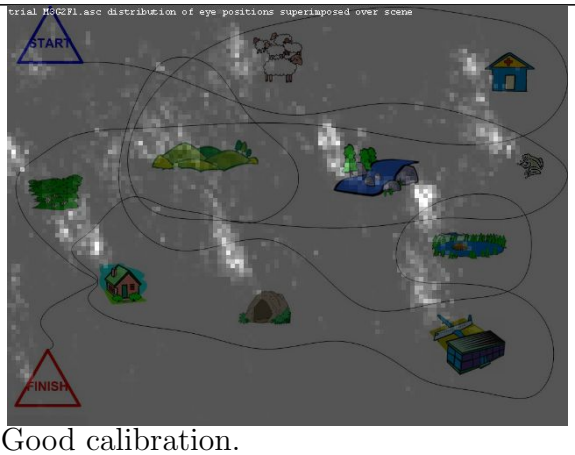
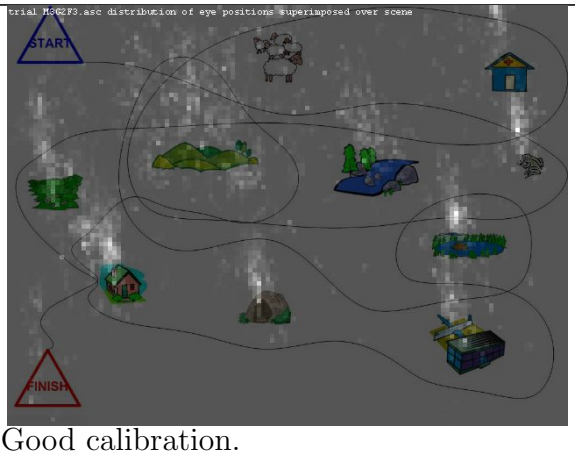
Continues on next page

Table 3.4 – continues from previous page

Session	Frequency Histogram	Duration (mm:hh)	Eye data usable
m2g3f2	 Flickering in eye data to top right corner from 3 min onwards. Complete loss about 4 minutes. Recover 5 minutes. Out of calibration.	20.48	
m3g1f3	 Calibration offset in horizontal plane. Some flickering.	20.22	

Continues on next page

Table 3.4 – continues from previous page

Session	Frequency Histogram	Duration (mm:hh)	Eye data usable
m3g1f3	 Poor calibration. Reject.	16.17	
m3g2f1	 Good calibration.	10.53	✓
m3g2f3	 Good calibration.	9.44	✓

Continues on next page

Table 3.4 – continues from previous page

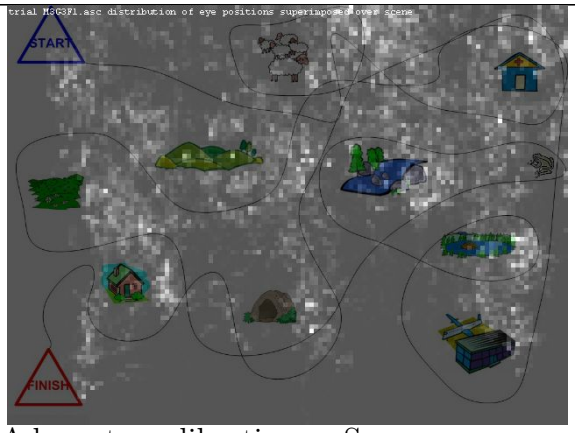
Session	Frequency Histogram	Duration (mm:hh)	Eye data usable
m3g3f1	 Adequate calibration. Some interim loss of data.	8.02	
m3g3f2	 Adequate calibration.	10.00	

Table 3.4: Quality evaluation of the eye/speech corpus.

3.5 Summary

This chapter documented the design of the eye/speech corpus. Some important lessons were learnt while collecting it. Tracking eye movements with current eye tracking technology is problematic and attempting to capture eye movement while participants undertook a task that involved conversational speech often resulted in a loss of eye tracking calibration. This loss was primarily due to the involuntarily movement of the participant, even if he or she was concentrating on the task in-hand.

The tracking of multiple modalities requires synchronisation schemes. In the eye/speech corpus, even though the computers used were obsolete specifications, com-

munications between the sub-programs capturing the individual modalities was fast enough to ensure that the latency in inserting a synchronisation tag into the data streams was not an issue.

There were two opposing constraints on the task duration. Capturing a large volume of data required long task durations, however the eye tracker's lack of reliability in maintaining calibration meant that the sessions that lasted the longest did not contain reliable eye tracking data [Hel61]. In rejecting some sessions due to the loss in calibration, some envisaged uses for the eye/speech corpus regarding person-specific eye movement analysis were compromised. The main objective of the corpus was not compromised though, and rejected sessions were used in the development of the baseline ASR system, which is described in Chapter 4.

4 Baseline ASR System

A speaker independent continuous ASR system was required to establish a baseline for ASR performance on the speech collected in the eye/speech corpus. This enabled measurement of the effect on performance of adding information from eye movements.

After covering some speech recognition basics, this chapter discusses:

- The corpora used for supervised training of the ASR (§4.3).
- Model building, structure, and parameter optimisation (§4.4, §4.5, §4.6).
- Performance benchmarking of the ASR against standard test sets (§4.7).
- Adaptation of the ASR system to the eye/speech corpus (§4.8).
- ASR performance of the system on the speech data in the eye/speech corpus (§4.10).

4.1 Speech recognition basics

The most common and successful ASR systems use Bayesian inference to find the most likely word sequence $\widehat{W}$ from a sequence of acoustic vectors Y:

$$P(\widehat{W}|Y) \propto \arg\max_i P(Y|W_i)P(W_i) \tag{4.1}$$

Speech is treated as a hidden stochastic process that represents a sequence of sub-word units and a visible stochastic process that is the acoustic vector sequence (the visible stochastic process being conditioned on the hidden). HMMs are used to find the class conditional probability of a sequence of acoustic vectors given a possible word sequence $P(Y|W_i)$, and a language model provides a prior for word sequence probability, $P(W_i)$. §2.4.3 described HMMs.

4.2 Performance measures

4.2.1 ASR

Two standard measures of ASR performance are used. Word Error Rate (WER) measures the proportion of the total number word substitutions, N_s, deletions, N_d and insertions, N_i, in the ASR output relative to the total number of words, N, in the transcription:

$$WER = \frac{N_i + N_d + N_s}{N} \tag{4.2}$$

A basic keyword spotting accuracy measure would be the ratio of true hits to false alarms of keyword occurrence over the entire test data. Motivated by the large variances (i.e. instability) in this statistic when evaluating word spotting systems in the early 1990s, the NIST derived Figure Of Merit (FOM) [NIS91] measures keyword spotting accuracy averaged over 1 to 10 false alarms per hour per keyword, with p_f being the % words spotted correctly before the f'th false alarm and T being the speech duration in hours:

$$FOM = \frac{1}{10T} \sum_{f=1}^{10} p_f \tag{4.3}$$

FOM is a useful complement to the more frequently used WER because it bases its estimate of performance on the ability to detect words pertinent to the application, regardless of their frequency of occurrence in common language. In selecting keywords, words that occur frequently but are unrelated to the application such as 'the', 'and' and 'at' etc. are ignored. From an information theoretic perspective, commonly occurring words have low information content and are susceptible to occurrence by chance in the WER measure. This is particularly so in small vocabulary systems. Unlike the standard WER, calculating the FOM requires time-aligned transcriptions. Reporting the component True Positive (TP) and False Alarm (FA) counts supplements the FOM.

For a gaze-contingent ASR system, the effect of integrating eye movements on performance can be measured by considering the set of words related to visual attention. Since the FOM enables the word recognition performance to be measured for a subset of the recogniser's vocabulary, it serves as the primary performance measure for the gaze-contingent ASR systems in Chapter 6.

4.2.2 Perplexity

Language model performance is measured using perplexity (PER), which represents the average predictability of a language model to generate each word for a given word sequence, $W = w_1, \ldots, w_N$, of N words:

$$PER = \frac{1}{\sqrt[N]{P(W)}} \tag{4.4}$$

PER values range from 1 to ∞. Higher values of perplexity indicate less predictability in the language model and are less desirable. Whilst the PER is useful in assessing the language model component of an ASR system independently of the acoustic models, empirical evidence suggests that a language model selected from a list of candidates chosen for having the lowest PER may not necessarily be the best candidate for lowest WER [CR97] [CR98]. Other measures of language model performance have been proposed that may better correlate with WER [CR99].

4.3 Training corpora

The baseline ASR system required supervised training using existing corpora[1] for language and acoustic models. The speech data collected in the eye/speech corpus was from British native speakers and did not provide enough data to train a large vocabulary ASR system. The WSJCAM0 [RFP+95] corpus was used for training the acoustic model set. The BNC [Bur00] and HCRC map task [ABB+91] corpora were used for language modelling. The selected corpora are summarised in the following sections.

4.3.1 WSJCAM0

The Wall Street Journal Cambridge '0' (WSJCAM0) [RFP+95][2] corpus was published in 1994 by Cambridge University. It contains recordings of read speech from 140 native British speakers and corresponding word and phone-level transcriptions. The text that speakers read was sourced from the Wall Street Journal, an American

[1] All the corpora in this section are listed and available from the Linguistic Data Consortium (LDC) `http://www.ldc.upenn.edu`

[2] The WSJCAM0 corpus is based on a subset of the Wall Street Journal '0' (WSJ0) corpus[PB92]. WSJ0 consists of WSJ read text spoken by American native speakers. The University of Pennsylvania published WSJ0 in 1993. It is a product of the American Defence Advance Research Projects Agency (DAPRA) Continuous Speech Recognition (CSR) project.

business and financial news publication. The corpus is split into training, evaluation and test sets. There are test sets for a 5,000-word and a 20,000-word vocabulary. These test sets are used for benchmarking of ASR systems. The audio bandwidth is 8 kHz.

WSJCAM0 was selected for training the acoustic model set based on the criteria that the corpus should have recordings spoken in British English, with similar dialects and similar accents to those in the eye/speech corpus. It also has a comparable audio bandwidth.

It is also possible to derive a language model from the WSJCAM0 data[3], and this was done to enable the acoustic model set to be benchmarked against other systems in the research community[4]. This language model was only used for testing against the WSJCAM0 test sets for benchmarking, since the vocabulary and language structure was specific to financial news.

4.3.2 British National Corpus

The British National Corpus (BNC) [Bur00] was first produced and published by Oxford University Press in 1994. It consists of texts of spoken and written modern British English from the late 20th century. The texts are marked with meta data using Corpus Document Interchange Format (CDIF). CDIF is an application of Standard Generalised Mark-up Language (SGML) - a precursor of today's widely used Extensible Mark-up Language (XML). The entire corpus consists of 4124 texts containing 100,106,008 words. The spoken part of the corpus comprises of 915 texts containing 10,365,464 words. The texts corresponding to the spoken part of the BNC is suitable for constructing a language model of general spoken British English.

The selection of the BNC corpus had different criteria from that of selecting corpora to train the acoustic model set. Word sequence probabilities depend on vocabulary, syntactic and semantic constraints. These constraints in the eye/speech corpus relate to the task domain, i.e. describing the map, map features, relative spatial positions and map route. Ideally, a language model for use in an ASR system must be trained using language that arises in the task domain where the ASR is used; however, such data was not available in sufficient quantity, if at all. To alleviate the problem of data sparsity, transcripts of speech from multiple task domains were used,

[3] §4.6
[4] §4.7

leading to a language model representing common-use spoken English. The BNC corpus fulfilled this role.

4.3.3 HCRC Map Task

The eye/speech corpus is based loosely on the HCRC map task [ABB+91][5]. The HCRC map task transcriptions consisted of 18,013 utterances with 149,859 words in total and 2160 unique words.

The BNC-based language model for common use English was adapted to the task domain using the HCRC map task. The HCRC map task has insufficient data for developing a language model covering vocabulary in the eye/speech corpus but it contains language relating to describing maps that provides some useful syntactic information. In this regard, language-modelling adaptation utilizes the HCRC map task, described in §4.8.

4.3.4 Other corpora considered

Other corpora were considered for training the baseline ASR. The TIMIT corpus [GLF+93] contains read speech from 630 American adult speakers, with an audio bandwidth of 8 kHz. The Switchboard corpus [GHM92] consists of 120 hours of telephone conversations between American adult speakers; with an audio bandwidth of 4 kHz, WSJCAM0 was chosen over these alternatives primarily because it uses British native speakers. The Switchboard corpus was deemed unsuitable due to its lower audio bandwidth.

The preference for using corpora with either read or spontaneous speech would ideally be spontaneous speech. In the absence of such an available corpus, British read speech (WSJCAM0) was preferred over using a corpus of spontaneous speech with lower audio bandwidth and American accent speakers.

4.4 Development platform

4.4.1 Software tools

The ASR system was built using the Hidden Markov model ToolKit (HTK) [ea02]. HTK is a set of software programs used to create, train, and evaluate small to large sized vocabulary HMM-based continuous ASR. It originated from the Cambridge

[5] §3.2

University Engineering Department (CUED) in 1989 and is today licensed to them from Microsoft for redistribution to the research community. CUED maintains control of releases and periodically issues updates. HTK capabilities lag behind state-of-the art ASR technology - e.g. the current release of HTK (3.1) does not support trigram language models. However, HTK is mature and reliable due to its widespread use.

The process of building an ASR system requires various text processing and extraction utilities. These were developed using Perl, Shell script, and C# programming languages. For the sake of brevity, the process description omits details and use of these utilities.

4.4.2 Hardware

The ASR was developed under the Linux Red Hat 9.0 operating system using a cluster of 7 Intel Pentium-4 GHz-class PCs running Sun Microsystems' Grid Engine Distributed Source Management (DRM) software. HTK was compiled using the Intel C++ Compiler (ICC), which incorporates processor optimisations into the compiled HTK binaries. The use of compiler optimisations resulted in reduced processing time compared with the standard GNU GCC compiled binaries. For example, a Viterbi decoding task using ICC compiled version of the HTK tool HVite took 49 minutes - a reduction from 67 minutes in processor time using the GCC compiled binaries. For training HMM models the GCC and ICC compiled version of the HTK Tool HEREST performed similarly.

4.5 Acoustic model set

Developing the acoustic model set required training, parameter selection, and performance benchmarking. An acoustic model represents a distinct sub-word unit, SW_i, of speech and enables the class conditional probability of the sequence of acoustic vectors given the sub-word, $P(Y|SW_i)$ to be calculated.

A left-to-right 3-state HMM model with parameter set $\{A, \pi, B\}$ was used for each triphone. State observation PDFs, $b_i(y)$, were modelled as N Gaussian mixtures, $b_{i,k}(y)$:

$$b_i(y) = \sum_{k=1}^{N} w_k b_{i,k}(y) \tag{4.5}$$

Where $b_{i,k}(y)$ has D dimensions and a mean, $\mu_{i,k}$ with diagonal covariance $\Sigma_{i,k}$[6,7]:

$$b_{i,k}(y) = \frac{1}{|\Sigma_{i,k}|^{\frac{1}{2}}(2\pi)^{\frac{D}{2}}} \exp \frac{1}{2}(y - \mu_{i,k})^T \Sigma_{i,k}^{-1}(y - \mu_{i,k}) \tag{4.6}$$

The process for training the acoustic model set closely follows the building of triphone acoustic models in the HTK tutorial [ea02]. There are the following additions: Front end processing uses cepstral mean normalisation. State output PDFs use Gaussian mixture densities rather than a single Gaussian for a finer parametric estimation. 'Skip' transitions are incorporated to better model variability in speaker pronunciation and speech rate. This section describes these enhancements together with the training process.

4.5.1 Acoustic feature extraction

The WSJCAM0 speech recordings for each trial consist of compressed waveforms for each recording from both a head and desk-mounted microphone at 8 kHz audio bandwidth. The head mounted microphone data was used. The waveforms were decompressed using the National Institute for Standards and Testing (NIST) Speech group's corpus building software tool 'SPeech HEader REsources' (SPHERE) [GF93].

The feature vector representation chosen was Mel-Frequency Cepstral Coefficients (MFCC) [DM80]. This is the de-facto method for feature extraction in ASR as MFCCs are not strongly correlated. The human perceptual mechanism and the source-filter model of human speech production motivate its use.

The MFCCs were extracted as follows. A discrete Fourier transform $\mathcal{F}_\omega$ was applied to overlapping windows of sampled speech, with waveform x_t and length T, to obtain the short-term log-power frequency spectrum for each window:

$$\mathcal{F}_\omega = \sum_{t=0}^{T-1} x_t e^{\frac{\omega j t}{T}} \tag{4.7}$$

Where ω represents the angular frequency. A Hamming window function, $w(t)$, was applied to each window to attenuate samples from the centre of the window towards the window edge as a function of time t:

[6] Diagonal covariance is where it is assumed that dimensions vary independently of one another.
[7] $\sum_{k=1}^{N} w_k = 1$

$$w(t) = 0.54 - 0.46 \cos\left(\frac{2\pi t}{T-1}\right) \tag{4.8}$$

The window length, T, was constant and chosen under the assumption that the articulatory system should be approximately stationary throughout the window. The commonly used value for T of 25ms was used. The windows overlapped in time by half the window length.

26 log-power Mel-spectrum features were obtained for each speech window by applying to the short-term log power spectrum the same number of triangular band-pass filters spread out along the Mel-scale [SVN37] [ea02]. The Mel-scale is motivated by humans discerning frequencies in the lower spectrum better than higher frequencies.

The filter bank responses were heavily correlated due to overlapping frequency bands. Decorrelation and compression of the feature space was achieved by applying the 'discrete cosine transform' (DCT) [Bli93] which yielded, by definition, a set of 12 Mel frequency cepstral coefficients $\{mfcc_1, \ldots, mfcc_{12}\}$:

$$mfcc_i = \sum_{j=0}^{25} m_j \cos\left(\frac{\pi i(j+\frac{1}{2})}{26}\right) \tag{4.9}$$

Where $mfcc_i$ is the i^{th} MFCC derived from the set of 26 Mel frequency features $\{m_1, \ldots, m_{26}\}$.

In addition to the 12 MFCCs[8] an energy term ($mfcc_0$) was also generated, providing an indication of the speech level for the speech window relative to the entire speech segment. It is proportional to the average log-power spectrum energy of the filter banks, normalised against the maximum average energy observed in all speech windows.

4.5.1.1 Trajectory features and normalisation

The MFCC feature vectors were further refined by applying Cepstral Mean Normalisation (CMN) [Ata74] to produce a set of CMN MFCCs. CMN accounts for the transfer characteristics of the speech-recording environment by subtracting from each MFCC in each window the average value for the MFCC across all speech windows:

[8]The choice of 12 coefficients is based on consensus from empirical results. For a recent study on coefficient variation, see [GWKB04].

$$MFCC_i^{\text{cmn}} = MFCC_i - MFCC_i^{\text{norm}} \tag{4.10}$$

Where $MFCC_i^{\text{cmn}}$ is the i_{th} CMN MFCC and $MFCC_i^{\text{norm}}$ is the average value of $MFCC_i$ for all speech frames.

In addition to CMN, the trajectory in time for sequences of MFCC vectors was represented by calculating the first and second order time derivatives of each MFCC. This increased the dimensionality of MFCC feature space, as each derivative constitutes a new feature. The first order derivative is usually referred to as the delta coefficient and the second order derivative is called the acceleration (delta-delta) coefficient. These derivatives have been shown empirically to increase ASR performance [Fur86] and one potential explanation for this is that they partially alleviate the shortcomings of the piecewise-stationary assumption inherent within HMMs' independance assumptions. Applying CMN to the 13 MFCCs and calculating time derivates for each coefficient resulted in a 39 dimensional feature vector representing each speech window (frame).

4.5.2 Training procedure

The HMMs were trained with Baum-Welch Expectation-Maximisation (EM) in HTK, using the WSJCAM0 phonetic transcriptions. The training data consisted of 90 utterances from each of 92 speakers. The acoustic model set was built in steps by increasing the complexity of the model set at each step, then reestimating the model set parameters using Baum-Welch EM before adding more complexity.

Model complexity was built up as follows: A 44 monophone model set[9] was created, each with a single Gaussian PDF for each state; each model corresponded to a phone. Then, a 19,189-triphone model set consisting of all triphones that exist in the WSJCAM0 corpus was created from the set of 44 monophones.

Reestimation of the triphone model set was not reliable due to data sparsity as there were, by definition, fewer examples of each triphone in the training data compared to monophones. This resulted in minimal variances in the reestimated state observation distributions. To overcome this, the states in the triphone model set were tied[10] using a phonetic decision tree [YOW94]. The decision tree rules are

[9]The phone set used is from the International Phonetic Alphabet (IPA) [Ass99] based on phonemes in Received Pronunciation (RP) English.

[10]Tied states are HMM states grouped together based on the similarity of their state output PDF.

supplied with HTK. State tying enabled robust reestimation, as triphones unseen in the WSJCAM0 data could be incorporated by tying their states to those of seen triphones. Triphones that occur in the eye/speech corpus but not in WSJCAM0 were added. This increased the number of triphones supported (logical models) to 25,321. The actual number of HMMs used (physical models) was 12,015.

Finally, the state observation single component Gaussian PDFs in the triphone model set were replaced by Gaussian mixture PDFs with diagonal covariance, enabling a finer-grain parametric estimation of the state output PDF.

After each step of adding complexity to the acoustic model set, the model set was reestimated using the training data and cross validated against the WSJCAM0 evaluation data using Viterbi decoding. After adding complexity, the performance would increase during subsequent reestimations before falling back slightly afterwards, indicating that reestimation was over fitting the model set to the training data. Over fitting undermines the model's ability to generalise to unseen data. Consequently, the reestimated model set that gave lowest WER on the evaluation set was carried through to the next step.

4.5.3 Parameter selection

4.5.3.1 Mixture components and decision tree thresholds

Optimum parameter values were determined empirically throughout the reestimation process. The key model parameter values required were the minimum number of states per node in the phonetic decision tree for state-tying, and the number of Gaussian mixture components for each state PDF. There is a theoretical trade-off between the two. If the maximum number of states per node in the phonetic decision tree is low, less tied-states (thus more distinct states) will result. Increasing the threshold for the minimum number of states per node in the phonetic decision tree leads to more tied-states (thus fewer distinct states). It is reasonable to assume that the effect of increasing the amount of state tying will mean that more Gaussian mixture components in each state output PDF are required to sustain an adequate level of parsimony across the model set. In addition, the training of fewer triphones also supports reliable estimation of a state output PDF with a higher number of mixture components.

To test this hypothesis, 4, 8 and 16 mixture component models were generated from the single component tied-state model set using phonetic decision tree minimum node thresholds of 175, 350, and 700. Thus, six model variants were compared.

Mixture components were added incrementally. A 2-mixture component model was generated from a single component model, a 4 mixture from the 2-mixture model after reestimation and so on. The model variants were reestimated using WSJCAM0 training data until over fitting was observed[11]. The reestimated model giving best performance against the WSJCAM0 evaluation set in each variant was selected for comparison with others. A 12,625-word uniform language model was used based on WSJCAM0 vocabulary. The results are summarised in Table 4.1.

Tree	Mixture components		
threshold	4	8	16
175	69.8%	69.6%	72.0%
350	68.1%	68.7%	70.9%
700	70.5%	68.1%	72.1%

Table 4.1: WER for different combinations of decision tree threshold and number of state PDF mixture components.

For a threshold of 175 states per node, 4 and 8-mixture component based model sets performed similarly. As state tying was increased using a decision tree minimum threshold of 350 states per node, the 4 and 8 mixture component model set both showed improvement, more so with 4 mixtures. Increasing state tying again using a minimum threshold of 700 states per node worsened the performance of the 4-mixture component model set but increased performance of the 8-mixture set. The 16-mixture component based model performed worse.

The results in Table 4.1 show the trade-off between the number of mixture components and degree of state tying. A single model was chosen for the baseline system from 2 candidates with the lowest WER (68.1%) - the 4 mixture component per state model set using a phonetic decision tree with the number of states per node minimum threshold of 350, and an 8 mixture component model with 700 minimum states per node threshold. It was not until performance was measured later in the process by incorporating a language model[12] that the 8 mixture model proved marginally better

4.5.3.2 Word insertion penalty and language model scale factor

During Viterbi decoding[13] there are two parameters used in decoding that affected performance - a word insertion penalty and language model scale factor. The word

[11] §4.5.2
[12] §4.6
[13] §2.4.3

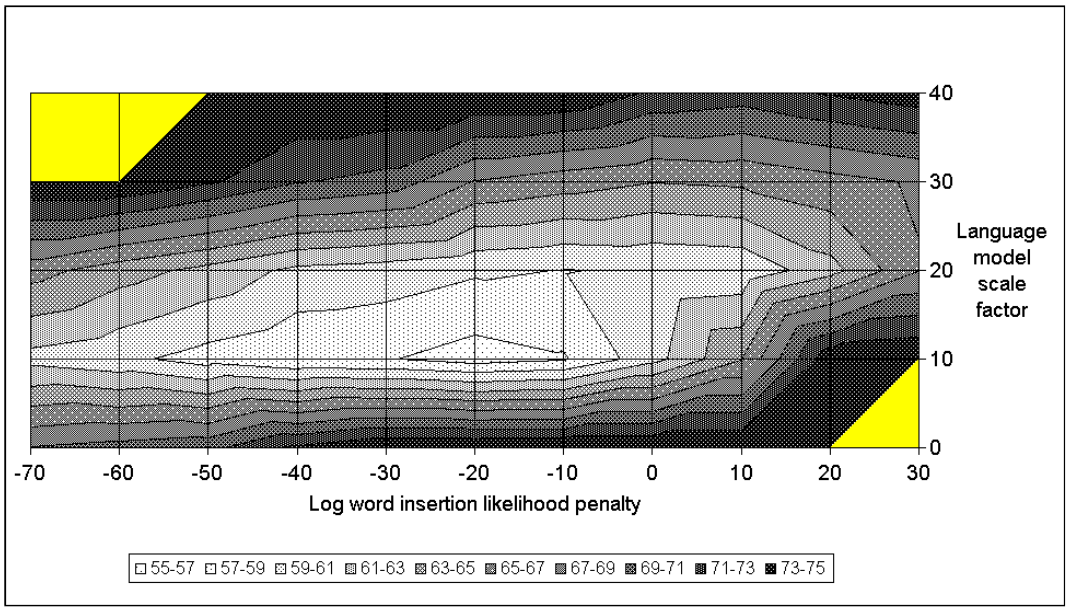


Figure 4.1: Surface plot showing WER as a function of word insertion penalty and language model scale factor. Optimum value corresponds to a log word insertion penalty of -10 and language model scale factor of 10. WER was measured at gridline intersections.

insertion penalty is an additive constant penalty incurred each time the Viterbi decoder hypothesises a transition between words. Its function is to balance the word insertion and deletion rates. In fact, the name 'word insertion penalty' is a misnomer as word insertions may be encouraged by adding positive likelihood. This would be desirable if the number of word deletions was in excess of the number of word insertions.

The language model scale factor also affects word transitions by multiplying all word transition probabilities by a constant factor. In essence, this balances the effect between acoustic model and the language model on ASR output, as a high value for the language model scale factor increases the language models affect.

The optimum value for the word insertion penalty and language model scale factor was determined empirically. It varied for each model set during training. For all model sets, it was reliably estimated since the performance of the recogniser in terms of WER has only one maximum as a function of word insertion penalty and language model scale factor. Figure 4.1 shows an example surface plot of WER as a function of word insertion penalty and language model scale factor. The computation is carried out in the log domain.

4.5.3.3 Beam pruning

Beam pruning is used during Viterbi decoding to reduce computational requirements by filtering out word sequences with forward ($\widehat{\alpha}$) probabilities that fall below a certain likelihood threshold. The HTK default values for beam pruning were deemed sufficient. Decoding without beam pruning did not yield significantly different performance in WER.

4.5.4 Acoustic model performance

Figure 4.2 shows the improvement in performance observed against the WSJCAM0 evaluation set during training. A 12,625-word uniform language model was used during development due to practical considerations. Consequently, the lowest WER observed was 70.5%[14,15]

4.6 Language model

A language model provides the a-priori probability of a word sequence, $P(W_i)$. It does this by providing the word boundary transition probabilities during Viterbi decoding. An N-gram language model approximates the joint probability of the sequence of J words $w_1, \ldots, w_J$ in the sequence W_i by conditioning each word on the previous $N-1$ words only:

$$P(W_i) = \prod_{j=1}^{J} P(w_j | w_{j^*}) \tag{4.11}$$

Where:

$$j^* = \begin{cases} j - N + 1 & \text{if } j - N + 1 \geq 1 \\ 1 & \text{otherwise} \end{cases} \tag{4.12}$$

Two language models were built: One for benchmarking the ASR against other systems using the WSJCAM0 test sets, the other for performing ASR with the

[14]The WER is produced using a sub-optimal choice of word insertion penalty. §4.6 improves on this performance by adding a non-uniform language model.

[15]Figure 4.2 shows the acoustic model set with 8 mixture components and a decision tree threshold of 700. Thus, the minimum WER of 70.5% (iteration 28) corresponds to the performance reported in Figure 4.1.

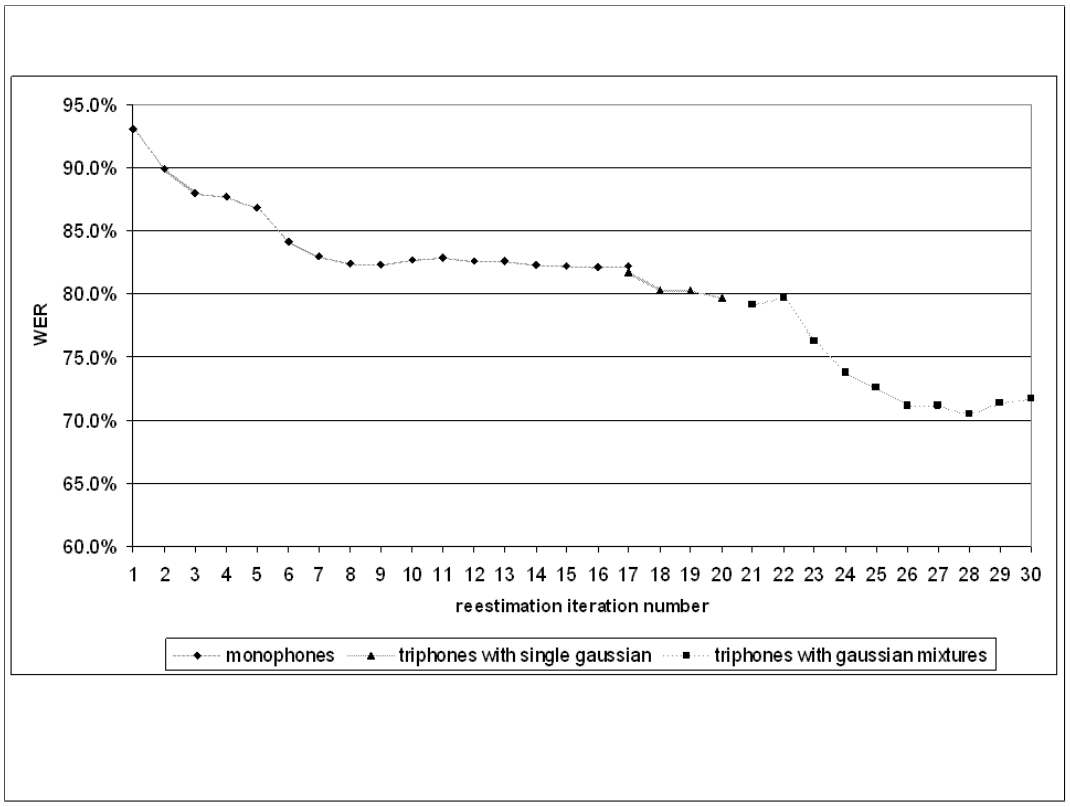


Figure 4.2: Performance improvement of acoustic model set during training. The x-axis represents each reestimation step. The Y access represents the WER. The legend at the bottom indicates the complexity of the model. Triphones were added at reestimation step 17 and Gaussian mixtures at step 22. Reestimation was shown to improve performance indicated by the downward trend in WER. When complexity was added, WER decreases on subsequent reestimations. Over fitting was observed when WER increases (i.e. iteration 22, 29, and 30).

eye/speech corpus. Both models were bigrams ($N = 2$), with unseen bigrams backed off to unigrams. Backed-off bigrams, $P_{\text{bo}}(w_i|w_j)$, were produced using frequentist estimates of probabilities for seen bigrams, $P_{sn}(w_i|w_j)$, and unigrams, $P(w_i)$, in the training data:

$$P_{\text{bo}}(w_i|w_j) = \begin{cases} P_{\text{sn}}(w_i|w_j) & \text{if bigram seen} \\ BOW(w_j)P(w_i) & \text{if bigram unseen} \end{cases} \tag{4.13}$$

For bigrams unseen during training, in each bigram word pair sequence, (w_j, w_i) a back-off weight [Kat87], $BOW(w_j)$, was used. The back-off weight ensured that the unseen bigrams estimated from unigrams maintained a row stochastic word transition

probability matrix:[16]

$$BOW(w_j) = \frac{1 - \sum^{\text{seen bigrams for } w_i} P_{sn}(w_i|w_j)}{1 - \sum^{\text{seen bigrams for } w_i} P(w_i)} \quad (4.14)$$

The WSJCAM0 test set backed-off language model was built from all training and evaluation data in the WSJCAM0 corpus, comprising of 11,220 utterances with 181,877 words. The eye/speech corpus language model, which is used for the integration experiments, is described later in this chapter (§4.9).

4.7 Benchmark performance

The baseline ASR system supports 25,321 triphones made up from all triphones derived from the British English Example Pronunciation (BEEP) dictionary plus those which occur in words observed in the eye/speech corpus speech data. The BEEP dictionary is a word to phone dictionary developed in conjunction with the WSJCAM0 corpus.

The WSJCAM0 test data consists of two sets of speech data from 48 speakers, with each speaker reading 40 sentences from a 5,000-word vocabulary. The test sets are referred to as si_dta and si_dtb.

HMM model variants with 2, 4, 8 and 16 Gaussian mixture components per state output PDF were tested. Table 4.2 shows the baseline recognition performance using a bigram backoff language model trained on WSJCAM0 training and evaluation data. The WER decreased as a function of the number of Gaussian mixture components per state output PDF from 2 to 8 mixtures, however increasing the number of components to 16 resulted in a higher WER. The 8-mixture component acoustic model set was selected for use with the eye/speech corpus. Table 4.3 gives the acoustic model performance (i.e. ASR using a uniform language model). Comparing the results in Table 4.3 and Table 4.2 shows that using a non-uniform language model approximately halves WER.

4.7.1 Comparison with other systems

The baseline system was compared against other systems that use the WSJCAM0 test corpus. The original WSJCAM0 baseline WER cited by Robinson of 11.4%

[16] $B = [P(w_i|w_j)]_{i,j=1,\ldots,V}$ for bigrams with vocabulary size V

[RFP+95] was achieved with the Abbot ASR system, not generally available to the research community. Van Bael [Bae02] [BK03] used HTK and achieved 31.4% WER using the BEEP pronunciation dictionary and a WSJ-derived bigram language model. Likewise Humphries [HW97] achieved a similar WER of 30.9%. Yan [YV02] achieved 12.8% WER using an ASR system with 20 Gaussian mixture components per state.

The baseline system's combined WER of 20.9% on WSJCAM test sets si_dta and si_dta compares favourably with other systems. It shares common properties with the Van Bael baseline system, documented in detail in his 2002 Masters thesis [Bae02].

A recent (2002) IEEE magazine article [Pad02] about speech recognition summarised the previous decade's research in terms of improvement to WER for decoding telephone voice-mail speech. Telephone voice mail speech is more of a challenge to decode due to lower bandwidth (4 kHz compared to 8 kHz), background noise and its spontaneous nature. A baseline WER of 40.5% was stated for a triphone-based HMM system with a trigram language model. Improvements in feature extraction, acoustic adaptation, and combining ASR output from multiple systems lead to an improvement in WER to 32.7%. Although this does not directly compare with WSJCAM0 test benchmarks, it demonstrates that current credible large vocabulary ASRs

Mixture	%WER		
Components	si_dta	si_dtb	si_dta + si_dtb
2	24.66	25.56	25.09
4	22.43	21.94	22.20
8	21.01	20.81	20.91
16	21.26	21.48	21.41

Table 4.2: Baseline recognition performance using WSJCAM0 5k word test data sets for ASR system comprising of a HMM triphone model set with 2,4, 8 and 16 gaussian mixture components per state output PDF, using a bigram backoff language model trained on WSJCAM0 training and evaluation data. Test results are shown for each set and combined for benchmarking purposes.

%WER		
si_dta	si_dtb	si_dta + si_dtb
47.99	49.41	48.67

Table 4.3: Baseline recognition performance using WSJCAM0 5k word test data sets for HMM triphone model set with 8 Gaussian mixtures components per state output PDF and a uniform language model. This gives an indication of the acoustic model ASR performance as the WSJCAM0 language model is not used in the baseline ASR.

typically perform between 10% and 30% WER, and that baseline system developed for this study is credible.

4.8 Acoustic adaptation to the eye/speech corpus

The WSJCAM0 corpus speech differs from the speech in the eye/speech corpus. They have different speakers and acoustic channel characteristics. The speech in WSJCAM0 is read rather than spontaneous. Running the ASR system on the eye/speech corpus in its native (trained and tested on WSJCAM0) form yielded poor performance - on average 96% WER. Adaptation to the eye/speech corpus was therefore necessary. This section details the improvements made by adapting the acoustic model set.

4.8.1 Adaptation techniques - MLLR and MAP

CMN accounted for the transfer characteristics (i.e. convolutive noise) in the acoustic channel during front-end processing. The acoustic HMMs parameter sets were also adjusted to account for the difference between the speech data in WSJCAM0 and that used in the eye/speech corpus [HH01].

To change HMM parameters, a Baum-Welch EM reestimation of the model set using the eye/speech data was not feasible as the volume of data available was insufficient to present multiple, or in most cases any, examples of each triphone in the model set. As an alternative, supervised acoustic model adaptation was applied to the HMMs using Maximum Likelihood Linear Regression (MLLR) and Maximum-A-Posteriori (MAP) adaptation. This section presents the performance improvements from applying these techniques leading to a selection of the best method.

4.8.1.1 MLLR

The MLLR technique [LP94] [GW96] computes a linear transformation of the means of the tied-state Gaussian observation PDFs in the HMMs, in order to maximise the likelihood of the adaptation data. This linear transformation matrix enables a corresponding linear transformation matrix for the variances to be computed. This technique will be henceforth referred to as 'global MLLR' as the same transformation matrix is used to update all of the tied-state Gaussian observation PDFs.

A refinement of MLLR is to compute mean and variance transformations for a subset of the HMMs that have similar valued parameters for their observation PDFs.

The HMMs are partitioned using a binary regression (decision) tree according to acoustic similarity. The root node of the regression tree represents all HMMs, and branch nodes smaller subsets of the HMMs as the tree is traversed downwards. Mean and variance transformation matrices are calculated at all nodes where there exists a sufficient number of states for adaptation. The threshold for a sufficient number of states is determined empirically. The use of a binary regression tree with MLLR is henceforth referred to as 'regression MLLR'.

MLLR benefits from requiring only a small amount of adaptation data to make a robust transformation: Instead of reestimating the large number of acoustic model parameters directly, a linear transformation matrix is estimated, which can be applied to every acoustic model, in the case of global MLLR, or a large group of acoustic models, in the case of regression MLLR. In applying the same transformation to many models, it performs best at adapting models to recording apparatus and environment, as opposed to speaker variation.

4.8.1.2 MAP

MAP adaptation updates the HMMs tied-state observation PDF means using Bayesian learning, to maximise the probability of the model given the adaptation data. Unlike EM, the degree of updating is governed by the number of examples in the adaptation data for each mean. MAP therefore performs best when a large amount of adaptation data exists. In this regard, MAP primarily accounts for speaker variation between test and training data, as opposed to the variation in recording apparatus and environment.

4.8.2 Adaptation procedure

ASR was performed on the Instruction Giver part of the eye/speech corpus using:

1. No acoustic adaptation.
2. MLLR with a global mean and diagonal covariance transformation as described in §4.8.1.1 ('Global MLLR').
3. MLLR with mean and diagonal covariance transformations as described in §4.8.1.1 ('Regression MLLR').
4. MAP adaptation as described in §4.8.1.2.
5. Regression MLLR followed by MAP.

4.8.3 Results

Table 4.4 and Figure 4.3 show the results for adaptation. Global MLLR gave an average 21.0% improvement in WER over no adaptation. Regression MLLR performed better showing an average 34.9% improvement over no adaptation, indicating that there was enough adaptation data to form a binary regression tree with multiple nodes that differed in their transformation matrices. MAP adaptation beat MLLR showing an average 40.6% improvement, indicating that there was sufficient adaptation data for modifying the state output PDFs to account for speaker variation. Combining Regression MLLR and MAP gave the overall best improvement in performance of 51.8% over no adaptation, resulting in an average WER across all sessions of 46.3%. Consequently, Regression MLLR followed by MAP was chosen as the preferred method for acoustic adaptation of WSJCAM0 acoustic models to the eye/speech corpus.

Note that the adaptation data used in these experiments is speech data from all sessions in the eye/speech corpus. Speech data is the same data used for subsequent eye movement ASR integration experiments in Chapter 6, which means the results presented in Table 4.4 are optimistic and do not necessarily reflect performance on unseen data. In subsequent experiments on a particular test session, a HMM triphone set using Regression MLLR and MAP was produced using all other session's speech recordings as adaptation data. Thus, a separate HMM triphone set was used for each eye/speech corpus session.

System performance is heavily dependant on the speech data - i.e. the eye/speech corpus' spontaneous and conversational speech data was harder to recognise than the WSJCAM0's read speech data used to benchmark the system. In this context, the average WER of 46.3% for the baseline performance given in Table 4.4 is acceptable against the WSJCAM0 benchmark of 20.9% WER given in Table 4.2.

4.9 Language model adaptation to the eye/speech corpus

4.9.1 Corpora used

In an N-gram language model [17], the word sequence probabilities will vary according to the task domain. To construct a baseline language model for this study, a task-

[17] §4.6

Session	WER for adaptation method (%)				
	no adaptation	Global MLLR	Regression MLLR	MAP	Regression MLLR + MAP
m1g1f2	84.1	65.9	59.8	62.8	52.2
m1g1f3	71.0	60.2	48.3	48.0	34.1
m1g2f1	99.8	73.0	53.8	39.5	32.9
m1g2f3	105.3	77.7	58.4	45.7	38.7
m1g3f1	77.9	59.0	50.5	51.2	44.3
m1g3f2	91.0	71.7	63.6	69.1	55.0
m2g1f2	112.1	88.8	72.7	76.6	50.2
m2g1f3	119.1	95.7	75.5	70.3	52.6
m2g2f1	93.8	76.8	58.8	47.8	40.3
m2g2f3	97.1	78.4	62.8	54.0	42.2
m2g3f1	97.2	73.0	58.3	49.8	38.8
m2g3f2	95.1	74.4	60.9	55.5	42.6
m3g1f2	105.3	81.8	71.0	65.9	55.1
m3g1f3	97.7	71.3	58.3	53.0	48.3
m3g2f1	96.9	73.9	62.5	63.3	53.7
m3g2f3	102.6	79.8	69.1	65.0	54.2
m3g3f1	89.6	83.8	70.1	54.4	57.8
m3g3f2	91.6	79.9	70.0	54.7	40.3
Mean	96.0	75.8	62.5	57.0	46.3
%Improvement		21.0%	34.9%	40.6%	51.8%
StDev	11.4	9.0	7.7	9.9	7.8

Table 4.4: ASR performance on eye/speech corpus speech data session for adaptation schemes. Results show that using Regression MLLR and MAP gives the best performance overall.

domain specific language model relating to participants describing landmarks and routes on the eye/speech corpus maps was required. The language model vocabulary was constrained to be the vocabulary of the eye/speech data (1,108 words) in order to minimise WER.

The HCRC Map Task was a potential source of data to construct such a language model, as the task domain was similar to the eye/speech corpus task. The HCRC Map Task transcriptions consisted of 18,013 utterances with 149,859 words in total and 2,160 unique words[18]. 394 of the words in the eye/speech corpus did not occur in the HCRC map task corpus transcriptions.

[18] §3.2

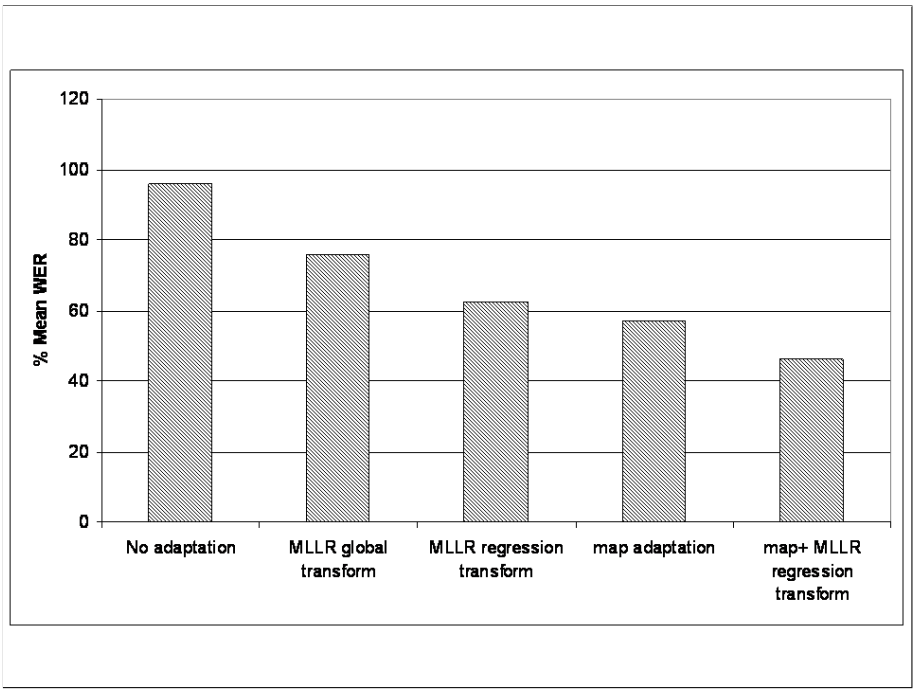


Figure 4.3: Bar graph showing effect of acoustic adaptation schemes on WER for all sessions in the eye/speech corpus.

The BNC corpus[19] provided a large amount of spoken English. From the spoken language part, 6,105,876 utterances with 340,429 unique words were extracted. This vocabulary included all of the words from the HCRC Map Task corpus and the eye/speech corpus.

The BNC corpus provided a large amount of data for language model construction of non-task specific general spoken English. The HCRC Map Task contained task-domain specific spoken English, but a small vocabulary that did not cover all of the eye/speech corpus vocabulary. Therefore, the construction of the eye/speech corpus task-specific language model used both corpora.

4.9.2 Optimisation

Combining the BNC and HCRC Map Task corpus into a single set of data in order to build the task-specific language model would be futile, as the volume of BNC corpus data is some 300 times greater than that of the HCRC map task. In such a scenario, the HCRC Map Task's effect on language model probabilities would be minimal. Therefore, to optimise ASR performance, and maintain the influence of

[19] §4.3.2

both corpora, separate bigram language models were created from the HCRC Map Task and BNC corpora. Each language model covered the 1,108 vocabulary of the speech in the eye/speech corpus. ASR performance was measured in terms of WER and FOM. Language model performance was measured using PER[20].

The bigram language model implementation was backoff with Turing-Good discounting [Goo53]. Bigrams for each word pair, $P_{\text{MapTask}}(w_i|w_j)$ and $P_{\text{BNC}}(w_i|w_j)$, were merged using linear interpolation:

$$P_{\text{merged}}(w_i|w_j) = \nu P_{\text{MapTask}}(w_i|w_j) + (1-\nu)P_{\text{BNC}}(w_i|w_j) \quad (4.15)$$

The resulting bigrams, $P_{\text{merged}}(w_i|w_j)$, formed a single backoff bigram language model which covered the eye/speech corpus vocabulary. The interpolation weight, ν, was defined as the % of Map Task language model used. The optimum interpolation weight was determined empirically by finding the weight corresponding to the lowest WER for speech data in the eye/speech corpus. The PER, and FOM for keywords[21] were also measured for performance evaluation. Table 4.5 shows the results.

Interpolation weight (%)	PER	WER (%)	FOM (%)
0	165.0	46.3	60.6
5	133.1	44.1	60.3
10	121.4	42.9	58.5
20	108.5	42.0	58.0
30	101.4	41.5	56.3
40	97.2	41.2	55.9
50	95.0	40.9	55.9
60	94.4	40.7	55.4
70	95.5	40.7	55.1
80	99.1	40.9	54.2
90	107.4	41.1	53.6
95	116.5	41.5	52.7
100	161.5	42.3	51.6

Table 4.5: Interpolation of HCRC Map Task and BNC language model and its effect on PER, WER and FOM against eye/speech corpus session m1g1f3 speech data for various interpolation weights.

[20] §4.2 defined performance measures.
[21] §3.4.7 lists the keywords identified in the eye/speech corpus.

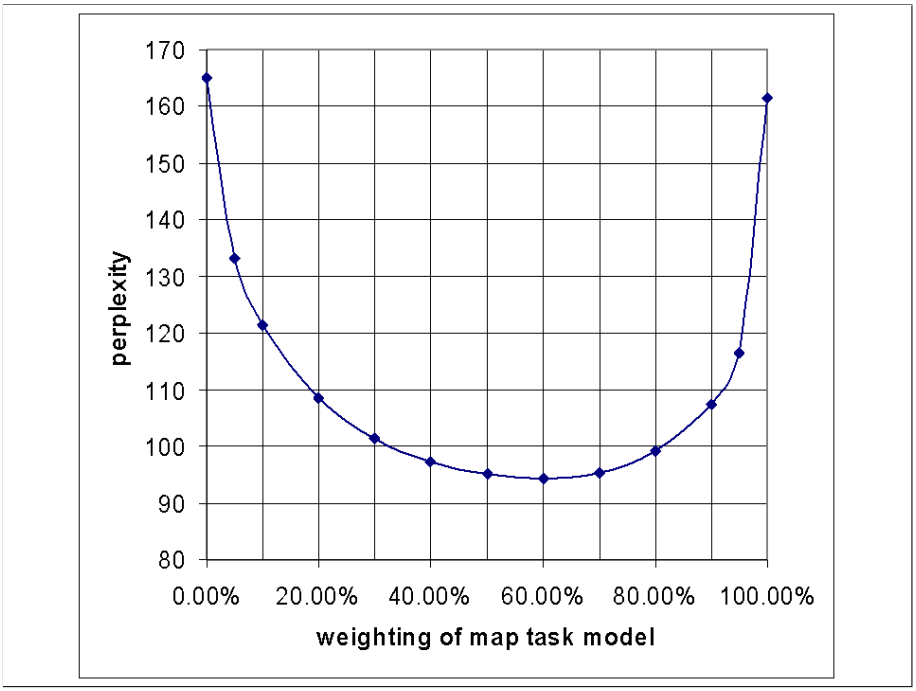


Figure 4.4: PER of the interpolated HCRC Map Task and BNC derived language model when using the eye/speech corpus session m1g1f3 speech data for various interpolation weights.

Figure 4.4 shows the perplexities of the language model against various interpolation weights. An all-BNC or all-Map Task language model (at 0% weighting and 100% weighting respectively) show similar perplexities of 165 and 161 (to 0 dp) respectively. Combining the models results in a decrease in perplexity. Perplexity declines exponentially as the weights of BNC and HCRC Map Task model become more similar. Lower perplexities were obtained where the Map Task has a slightly higher weighting than the BNC. A weighting of 60% Map Task gives the lowest perplexity.

As discussed previously[22], perplexity is not necessarily an indication of ASR performance. However a clear improvement in PER was observed by interpolating the language models. This indicated that the language used in the eye/speech corpus task was being modelled better by mixing the BNC and Map Task models in roughly equal proportions, compared to using either alone. When using only the BNC language model, perplexity was high due to an under-representation of task-domain grammar in the bigram probabilities. Conversely, when using only the Map Task language model, perplexity was high because words not present in the Map Task vo-

[22] §4.2.2

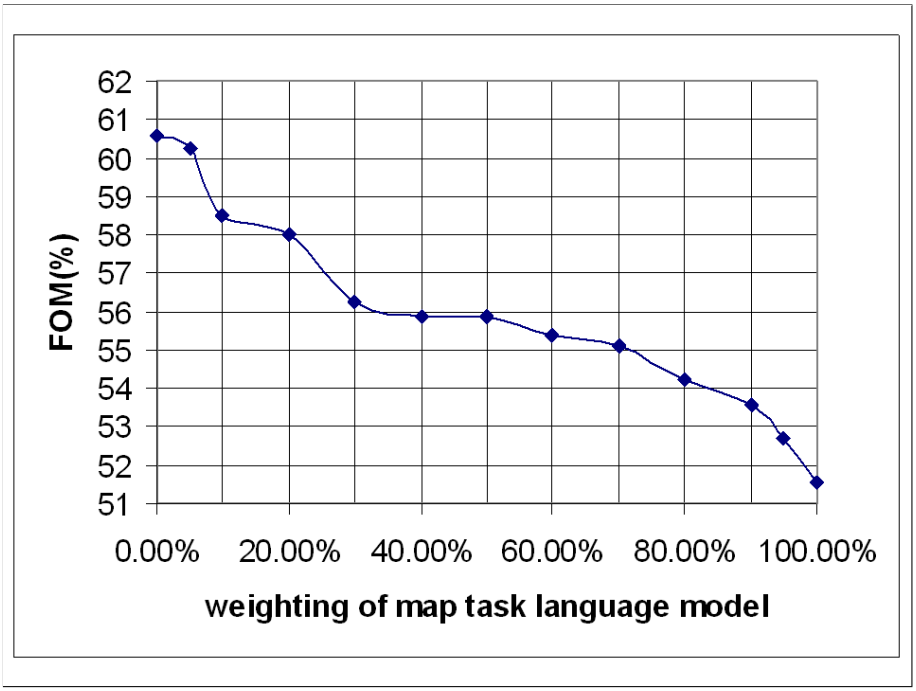


Figure 4.5: FOM of the interpolated Map task and BNC derived language model when using the eye/speech corpus speech data for various interpolation weights.

cabulary, but present in the eye/speech corpus, were under-represented in the bigram probabilities[23].

Figure 4.5 and Figure 4.6 show ASR performance on the eye/speech corpus speech data, against various interpolation weights. The FOM falls as the interpolation weight increases as many keywords (map landmarks) were unseen in the Map Task. The change in WER against interpolation weight shows a similar form to PER. However, the WER for ASR using only the Map Task language model showed better performance than using only the BNC language model, unlike using PER where both were similar. This indicated that task-domain grammar representation in the language model was more important for resolving the difference between acoustically similar words than accounting for unseen words.

The language model chosen for the baseline ASR uses interpolation levels of 60% HCRC Map Task and 40% BNC ($\nu = 0.6$). The HTK language model tools were used to generate the language model set, with choice of a backoff bigram language model using Turing-Good discounting being due to the limitation of HTK supporting only bigram and unigram language models.

[23]Uniform word probabilities were assumed for unseen words in the Map Task language model

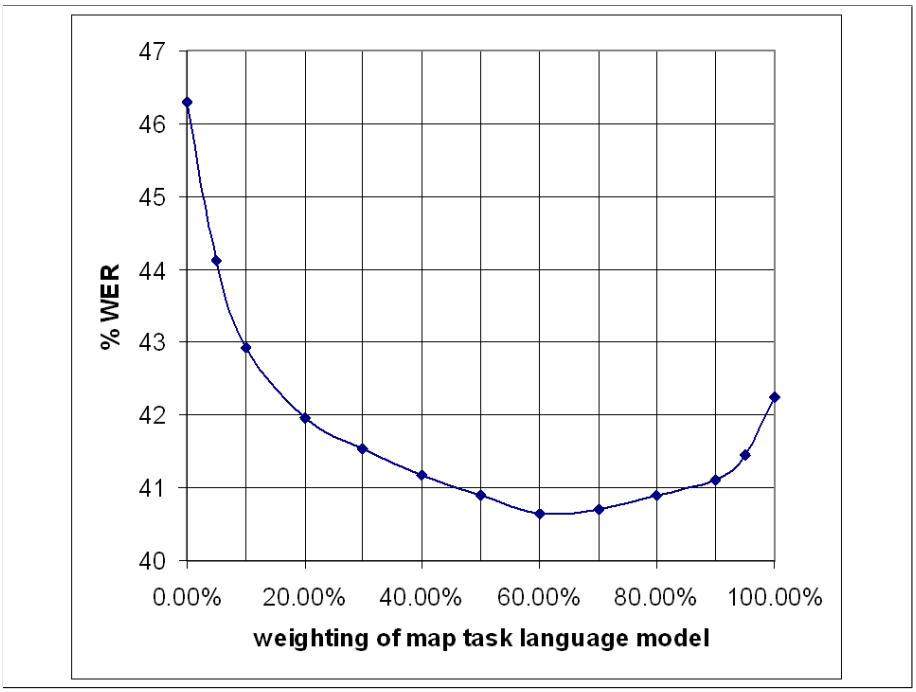


Figure 4.6: WER of the interpolated Map task and BNC derived language model when using the eye/speech corpus speech data for various interpolation weights.

4.10 Baseline results

Baseline results for the eye/speech corpus were obtained using the ASR described in this chapter, Table 4.6 presents these results. The results indicate a mean WER of 50.5% with a standard deviation of 6.64%. These results are used in Chapter 6 as a basis for integration experiments.

Session	WER(%)	Session	WER(%)	Session	WER(%)
m1g1f2	49.7	m2g1f2	64.7	m3g1f2	54.5
m1g1f3	46.3	m2g1f3	60.9	m3g1f3	50.9
m1g2f1	42.9	m2g2f1	42.3	m3g2f1	54.5
m1g2f3	45.4	m2g2f3	43.3	m3g2f3	56.5
m1g3f1	45.2	m2g3f1	45.9	m3g3f1	53.3
m1g3f2	52.5	m2g3f3	43.0	m3g3f2	56.6

Table 4.6: Baseline ASR performance against the eye/speech corpus.

4.10.1 Time-aligned transcriptions

The eye/speech corpus was time-aligned transcribed by humans[24]. Regenerating transcriptions using baseline ASR output eliminated the human error in time aligning words. For each speech segment, a language model corresponding to the transcribed word sequence for that segment was generated[25]. The ASR, using the segment-specific language model, produced a time-aligned transcription for each segment.

The results of this automatic alignment were beneficial. Human transcribers did not encode all pauses between words, tending to specify instead a mid-point during the pause between words as the word boundary. Thus, pauses were omitted altogether. A benefit of automatically generating time-aligned transcriptions was that all silences between words were identified. Consequently, word boundary timings were more accurate. The improved keyword spotting (FOM) accuracy rates during integration experiments described in Chapter 6 reflected this.

4.11 Summary

A typical HMM large vocabulary continuous ASR system was implemented using HTK. The system was trained using the WSJCAM0, BNC, and HCRC Map Task corpora. Audio was parameterised using static and dynamic MFCCs with CMN. The ASR used phonetic decision tree clustered tied-state triphones with 3-state left-to-right topology HMMs with 8 component Gaussian mixtures for state output PDF. The HMMs incorporated skip transitions to handle deviations in speaking rate and pronunciation from the training data, by allowing bypassing of states in the left to right HMM topology. The system was benchmarked against standard WSJCAM0 5k test sets and shows credible performance (20.8% WER) against the published results of other ASR systems. The acoustic model set was adapted to the eye/speech corpus using MLLR and MAP adaptation and a mean baseline of 50.5% WER obtained for eye/speech corpus sessions, using a 1108 word language model.

Baseline performance against the eye/speech corpus compared to WSJCAM0 performance is relatively low, although credible. There are two likely reasons for this: deviations from training corpora and transcription errors made by inexperienced transcribers.

[24] §3.4.5

[25] As HTK does not provide this feature, this was realised by software written in C#

Whereas the corpora used to train the baseline ASR was matched as much as possible to the eye/speech corpus, there were some deviations which affected the performance of the acoustic and the language models. The WSJCAM0 speech, whilst from British native speakers, was read speech and not spontaneous as in the eye/speech corpus. Spontaneous speech is prone to partial pronunciations, disfluencies and speech rate variation that were not prevalent in WSJCAM0. The BNC corpus was mostly speech from radio and television broadcasts and not conversations between teenagers as in the eye/speech corpus. The BEEP pronunciation dictionary, whilst listing common British pronunciations that correspond to words, does not necessarily indicate the actual pronunciation used in the eye/speech corpus, or in common English either, since it was developed semi-automatically from WSJCAM0 using ASR.

Inexperienced transcribers who varied in their handling of disfluencies carried out transcripts of the eye/speech corpus. It is not unreasonable to assume that the transcription errors, while mostly detected in the 2nd pass of the data, were probably not all eliminated. It is also likely that WSJCAM0 and the HCRC Map Task corpora contain transcription errors. The baseline ASR system enabled accurate time-aligned transcriptions of eye/speech corpus sessions to be generated, improving transcription accuracy with respect to word boundaries. Transcription errors affected the adaptation of the acoustic models to the eye/speech corpus and performance measuring.

Whereas this chapter has presented the building of an ASR system in an orderly fashion, there were many iterations of the recogniser. Initially monophone acoustic models with a unigram language model were envisaged, but it was found that model complexity had to be increased in order to recognise spontaneous speech with any success.

ASRs using MFCC acoustic feature vectors with triphone HMMs and an N-Gram language model have dominated speech recognition research for the last decade. The interim has seen much research in engineering refinements but the foundations remain intact. Chapter 5 studies information extraction from eye movement and Chapter 6 integrates the information from eye movements into this baseline ASR system.

5 Eye Movement Analysis

This chapter concerns extracting information from natural eye movement data for integration into an ASR system. Three forms of information from eye movement data were uncovered:

- The eye movement type, such as fixations or saccades.
- The identification of areas of interest in the visual scene or more precisely, the focus of visual attention (FOVA)[1].
- The classification of the sequence of foci of visual attention to identify a user goal or behaviour.

In previous studies, only the standard HMM was used to classify eye movement data. The main body of this chapter describes the novel application of the hidden semi-Markov model (HSMM) to eye movement data and asks what advantages do the Markov chain in the HMM and the semi-Markov chain in the HSMM yield in terms of facilitating noise-robust decodes of eye movement type and FOVA sequence. The chapter finishes with some preliminary studies addressing the use of HMM models to recognise user goals or behaviours.

5.1 Previous HMM eye tracking studies

Salvucci successfully applied HMMs to eye movements for eye-movement-type classifiers and for uncovering the sequence of foci of visual attention [Sal99] [Sal00]. His motivation was primarily to automate the analysis of eye movement data collected in cognitive psychology research, as analysis is a time consuming, human-intensive and error-prone activity. His work applying HMM to eye movements is the most comprehensive published to date and relevant details of it are given later in this chapter in §5.6.1 and §5.7.1.

[1] A definition and discussion of the 'focus of visual attention' was given in §2.2.3.

The author's Masters Thesis prototyped the software used for eye movement analysis in this chapter and applied HMMs to eye movement to uncover the focus of visual attention [Coo02]. A conference paper outlining initial work in using HSMM to uncover focus of visual attention (§5.7) was published in the Eye Tracking Research and Application symposium (ETRA) in 2004 [CRM04], and can be found in appendix B.

5.1.1 Types of eye movement

There are five main eye movement types - fixations (gazing at a single point); pursuit (tracking a moving object); vestibular (rotation to compensate for head rotation); vergence (disconjugate rotation of each eye to focus on near objects) and saccades (rapid movement between points) [Ray98]. Visual processing is attenuated during saccadic eye movement [Mat74] and this distinction makes these classifications and their relative latencies useful units of analysis for gaining insight into cognitive processes. Three further types of eye movement are not generally considered a separate classification, as they co-occur with a main type. These are nystagmus (tremor in eye); drifts (slow movements due to oculomotor imperfections); and micro-saccades (rapid eye movements to correct drifts).

5.2 Eye movement data sets

There is little in the way of published corpora of eye movement data compared to that of speech. In these experiments, two unpublished datasets were used: The eye/speech corpus (§5.2.1) and the Psycholinguistic study data (§5.2.2). A third dataset, the Smart Eye data (§5.2.3) was used to characterise noisy eye tracking data (§5.5).

All data sets used in this study record fixations and saccades as the main movement types. Pursuit movements were absent as the visual scenes were static. Vergence is not observed as the visual scene, presented to the user on a VDU, is at a fixed distance from the subject. Vestibular movements were minimal as head movements were infrequent and the subject was seated facing the same direction throughout.

5.2.1 Eye/speech corpus

The eye/speech corpus contains 7 sessions of high quality samples of eye movement. The eye movement data indicates the direction of gaze relative to a visual scene

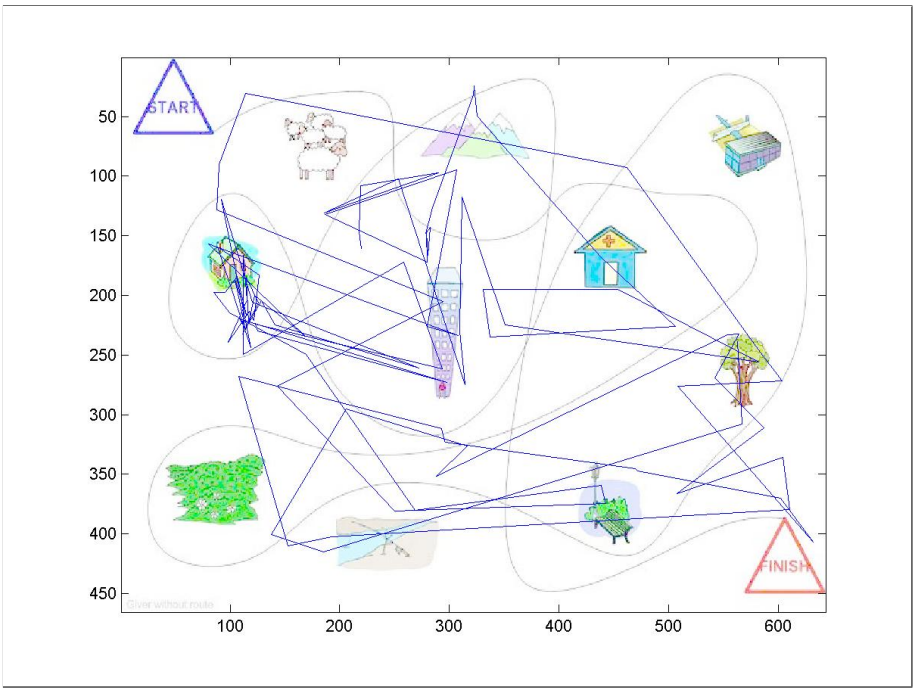


Figure 5.1: A scene from the eye/speech corpus with participants eye movement (scan paths) superimposed over the scene.

displayed on a computer screen (Figure 5.1). This data was recorded by an Eyelink [Gmb99] eye tracker at 250 Hz sample frequency. In addition to eye position, the data also contains event information for onset and duration of fixations and saccades estimated by the eye tracker in real-time during capture. Chapter 3 described the corpus in detail.

5.2.2 Psycholinguistic study dataset

Eye tracking and speech data were collected as part of a psycholinguistic study at the School of Psychology in the University of Birmingham during 2002. The eye data was recorded using the same Eyelink eye tracker as that used to collect eye data in the eye/speech corpus. Participants viewed sets of 4 objects (Figure 5.2) and named them in short utterances such as 'bell, web, window, door'. The session durations were typically much shorter than those in the eye/speech corpus (i.e. no longer than 1 minute). 50 trials were used including 80 different word tokens (i.e. objects).

This data was originally used to examine cognitive processing in language production. The data forms part of a greater research effort stemming from 'The theory of lexical access in speech production', proposed by psycholinguistic researchers Levelt

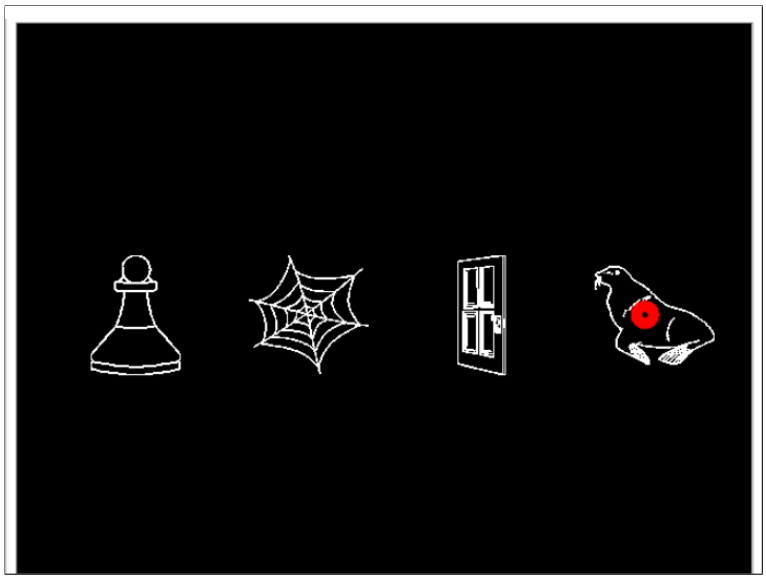


Figure 5.2: A typical screen presented to users in the psycholinguistic study data. The 4 objects are named by the subject in sequence from left to right.

et al in their seminal paper [LRM99]. Details of the theory were given in §2.2.4. To date, published research using the Psycholinguistic study dataset includes analysing the effect of word length on fixations during naming objects [MRL03], which finds that fixation time increases as word length does.

In this study, the Psycholinguistic study data serves the purpose of evaluating HMM-based eye movement classifiers. The eye tracking data has sequential (the order of objects being named) and durational (object recognized and named) task-dependent characteristics. The data was captured in more constrained conditions than the eye/speech corpus - task duration was shorter and the visual field was less complex. The speech recordings were not used, consisting of single words corresponding to the objects viewed.

5.2.3 Smart Eye dataset

The Smart Eye dataset is used to demonstrate the validity of adding Gaussian noise to the Psycholinguistic study data to simulate noisy eye tracking data[2] in the FOVA classification experiments[3].

The Eye/speech corpus and Psycholinguistic dataset used a head-mounted eye tracker, developed during the 1990s. Less intrusive eye tracking technologies have

[2] §5.5
[3] §5.7

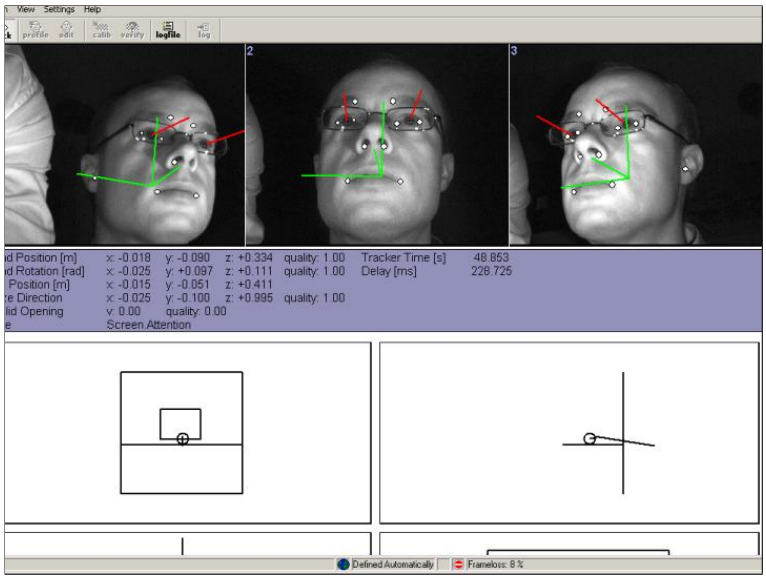


Figure 5.3: The Smart Eye Pro eye tracking system software estimating gaze direction from 3 IR camera images.

since emerged which rely solely on multiple remote cameras capturing images of a person's head [SYW97]. 'Smart Eye Pro' [AB04] is one such system which was developed for research to monitor car driver behaviour and improve car driving controls. Smart Eye Pro processes the video images from these cameras in real-time, estimating the head position, gaze direction, and facial features. Figure 5.3 shows the system in action. 'Smart Eye Pro' was demonstrated at Birmingham University in February 2004 and sample data was collected. This data consists of eye tracking data for a user reading the digits 1 to 4 positioned horizontally on a VDU in both forward and reverse sequence (i.e. left to right or right to left), akin to the objects in the Psycholinguistic study data. The data was recorded at 25 samples per second.

5.3 The Hidden semi-Markov model

With a standard HMM (introduced in §2.4.3), the probability distribution of state occupancy duration decays geometrically from the time of entering the state (i.e. from time t). This is not the observed behaviour of eye movements, where state durations correspond to fixation and saccade durations. The HSMM [RM85] [BW86] [Lev86] (Figure 5.4) provides a method to incorporate a more flexible state duration PDF in a HMM framework.

The durational distribution, $p_i(d)$ describes the statistical distribution of how long

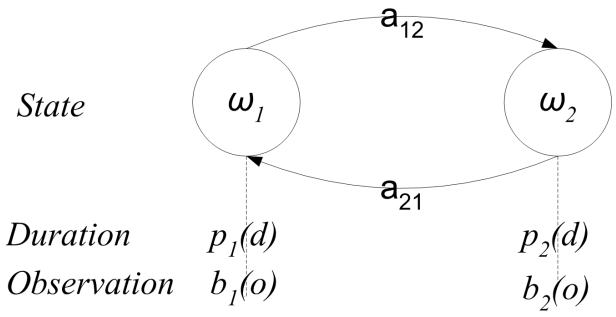


Figure 5.4: A 2 state HSMM.

attention is on an object before the user looks at something else. In common with the Observation PDF, a parametric distribution such the Gaussian is frequently used:

$$p_i(d) \quad = \quad p(d|\omega_i) \tag{5.1}$$

Where d is the duration and ω_i is state i. Consequently, it is convenient to define,

$$a_{ij} = 0 \qquad \text{for } i = j \tag{5.2}$$

Where a_{ij} is the state transition probability from state i to state j. Viterbi decoding in HSMMs incorporates the explicit state duration model:

$$\hat{\alpha}_t(i) \quad = \quad \max_j \max_d (\hat{\alpha_{t-d}}(j) p_i(d) a_{ji} b_i(o_{t-d+1}, \ldots, o_t)) \tag{5.3}$$

Where $\hat{\alpha}_t(i)$ is the maximum calculated value of the forward probability for state i at time t and $\hat{\alpha_{t-d}}(j)$ is the maximum calculated value of the forward probability for state j at time $t-d$. At each node (i,t) of the state-time trellis a record is kept of d_{max}, the duration, and j_{max}, the state, j at $t-d$ that gives the maximum calculated value of $\hat{\alpha}_t(i)$. As for the HMM, the optimum hidden state sequence is thus recovered by tracing back from the final node in the trellis, using j_{max} and d_{max} to identify the previous state.

Viterbi Reestimation is covered in Appendix C.

5.4 Analysis software

HMM decoders were implemented in C# running on the Microsoft .NET software platform on a 2.4 GHz Pentium 4 PC. The HMM library developed for this is detailed

in Appendix A. Its initial conception and software design was outlined in the authors Masters thesis [Coo02]. Algorithms implemented in MATLAB [Mat02] and C# were used for analysis. HTK was not used because it lacks an HSMM implementation.

5.5 Adding noise to eye-tracker data

The experiments in this chapter add noise to eye movement data in order to measure the robustness of classifiers. Noise was added as a deviation from the original eye position in the horizontal and vertical plane (5.4) using single Gaussian random variables, $G(0, \sigma)$, for each plane with standard deviation, σ equal to the original eye position and the noise level respectively. Horizontal and vertical plane noise was varied independently. A diagonal covariance matrix was assumed.

$$(x', y') = (x + G(0, \sigma), y + G(0, \sigma)) \tag{5.4}$$

To ascertain whether adding Gaussian noise to deviate eye positions simulates real eye tracker noise, the Smart eye dataset (§5.2.3), was compared to the Psycholinguistic data (§5.2.2). As described previously, in each data set the subjects faced a screen and looked at 4 objects, in turn, from left to right.

Figure 5.5 and Figure 5.6 show the output of the psycholinguistic and Smart Eye data respectively. The Psycholinguistic dataset shows eye position in the horizontal plane progressing in four steps corresponding to fixating on 4 objects between 3976ms and 5096ms. The Smart eye dataset has similar output between 1600ms and 4100ms. In both, the vertical plane remains relatively constant as expected. The Smart Eye data is visibly noisier and shows greater variation. The frequency histogram plot of this simple comparison in Figure 5.5 shows that eye tracker noise broadly follows a normal distribution about the actual eye position. Therefore adding random Gaussian noise to simulate noisy eye tracker output from laboratory studies is a valid approach. This approach does not account for the calibration errors observed in the eye/speech corpus eye movement data[4].

[4] §3.4.3

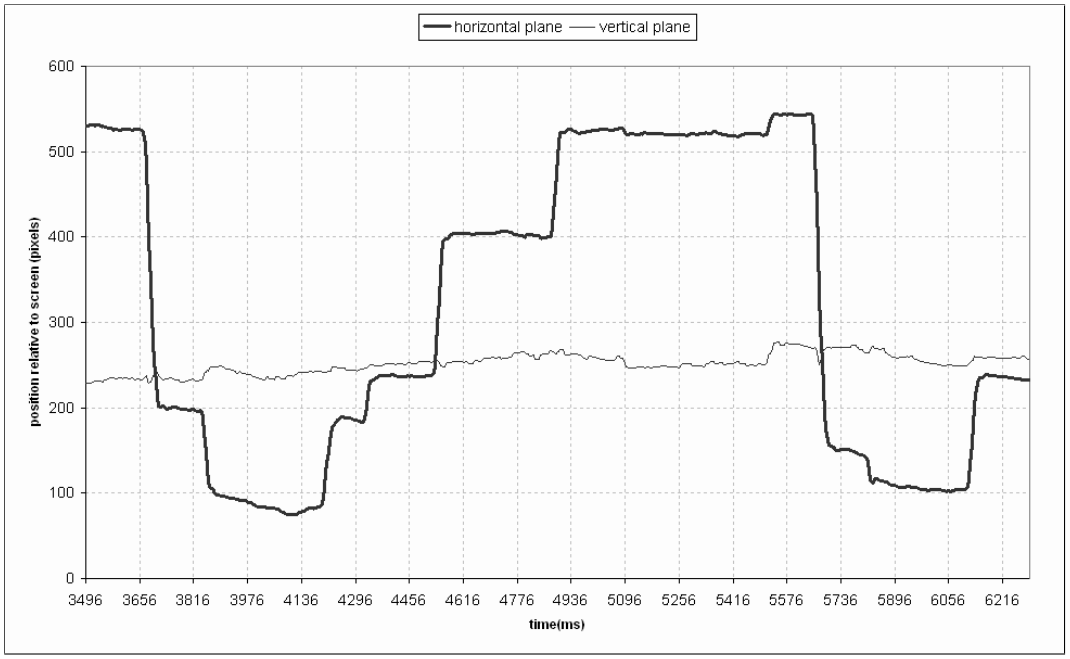


Figure 5.5: Sample output from the Psycholinguistic dataset.

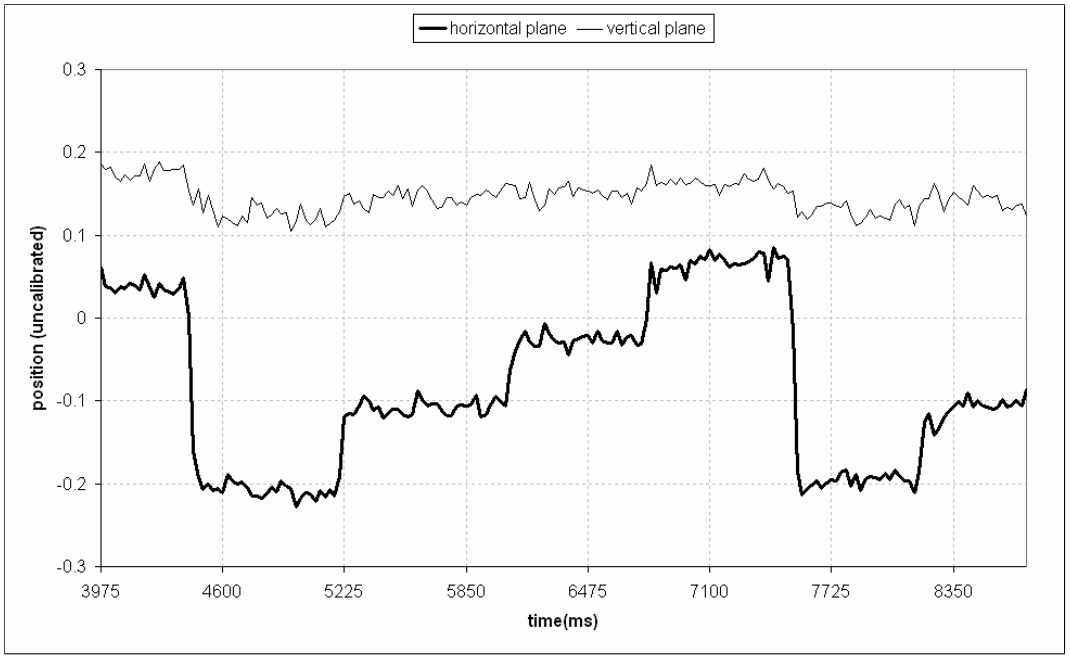


Figure 5.6: Comparable output from Smart Eye data set.

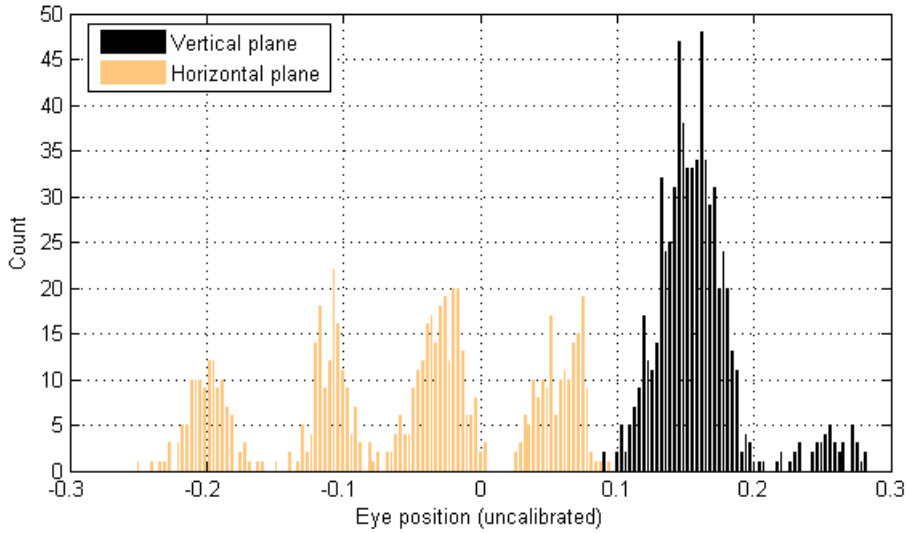


Figure 5.7: Frequency histogram showing the eye position variation in the Smart eye dataset. Peaks in frequency indicate fixation positions. The frequency distributions indicate that a Gaussian distribution would approximate variation from fixation positions.

5.6 Eye-movement-type classification experiments

This section asks what advantages do the Markov chain in the HMM and the semi-Markov chain in the HSMM yield for eye-movement-type classification. Whereas in previous research HMMs were employed as eye-movement-type classifiers, there has been no attention given to HSMMs and their measurable benefits. There is a growing requirement to classify eye tracker data captured in environments where sensor noise is present, such as outside the laboratory using wearable or remote eye tracking equipment. In these situations, the benefits of using HMM or HSMM may become apparent. To explore these benefits, GMM, HMM and HSMM eye-movement-type classifiers were built. Their performance was compared as noise was added to the eye tracker data.

5.6.1 Previous studies

In 1984 Kowler proposed a 2-state Markov model to predict a person's smooth (pursuit) eye movements based on the motion of a moving target [KAM84]. Eye movement classification is useful unit of analysis in cognitive science where fixation durations

help to infer cognitive processes and this inspired Salvucci to propose a 2-state eye movement-type HMM which classified saccades and fixations [Sal99] [Sal00]. The HMM was compared with classifying eye movements based on three techniques involving thresholding. In summary, these techniques were:

- Measuring the velocity of the eye and classifying a fixation based on a threshold for maximum velocity.
- Measuring the variance of eye positions within a moving time window and classifying small variances as fixations.
- Associating eye position during movement to the known visual artefacts in the field of view, and classifying a sequence of eye positions as a fixation if the eye positions in the sequence correspond to the same artefact and exceed a minimum (pre-determined) duration for a fixation.

The main findings from the comparison of classifier models was that two models were preferred - The HMM and measuring eye position variance over a time window and applying a maximum threshold to identify a fixation.

Outside of cognitive studies, HMMs were subsequently used for fixation detection as part of user interface studies [YBZ02] and more recently, the models have been expanded to account for vestibular and smooth pursuits eye movements [CAR04] using an observation vector comprising of head movement in addition to eye movement data. In all previous studies, and in common with HMM applications in most fields, the Gaussian distribution was used to represent state output PDFs and model parameters were estimated using supervised training data.

5.6.2 Distinction based on time derivatives

Distinctions in eye movement types are based on the velocity and acceleration of the eye movement. An eye tracker provides information regarding direction of gaze in the form of a 2-dimensional feature vector representing the horizontal and vertical eye position of the eye relative to the visual field[5].

Let the observation symbol set, O, define the set of possible observations for gaze direction. O is the product of a horizontal and vertical symbol set (X and Y respectively) on a visual field whose size is constrained by $r * c$ discrete points, where r,c are number of points in a row and column respectively:

[5]Depth is not accounted for in these experiments although was present in the Smart Eye dataset.

$$O = X \times Y = \{0, \ldots, r\} \times \{0, \ldots, c\} \tag{5.5}$$

The sequence of eye positions, o, represents the eye tracking data - horizontal (x) and vertical (y) position of the eye with respect to the visual field over time:

$$o = (o_1, \ldots, o_t, \ldots, o_T) \tag{5.6}$$
$$o_t = (x_t, y_t) \tag{5.7}$$

The method for classifying eye movement type is to derive a sequence of 1st-order time derivatives from the sequence of eye positions and classify each derivative (i.e. eye velocity observation) to an eye movement type:

$$o' = (o'_1, \ldots, o'_t, \ldots, o'_T) \tag{5.8}$$
$$o'_t = o_t - o_{t-1} \tag{5.9}$$
$$o'_t = (x'_t, y'_t) \tag{5.10}$$

The classification of eye movement type is based on the velocity or speed (i.e. the magnitude of the velocity, $\|o'_t\|$, of the eye and the eye movement type's expected duration. To demonstrate this empirically, fixation durations from all sessions deemed to have an adequate level of quality in the eye/speech corpus were determined from the horizontal and vertical eye velocities, determined from the change in eye position over time. The frequency distribution of eye speeds (Figure 5.8) shows a mode greater than 0 pixels/s indicating, as expected, that even during a fixation the eye is never static due to nystagmus and drifts. A large proportion of the eye movements are low speed, indicating fixations. Saccades make up the remainder of the distribution that decreases in a decaying fashion as a function of increasing speed.

Given the distribution of eye speeds in Figure 5.8, fitting a truncated 2-component Gaussian distribution with diagonal covariance yields estimations for the probability distribution function $b(\|o'_t\|)$ of the two eye movement types. The distribution is truncated so that negative values for speed have zero probability. A single component in the mixture density represents each type[6]:

[6]The maximum likelihood estimation for the Gaussian mixture was implemented in C# based on MacKay's soft K-means algorithm version 3 [Mac03].

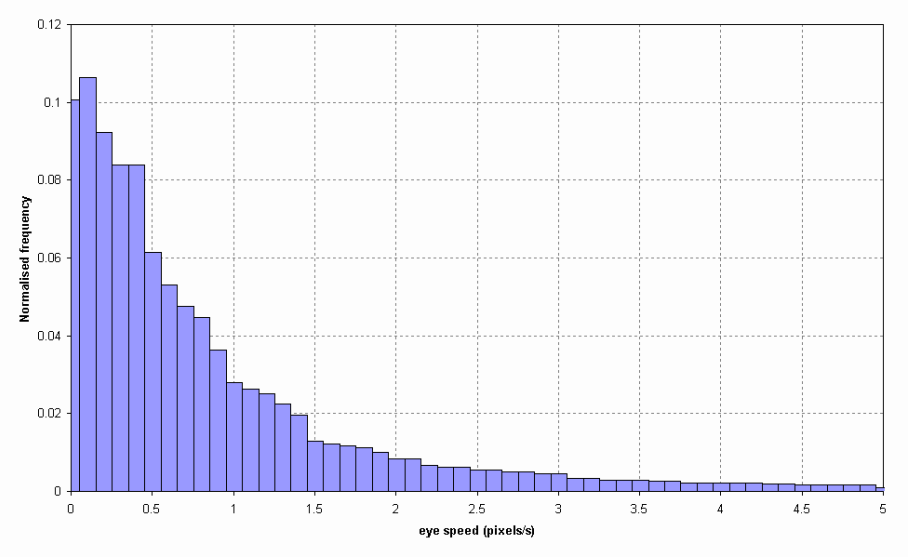


Figure 5.8: Normalised frequency histogram showing the eye speed in eye/speech corpus session m1g3f1.

Component	m1g1f3	m1g3f1	m1g3f2	m2g1f2	m2g2f3	m3g2f1	m3g2f3
μ_1	150	114	153	158	116	115	92
σ_1	97	76	113	104	81	80	71
w_1	.664	.576	.689	.561	.532	.746	.639
μ_2	768	1312	2009	1036	978	1670	1521
σ_2	1441	2727	6156	2846	2969	3524	3259
w_2	.336	.424	.311	.439	.468	.254	.361

Table 5.1: Estimation of eye speed probability distribution using 2 Gaussian mixtures. Subscript 1 represents component associated with eye speeds classed as a fixation. Subscript 2 represents component associated with eye speeds classed as a saccade.

$$b(\|o'_t\|) = \begin{cases} \sum_{k=1}^{2} w_k b_k(\|o'_t\|) & \text{if } \|o'_t\| \geq 0 \\ 0 & \text{if } \|o'_t\| < 0 \end{cases} \tag{5.11}$$

Table 5.1 shows 2-component truncated Gaussian mixture density estimations for eye speed in the eye/speech corpus sessions. The mixture weights reflect the proportion of eye movement attributed to saccades or fixations, with fixations carrying greater weight (w_1). Figure 5.9 shows a plot of the Gaussian mixture component density estimate for session m1g3f2.

Further evidence that there were two distinct types of eye movement (based on eye speed) present in the data was obtained by repeating mixture density estimation

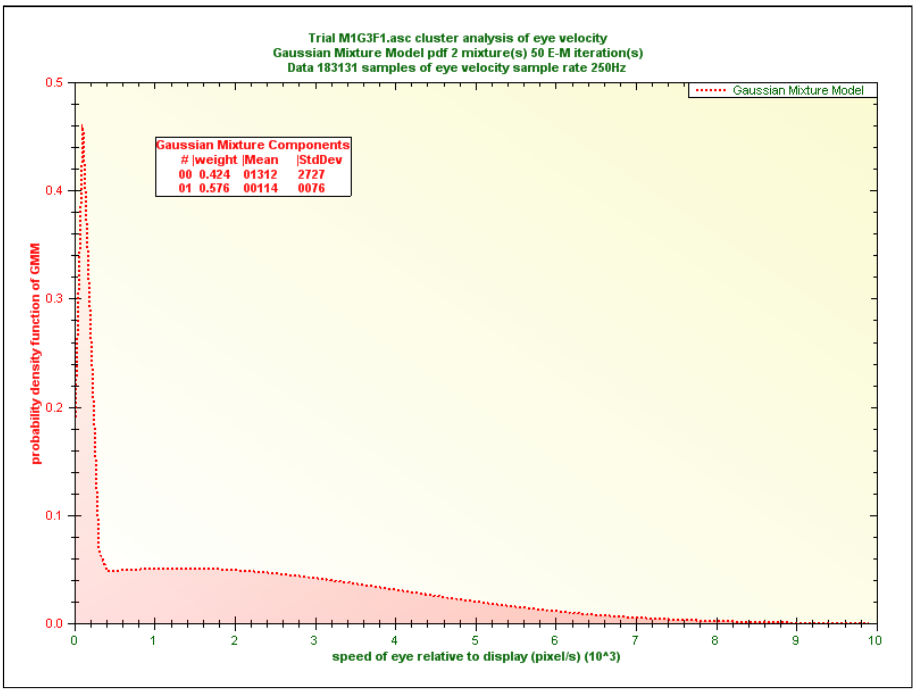


Figure 5.9: GMM estimation of eye speed distribution for eye/speech corpus session.

using a larger number of mixture components. Table 5.2 compares the weightings between 2 and 5 component mixture density estimation. Two components dominated the weightings. It is therefore reasonable to assume that the additional component's purpose of improving the model works by refining the existing structure rather than identifying other eye movement types.

The fitting of a 2-component Gaussian to eye data speeds is used to initialise the HMM model observation PDF in §5.6.4.

5.6.3 Distinction based on duration

The mean durations for fixations and saccades vary according to task. Typically, mean durations for fixations range from 225ms to 400ms. Saccades are ballistic and their durations are a function of the degree of eye movement. Saccades typically range from 30-50ms for $2-5^{\circ}$ movement of the eye. The average velocity of the eye during a saccade is typically $500^{\circ}s^{-1}$ [Ray98].

Figure 5.10 shows the frequency histogram of fixation durations for 7 sessions of the eye/speech corpus, which correspond to 11654 fixation events. A Gaussian distribution with a positive mean is a potential approximation for the fixation duration distribution.

Component	m1g3f2 (2 components)	m1g3f2 (5 components)
μ_1	153	193
σ_1	2009	117
w_1	.689	.568
μ_2	2009	2176
σ_2	6156	6432
w_2	.311	.472
μ_3	-	25
σ_3	-	0
w_3	-	.065
μ_4	-	56
σ_4	-	0
w_4	-	.043
μ_5	-	35
σ_5	-	0
w_5	-	.041

Table 5.2: Estimation of eye speed probability distribution using 2- and 5-component GMMs. In both models, 2 components dominate weightings.

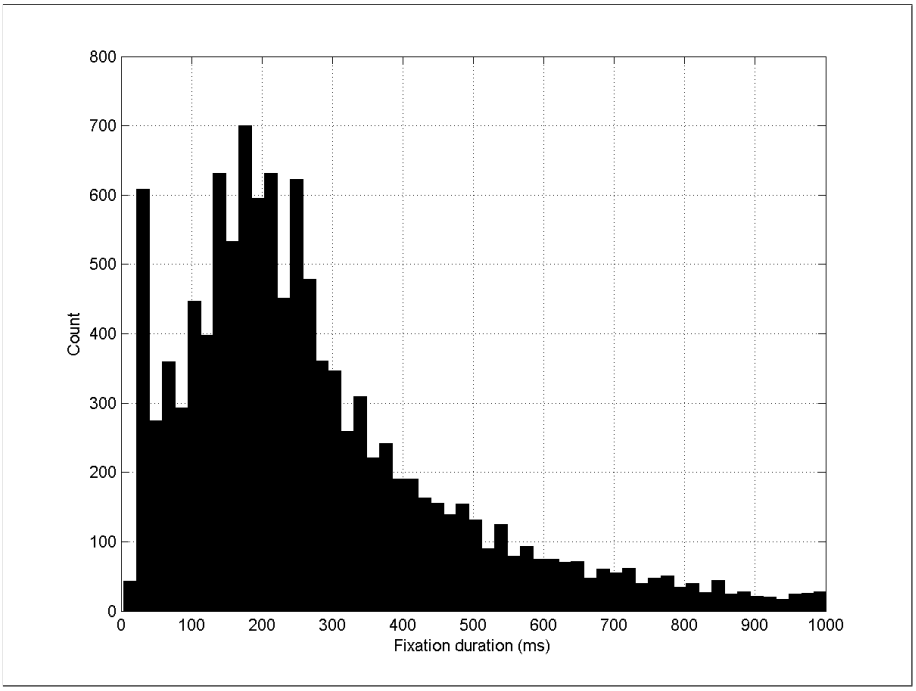


Figure 5.10: Distribution of eye fixation durations observed in the eye/speech corpus.

5.6.4 HMM and HSMM structure

A 2-state ergodic HMM eye-movement-type classifier was built to uncover fixations and saccades. The eye speed, $\|o'_t\|$, derived from successive eye positions (5.10), was the observation and each of the HMM state output PDFs was implemented as a single truncated Gaussian. Truncation ensures zero-probabilities for negative values. The HMM observation PDF parameters for fixation and saccade states were initialised by fitting a 2-component GMM to the observation set as described in §5.6.2, with each component representing a single state output PDF.

The 2-state HSMM eye-movement-type classifier similarly used a single truncated Gaussian state output PDF, initialised in the same way. Truncated Gaussian distributions were also used for the HSMM state duration PDFs. The distribution of eye fixations (Figure 5.10) was used to estimate an initial value for the mean duration (i.e. 200ms) of a fixation. An initial mean value for a saccade was set at 40ms, based on previous findings in the literature [Ray98].

Both models were trained using Viterbi Reestimation. Supervised and unsupervised training approaches were evaluated. The eye-tracker's in-built fixation detection records were used as data for supervised training. The in-built fixation detection algorithm for Eyelink is based on thresholds for moving average velocity [Gmb99]. For unsupervised training, the Viterbi decoded state sequence from the model was used to reestimate the model until the model likelihood reached a local maxima.

Supervised training resulted in unstable models, which made incorrect classifications and a high number of state transitions in brief succession. The most likely explanation for this is differences in transition times between states in the eye tracker's in-built fixation detection algorithm muddying the decision threshold. Supervised training was consequently abandoned and the experiment used unsupervised training.

5.6.5 Baseline GMM

A GMM-based model was implemented to provide an example of a simple detection method, mimicking the in-built eye tracker eye-movement-type detection algorithm. The GMM was used to provide a baseline for eye-movement-type classification performance. The GMM-based model used a 2-component GMM estimated from the data[7]. The GMM was the same as that used for initialising the HMM and HSMM

[7] 5.6.2

HMM	Fixation	Saccade
Observation PDF mean (pixels/sample)	0.718	10.20
Observation PDF standard deviation (pixels)	0.577	26.70
Self state transition probability ($a_{i,i}$)	0.766	0.935

Table 5.3: HMM parameters for observation state PDF and self state transition probability.

HSMM	Fixation	Saccade
Observation PDF mean (pixels/sample)	0.796	10.21
Observation PDF standard deviation (pixels)	0.758	27.23
Duration PDF mean (ms)	127.86	36.26
Duration PDF standard deviation (ms)	37.68	8.58

Table 5.4: HSMM parameters for observation state PDF and durational PDF.

observation PDF parameters for fixation and saccade states. Each component in the GMM represented an eye movement type. For a given observation of eye movement speed, the eye-movement-type classification was that represented by the component that provided the greatest contribution to the observation's GMM probability. In the decoded sequence of eye movement classifications for a given eye movement sequence, fixation durations of less than 100ms were reclassified as saccades, in common with the in-built eye tracker fixation detection algorithm.

The eye/speech corpus sessions M1G3F2 and M3G2F1 were used for training and testing the models respectively. The session's eye movement data contained 220,802 and 163,326 eye position samples respectively. These were chosen as representative data. Similar results were obtained using other eye/speech corpus sessions.

5.6.6 Model parameters

Table 5.3 and Table 5.4 list the trained HMM and HSMM model parameters respectively for eye-movement-type classification. The observation PDFs for HMM and HSMM were approximately equal. The state duration for HMM was calculated from the self-state transition probabilities. The state duration PDFs of the HMM and HSMM are shown in Figure 5.11 and Figure 5.12 respectively.

Referring to Figure 5.11, the HMM saccade and fixation state duration PDFs geometrically decay from 0ms, which asymptotes towards zero at around 80ms and 280ms for saccades and fixations respectively. Likewise for the HSMM (Figure 5.12), the state duration PDFs for saccades and fixations peak at 36ms and 128ms, which

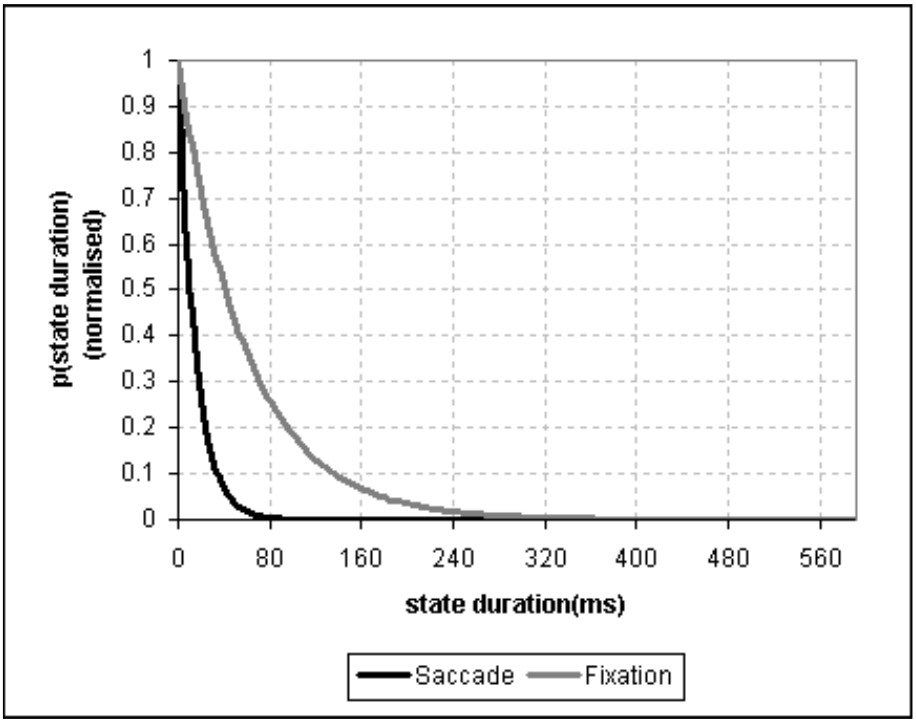


Figure 5.11: Geometric state duration probability distribution functions for HMM based eye-movement-type classifier.

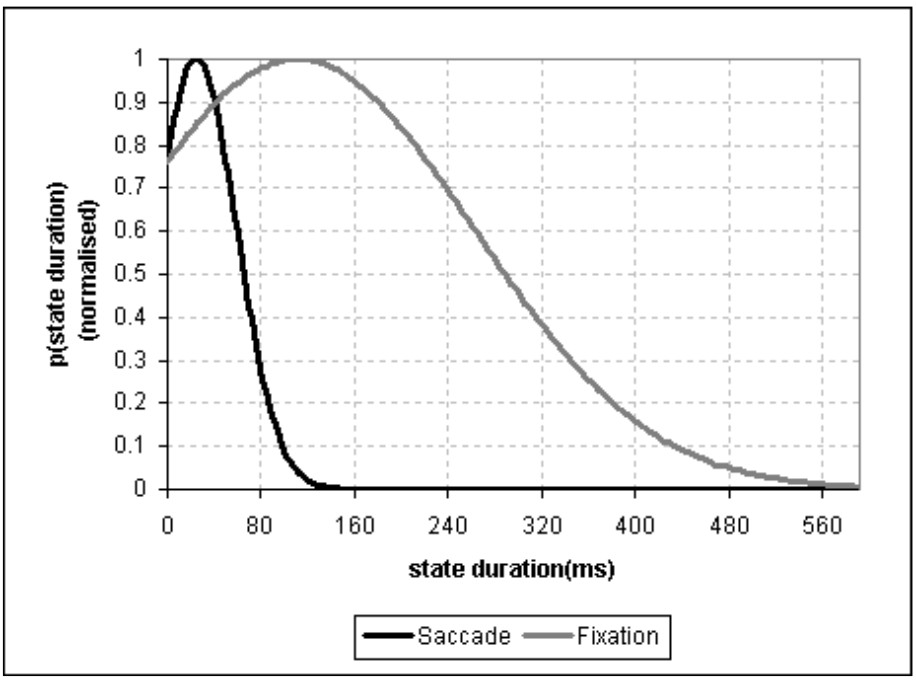


Figure 5.12: Truncated Gaussian state duration probability distribution functions for HSMM based eye-movement-type classifier.

asymptotes towards zero at approximately 160ms and 560ms respectively. From these observations the HMM state duration PDFs favour shorter saccades and fixations.

Comparing the HSMM fixation duration PDF in Figure 5.12 with the actual distribution in Figure 5.10 suggests that the HSMM fixation duration PDF better approximates the fixation duration distribution for eye movement. In addition to comparing against the eye/speech corpus, both the HSMM saccadic and fixation duration means are in line with published measures for average durations of these eye movement types.

5.6.7 Tests performed on eye/speech corpus data

The HMM, HSMM and GMM models were used to classify the eye movement type from eye/speech corpus eye data. There were two experiments:

Experiment 1: Classification Behaviour. Comparison of the eye-movement-type classifications between HMM, HSMM and GMM and the eye trackers' in-built classification method.

Experiment 2: Noise robustness. Noise was added to the eye tracking data[8]. The HMM, HSMM and GMM's relative noise robustness was evaluated using the measures described in the next section.

5.6.8 Evaluation measures

The following measures were defined for evaluating the performance of the GMM, HMM and HSMM in classifying eye movement type into either fixations or saccades for noisy eye movement data in experiment 2:

Accuracy: Proportion of eye positions with added noise classified correctly by a model compared to the same model classifying the same eye positions with no added noise.

Instability: Number of state transitions (i.e. classifications) made by a model compared to the same model classifying the same data with no added noise. State transitions represent onsets of fixations or saccades. A value less than 1 indicates fewer transitions occurring, greater than 1 indicates more transitions, hence greater instability.

The Accuracy measure is of primary importance for comparing models, with Instability providing a further insight into the models' classification behaviour. Both

[8] §5.5

measures are ratios that compare the model when classifying noisy eye movement data to the same model when classifying the same eye movement data with no added noise. The similar performance of all classification models in no noise (see §5.6.9 below) made this a reasonable approach since the manual labelling of eye movement data to eye movement type was not practical given the volume of data.

5.6.9 Results

In *Experiment 1: Classification behaviour* the models performed similarly to the eye tracker's own fixation-detection algorithm. Figure 5.13 and Figure 5.14 show a sample of the eye movement and the corresponding classifications respectively. Eight fixation onsets are shown over a 1800ms duration. As expected, the majority of classifications of eye position are fixations. The HMM and GMM appeared slightly more prone to classifying short saccadic movements (e.g. at 1200ms in Figure 5.13), suggesting a greater sensitivity of these models to detecting microsaccades. This is likely due to the HMMs geometric state duration PDFs and the GMM's uniform duration PDF. However, despite these small differences, the overall impression gained from these results is that the HMM and HSMM classifiers had comparable performance and behaviour to the GMM and the eye tracker's inbuilt algorithm. The benefit of using a Markov or semi-Markov chain was therefore not evident.

Experiment 2: Noise robustness showed that the HSMM exhibited a more graceful decline in accuracy as noise was added, compared to the other models. Figure 5.15 shows this change in accuracy. Adding Gaussian noise with a deviation of 1 pixel to the eye position data resulted in a decline in accuracy for all models, with the HSMM fairing best (95.6% accuracy), the HMM slightly worse (91.3%) and the GMM poorest (83.6%). Adding more noise resulted in a decline in accuracy for all models, which asymptotes to approximately 30%. This corresponds to chance accuracy for the 2-class problem - the classification of eye movement as saccades is more likely when the noise corrupts the eye movement data and a higher proportion of the eye movement consists of fixations rather than saccades. For all noise levels, the HSMM accuracy was highest. Measuring accuracy alone, the value of using HMM and HSMM is demonstrable as noise is increased.

Figure 5.16 shows the change in instability of the models to added noise. Adding Gaussian noise with a deviation of 1 pixel to the eye position data resulted in 73% more state transitions for the GMM, compared with using the same model for classifying the eye data with no added noise. Similarly, the HSMM classified with 13%

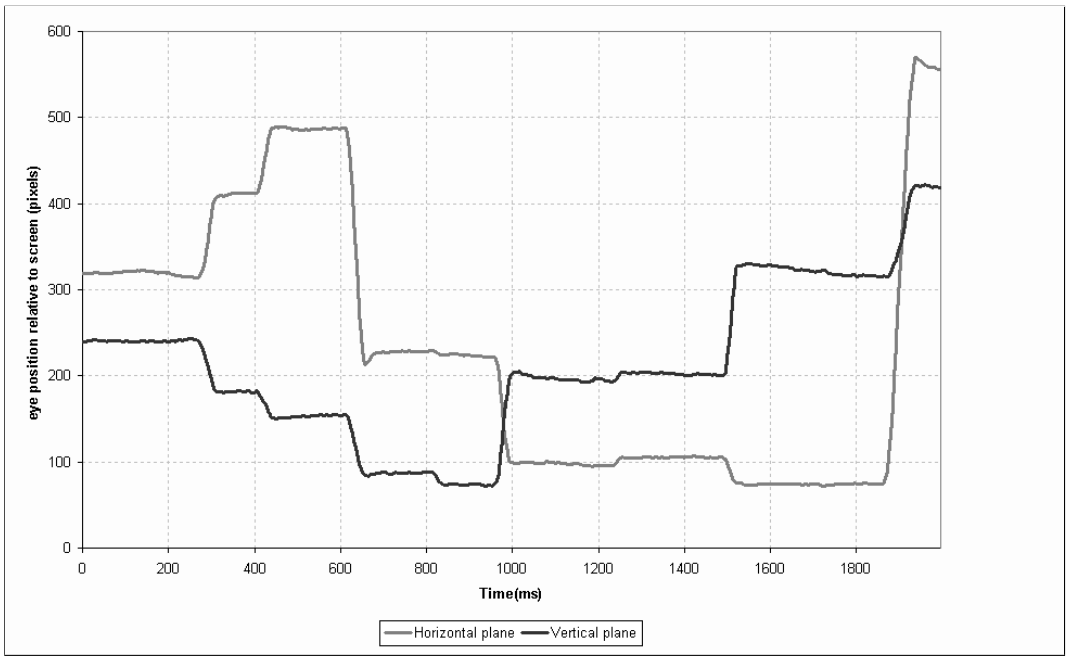


Figure 5.13: A typical example of the eye movement data used in the eye-movement-type classification experiments.

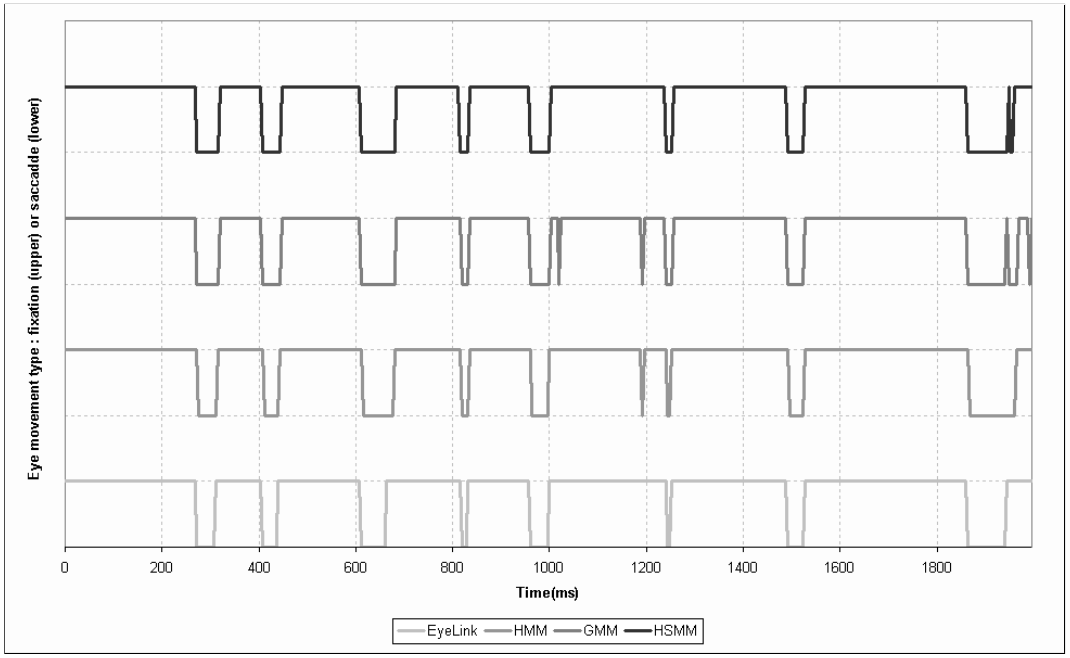
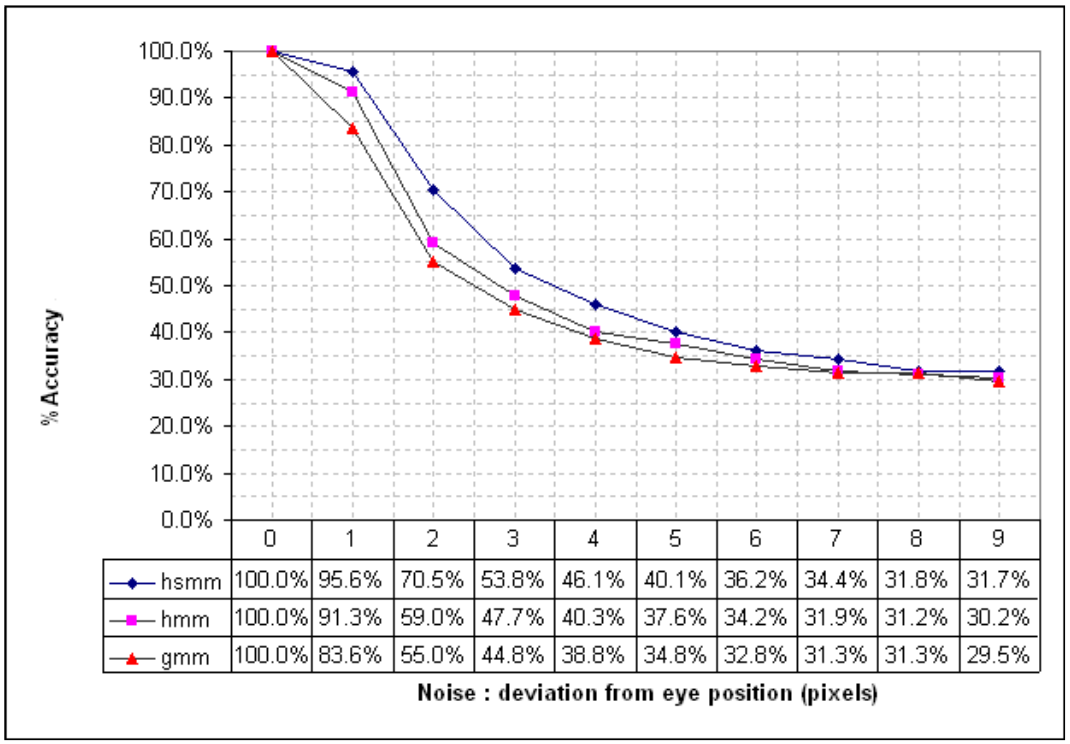


Figure 5.14: Eye-movement-type classification of eye data shown in Figure 5.13.

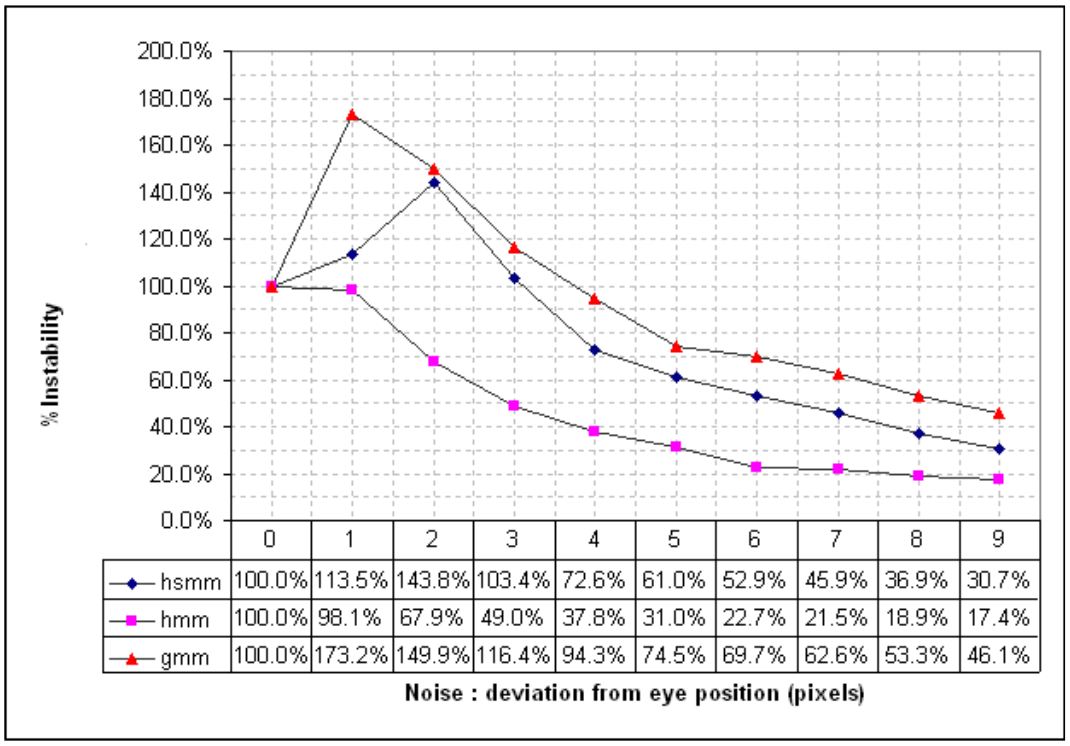


	0	1	2	3	4	5	6	7	8	9
hsmm	100.0%	95.6%	70.5%	53.8%	46.1%	40.1%	36.2%	34.4%	31.8%	31.7%
hmm	100.0%	91.3%	59.0%	47.7%	40.3%	37.6%	34.2%	31.9%	31.2%	30.2%
gmm	100.0%	83.6%	55.0%	44.8%	38.8%	34.8%	32.8%	31.3%	31.3%	29.5%

Figure 5.15: Accuracy of GMM, HMM and HSMM eye type classification models as a function of added noise.

more state transitions. The HMM classified with a 2% drop in transitions, thus at a noise level of 1 pixel deviation the HMM was the most stable. Doubling the noise resulted in the HSMM classifying with 44% more transitions compared to classifying eye data with no added noise. The other models showed a reduction in instability. The instability for all models asymptotes towards a low level as noise was increased, as most of the eye movement data was classified as saccadic. At a noise level of 9 pixels deviation, the HMM instability is at 17.4%, the HSMM is at 30.7% and the GMM is at 46.1%.

The relative levels of instability can be attributed to the difference in state duration modelling. The GMM had the highest instability because there is no duration modelled for either eye movement type (i.e. no Markov chain) - more transitions occur, as the probability of a self-state transition is equal to that of exiting the state. The HMM and HSMM condition each state on previous states (i.e. the 1st order Markov and Semi-Markov chain) and typically the self-state transition probabilities are greater than the probabilities for exiting the state. This results in the HMM and HSMM making less state transitions. The instability measure shows benefits of using



	0	1	2	3	4	5	6	7	8	9
hsmm	100.0%	113.5%	143.8%	103.4%	72.6%	61.0%	52.9%	45.9%	36.9%	30.7%
hmm	100.0%	98.1%	67.9%	49.0%	37.8%	31.0%	22.7%	21.5%	18.9%	17.4%
gmm	100.0%	173.2%	149.9%	116.4%	94.3%	74.5%	69.7%	62.6%	53.3%	46.1%

Figure 5.16: Instability of GMM, HMM and HSMM eye type classification models to added noise.

the HSMM and HMM in noisy data.

Combining the results for accuracy and instability gives an overall idea of the relative performance of the models. As noise was added, the accuracy tailed off for all models. The HSMM faired best, although it made more state transitions that the HMM. The HMM's accuracy fares worse than the HSMM's (although better than the GMM's) offsetting any apparent advantage in stability.

5.6.10 Discussion

These results show a benefit of using the Markov and semi-Markov chain for eye-movement-type classification but only if the eye tracker data is subject to noise - if the eye tracker data is noise-free then the GMM-based model has been shown to suffice. Due to the absence of a ground truth in labelling the data for eye movement type, unsupervised reestimation should be used to optimise the models.

This experiment's novelty has been to use the semi-Markov chain. Future experiments in this area could concentrate on comparing the HMM and HSMM models to models other than the GMM to further demonstrate their effectiveness.

Interest in eye movement classification is limited to research applications in cognitive psychology, where noisy eye tracking data is not common as experiments are conducted in controlled environments with precision eye trackers. In HCI applications the robust detection of fixations is less-controlled environments may desirable. To enhance the utility of this research, this comparative evaluation of the Markov and semi-Markov chain continues in the next section, where the models' performance is used to classify the focus of visual attention.

5.7 Focus of visual attention classification experiments

The use of HMM and HSMM extends beyond classifying eye movement type. Eye tracker data may be used as part of a multimodal system, where the information of interest in the data is the focus of visual attention (FOVA) sequence[9]. The classification task for determining the FOVA sequence from eye position data is to assign each fixation to an area of interest in the visual scene.

This approach is sufficient for eye movement data captured with a precision eye tracker used in laboratory conditions where the eye movement type can be classified accurately. The eye movement data in the eye/speech corpus falls into this category.

When considering noisier eye tracking data, such as that in the Smart Eye dataset[10] and remote/wearable eye trackers used outside of the laboratory, from a statistical pattern classification perspective, it is desirable to delay decision making until the final stage of decoding to avoid non-recoverable error propagation [Hun87]. A decoding scheme that uncovers the FOVA by classification of eye movement into fixations and saccades prior to FOVA assignment carries the risk of misclassifying eye movement type and subsequently misclassifying the correct FOVA.

The aim of this section is to explore whether the Markov and semi-Markov chains in the HMM and HSMM can be used to encode the expected shifts between visual foci and the expected durations of visual attention, facilitating a more noise-robust decode of the FOVA from eye position data. Encoding the expected duration of visual attention may also enable discrimination between different cognitive tasks, an application briefly explored in §5.10.

[9] §2.2.3
[10] §5.2.3

These models in this study differ from previous studies which classify FOVA using either a pre-segmentation stage [YBZ02] or assigning saccades and fixations as separate states [Sal99]. Rather than conduct experiments with the same HMM and HSMM model structure used to classify eye movement type[11], variants of the HMMs and HSMMs were implemented to enable the balance of power between the durational and observation PDF in Viterbi decoding to be controlled. The model tests were performed on the Psycholinguistic study dataset.

5.7.1 Previous studies

The sequence of FOVA was mapped to those predicted by cognitive process models for reading and equation solving by Salvucci [Sal99]. Different strategies were identified from recovered FOVA sequences - e.g. exploring how one fixated on an equation when expanding parenthesis in a particular order identified (by prior knowledge) the cognitive strategies employed. The 2-state HMM model used for identifying fixations and saccades[12] was utilised by Salvucci in two variants to recover the FOVA sequence. In the first variant, it was used as a pre-segmentation stage to discount eye movements classified as saccades, prior to assigning eye movements to the closest potential FOVA. The second variant enhanced the fixation/saccade HMM by including eye position in addition to the eye velocity in the observation vector. A fixation and saccade state represented each potential FOVA in the visual scene, with a transition from each FOVA's fixation state to every other FOVA's saccade state and vice-versa.

When applying the HMM models to equation solving, both approaches were found to be superior to a baseline model which used a threshold approach of determining sequence of foci of visual attention. This baseline model assigned eye positions to the nearest object and discounted sequences of eye movements assigned to the same object if the sequence duration was short. The second variant, in delaying the classification of saccades and fixations, increased the accuracy slightly however it was prone to classify short fixations as saccades and thus discounted them from the fixation sequence, leading the author to recommend the first variant over the second.

This experiment solves the problem of classifying short fixations as saccades using Viterbi Reestimation to train the models and/or replacing the Markov chain with the semi-Markov chain[13].

[11] §5.6
[12] §5.6.1
[13] §5.6

5.7.2 HMM and HSMM model structure and variants

Ergodic HMM and HSMM classifiers were built. The model states represented areas of interest (foci) in the visual scene and the observations were eye positions. In using a HMM or HSMM state to represent a FOVA and not classifying the eye movement type, FOVA shifts were assumed to be instantaneous. This was an incorrect model assumption, as FOVA shifts require saccadic eye movement during which human visual processing is heavily attenuated [Mat74][14]. In effect, the models treated saccades as random noise, detecting a FOVA shift on the onset of a new fixation[15].

The HSMM is better equipped than the HMM to treat saccades as random noise, rather than assign the eye movement to a visual focus which falls along the saccades' trajectory. This is because it has explicit state duration PDFs that can assign low probabilities to short state durations. The geometric state duration PDFs in the HMM assign the highest probability to the shortest state durations, and are therefore more likely to classify saccadic eye movement which passes over a visual focus as a FOVA.

There were six HMM/HSMM types:

- HMM - A standard HMM.
- HMMR - A HMM with Richter observation PDFs.
- HSMM -A standard HSMM.
- HSMMR - A HSMM with Richter observation PDFs.
- HSMMT - A HSMM with task-specific state duration PDFs.
- HSMMRT - A HSMM with Richter observation PDFs and task-specific state duration PDFs.

All model types had four states representing the four objects that made up the visual scenes in the Psycholinguistic study dataset. These were ergodic since FOVA shifts could occur between any pair of the four foci, and used eye position as the observation[16]. The HMM was implemented in two variants. The first (HMM) used a

[14] This is why we do not see an obvious blur every time we move our eyes.

[15] Assuming that HMM/HSMM decoding is forward pass (Viterbi). Baum-Welch decoding, which uses a forward and backward pass, would likely detect the FOVA while the eye position is half way between foci.

[16] (5.7)

standard radially asymmetric 2-dimensional Gaussian PDF with diagonal covariance to represent the distribution of eye positions over an object for the state observation PDFs. The justification for using single component Gaussian PDFs for observation probabilities was discussed earlier in this chapter[17].

The HMMR added an additional Gaussian component to each dimension in the observation PDF having the same mean but a much larger standard deviation. In these experiments the standard deviation was 10 times the standard deviation of the first component. The use of such a component was first proposed by Richter [Ric86] and is referred to as the Richter Distribution. This second component lowered the dominance of the observation PDF in Viterbi decoding, as it reduced the difference between the states' observation PDFs. This in turn allows the state transition probabilities (i.e. state duration PDFs) to play a more dominant role in decoding.

These two variations in observation PDF were also modelled for the HSMM, which was realized in four variants - HSMM, HSMMR, HSMMT, HSMMRT. The 'R' in the model names refers to the use of the Richter distribution and the 'T' stands for 'task specific'. The HSMM variants' observation PDFs were equivalent to the HMMs'. The HSMMs' state duration PDF was used in two ways. For the HSMM and HSMMR, the duration PDF for each state was uniformly distributed between 100ms and 600ms, representing the range of attention durations (i.e. fixation durations) expected from published findings on general eye movement for various cognitive tasks such as reading, visual search and scene perception [Ray98]. For the HSMMT and HSMMRT, each duration PDF was a single truncated Gaussian estimated using the psycholinguistic study dataset. The motivation behind the HMMT and HMMRT was to provide task-specific models, i.e. the task being that performed by subjects in the psycholinguistic dataset.

Observation PDFs and transition probabilities for all models were estimated using Viterbi Reestimation. The HSMMT and HSMMRT also used Viterbi Reestimation to estimate state duration PDFs. To initialise the models, the Psycholinguistic study eye position data was labelled with the FOVA for supervised training. Determining the FOVA from the psycholinguistic study data was straightforward because there were only four objects in the visual scene. The onsets of fixations were used to indicate the FOVA shift, and the FOVA was identified from the nearest object. The Viterbi decoded state sequence from this model was used to reestimate each model until its likelihood reached a local maximum. The labelled data served as a 'ground

[17] §5.5

truth' in the evaluation measures[18].

For Viterbi reestimation in the HSMMT and HSMMRT, the state duration was calculated as:

$$\mu_d = \frac{\sum_{n=1...N} d_i(n)}{N_i} \tag{5.12}$$

Where $d_i(n)$ is the n^{th} duration of state i from Viterbi decoded state sequence. N_i is the total number of durations of state i from Viterbi decoded state sequence.

Viterbi reestimation was carried out 10 times to train the models. This was determined from observing that reaching local maxima took on average 5 reestimations and that twice this number of reestimations was a suitable safety margin to ensure that a local maximum was reached.

5.7.3 Baseline 'Nearest Neighbour' model

In addition to the HMM and HSMM model variants, a baseline 'Nearest Neighbour' ('NN') model was built. In common with the HMM and HSMM, the model consisted of a radially asymmetric 2-dimensional Gaussian PDF for each object with the eye positions as the observation. The PDFs were the same as those used for the observation PDFs in the HMM. The FOVA for a given eye position was determined from the object observation PDF which had the highest probability. Any short duration ($< 100ms$) FOVA in the resulting sequence were reassigned to the previous FOVA to avoid misclassifying saccadic eye movement as a FOVA. The baseline allowed the performance of the Markov and semi-Markov chain to be measured against employing a simple threshold-type model to deal with saccadic movement. The NN model in the FOVA classifier is equivalent to a 4-component GMM with the addition of thresholding to filter out saccadic eye movement. Each component in the GMM represents a visual focus.

5.7.4 Evaluation measures

In a similar vein to the eye-movement type classification experiments, the following measures evaluated the performance:

Accuracy: Proportion of eye positions with added noise classified correctly by a model compared to the ground truth.

[18] §5.7.4

Instability: Number of state transitions (i.e. classifications) made by a model compared to the ground truth. State transitions represent onsets of FOVA.

The Accuracy measure is of primary importance for comparing models, with Instability providing a further insight into the models' classification behaviour. The ground truth FOVA sequence for these experiments was determined by allocating the eye positions to the nearest object[19] using the Euclidean distance.

5.7.5 Tests Performed on the Psycholinguistic study dataset

The following tests were performed on the Psycholinguistic study dataset[20]. There were 3 experiments:

Experiment 1: Individual session models, trained each model type for an individual session (total of 50 sessions) and then used these models to decode the same session data with progressive levels of Gaussian noise added[21]. While use of the same data source for training and evaluation was not desirable, the objective was to measure the models' relative resilience to noise and their performance with observation PDFs specific to the individual session.

Experiment 2: Consolidated models, used 25 of the sessions as a training set of data, with the remaining 25 used as an evaluation set. The training data was used to produce a single model for each model type, which was used to decode the evaluation set. Since the visual scene used in each session of the training set used different objects, it was anticipated that the models' accuracy and instability would be more dependent on state duration PDFs than observation PDFs compared to experiment 1. The rationale for this is described below.

For *Experiment 3: Additional noise*, Experiment 2 was repeated adding additional noise to observe the relative degradation of the models.

The results from Experiments 1 and 2 allowed the analysis of the effect of using HMMs and HSMMs for decoding eye data with observation PDFs specific to a session, albeit with a slightly compromised data source (Experiment 1), together with more general observation PDFs with a non-compromised data source (Experiment 2). Analyzed together, the results aim to give a good overall assessment of the utility of the Markov and Semi-Markov chain for uncovering the FOVA from eye-tracking

[19] §5.7.2
[20] §5.2.2
[21] §5.5

data when treating saccades as random noise. The baseline 'NN' model enabled the HMMs and HSMMs to be compared against a model that has the equivalent observation PDF but no Markov or semi-Markov chain or Richter PDF.

The dominance of the state transition probabilities and state duration PDF in decoding was controlled at both the model and experiment level. At the model level, the Richter distribution was intended to lower the effect of the observation PDF on decoding by lowering the discrimination between the state observation PDFs of the model. At the experiment level, in Experiment 2 the observation PDF's dominance in all HMM and HSMM models was relaxed by training a single model of each model type for all sessions. Although the HMM and HSMMs' state transition probabilities and state duration PDFs were session specific for Experiment 1, they were assumed equivalent to those in Experiment 2 because the participant's task of viewing objects in a set sequence was common to all sessions.

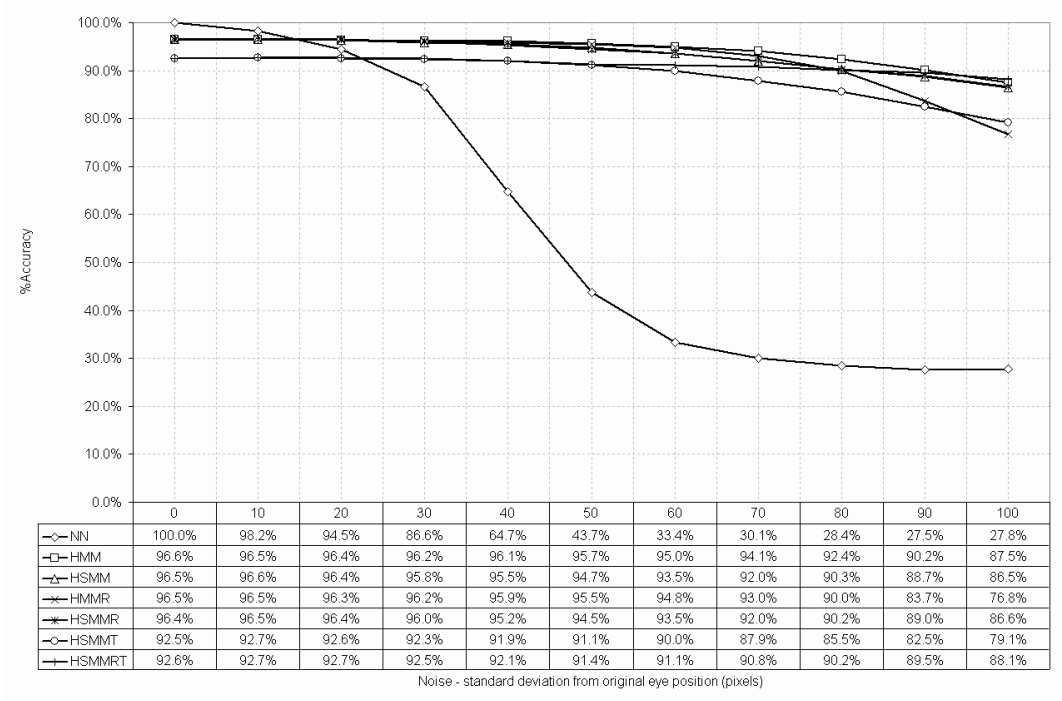
Since Experiment 2 may be regarded as a progression from Experiment 1 with relaxed observation PDFs in relation to decoding, only Experiment 2 was repeated with additional noise for Experiment 3.

5.7.6 Experiment 1 Results

(Separate models for each of the 50 sessions. Models' observation PDFs specific to each session.)

Figure 5.17 shows that all HMM and HSMM variants reacted similarly to noise. The HMMR lost some degree of accuracy at high noise levels compared to the other variants, demonstrating the negative effect of relaxing the observation PDF sensitivity to noise using the Richter distribution without the counterbalance of state duration PDF. Models with an explicit state duration PDF but without Richter observation PDFs (HSMMT and HSMM) performed marginally worse than those with Richter observation PDF (HSMMRT and HSMMR). The baseline 'NN' model performed worst as noise was added which made eye positions lie between objects (i.e. at a standard deviation of 30 pixels in Figure 5.19), resulting in accuracy that asymptotes to chance at 25%, given the classification task is a four-class problem and objects were viewed for similar durations and a similar number of times.

Figure 5.18 shows that the HSMMRT and the HMMR were the most stable, with instabilities of 116% and 126% respectively. The HSMMR's stability however coincided with a fall-off in accuracy. Other model variants were more unstable, although their accuracy was comparable to the HSMMRT. The HSMMRT outperforms the



	0	10	20	30	40	50	60	70	80	90	100
NN	100.0%	98.2%	94.5%	86.6%	64.7%	43.7%	33.4%	30.1%	28.4%	27.5%	27.8%
HMM	96.6%	96.5%	96.4%	96.2%	96.1%	95.7%	95.0%	94.1%	92.4%	90.2%	87.5%
HSMM	96.5%	96.6%	96.4%	95.8%	95.5%	94.7%	93.5%	92.0%	90.3%	88.7%	86.5%
HMMR	96.5%	96.5%	96.3%	96.2%	95.9%	95.5%	94.8%	93.0%	90.0%	83.7%	76.8%
HSMMR	96.4%	96.5%	96.4%	96.0%	95.2%	94.5%	93.5%	92.0%	90.2%	89.0%	86.6%
HSMMT	92.5%	92.7%	92.6%	92.3%	91.9%	91.1%	90.0%	87.9%	85.5%	82.5%	79.1%
HSMMRT	92.6%	92.7%	92.7%	92.5%	92.1%	91.4%	91.1%	90.8%	90.2%	89.5%	88.1%

Figure 5.17: Experiment 1 accuracy vs. noise showing all HMM models performing similarly to added noise.

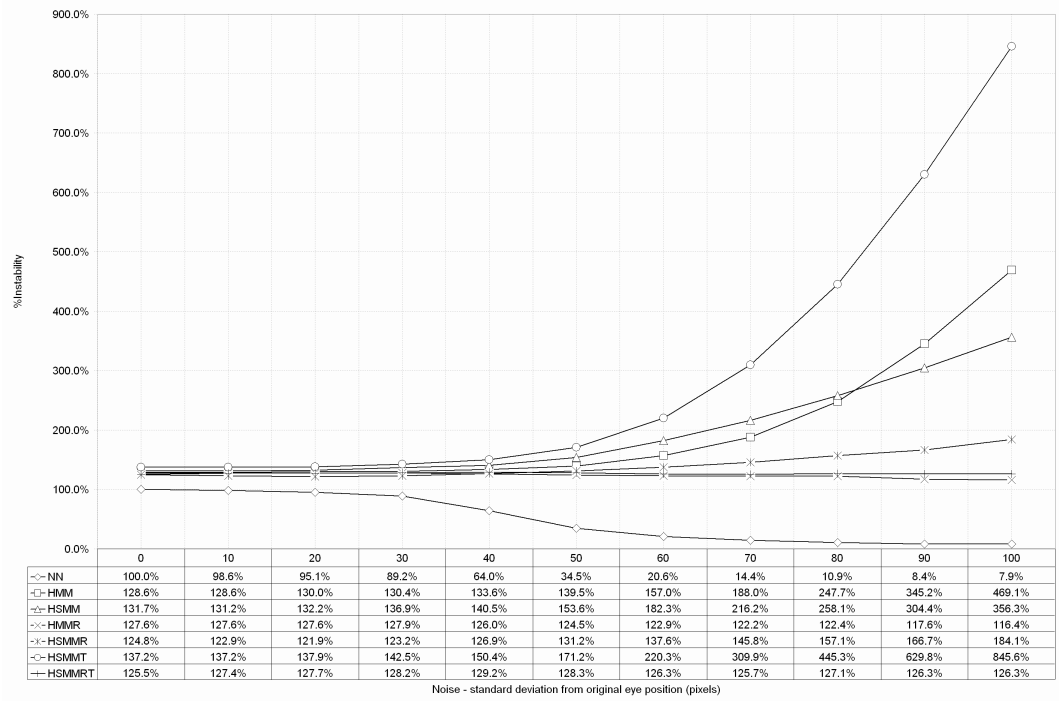
HSMMR in stability demonstrating that the task-dependent explicit state duration PDF increases overall performance over task-independent state durations.

The overall impression from Experiment 1 was that using models trained and evaluated on the same session data to ensure accurate observation distribution PDFs reduced the difference in effect between the Markov and semi-Markov chains during decoding. The presence of a chain was shown to be beneficial when the noise level was increased to the point where the observation PDF became an unreliable indicator of the nearest FOVA.

5.7.7 Experiment 2 Results

(Single models trained on 25 sessions and used to classify 25 sessions. Models' observation PDFs not specific to individual session.)

Compared to Experiment 1, the results in Experiment 2 better discriminated between the HMM and HSMM variants. All HMM and HSMM variants outperformed the baseline 'NN' method. The HSMMR and HSMMRT were the most resistant to noise. They maintained relatively high levels of accuracy (89% and 83% respec-



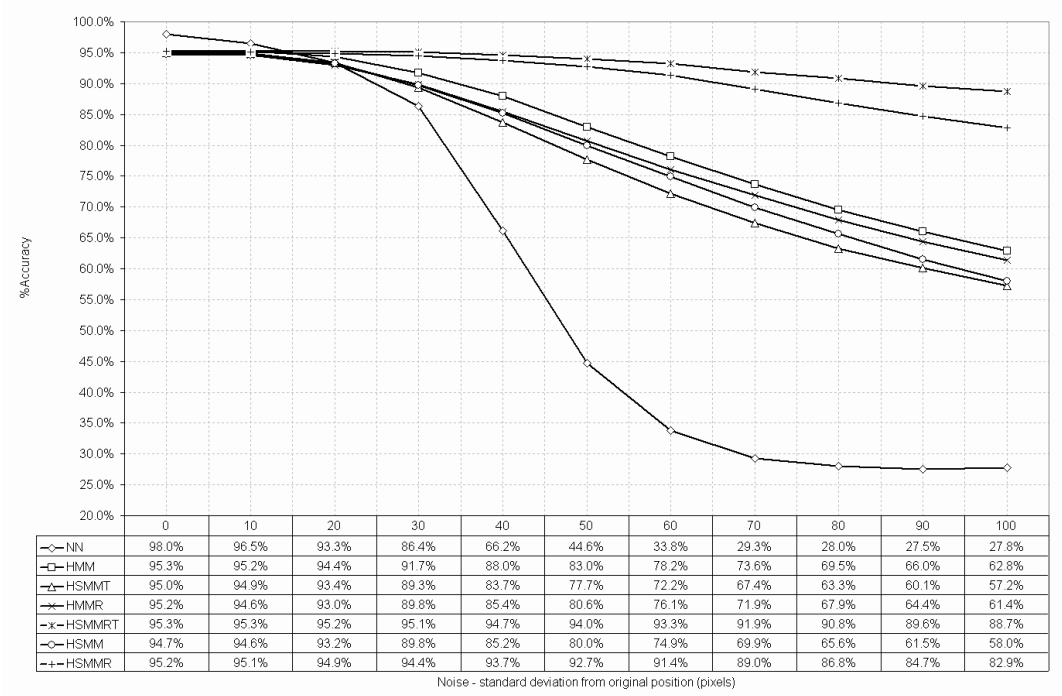
	0	10	20	30	40	50	60	70	80	90	100
NN	100.0%	98.6%	95.1%	89.2%	64.0%	34.5%	20.6%	14.4%	10.9%	8.4%	7.9%
HMM	128.6%	128.6%	130.0%	130.4%	133.6%	139.5%	157.0%	188.0%	247.7%	345.2%	469.1%
HSMM	131.7%	131.2%	132.2%	136.9%	140.5%	153.6%	182.3%	216.2%	258.1%	304.4%	356.3%
HMMR	127.6%	127.6%	127.6%	127.9%	126.0%	124.5%	122.9%	122.2%	122.4%	117.6%	116.4%
HSMMR	124.8%	122.9%	121.9%	123.2%	126.9%	131.2%	137.6%	145.8%	157.1%	166.7%	184.1%
HSMMT	137.2%	137.2%	137.9%	142.5%	150.4%	171.2%	220.3%	309.9%	445.3%	629.8%	845.6%
HSMMRT	125.5%	127.4%	127.7%	128.2%	129.2%	128.3%	126.3%	125.7%	127.1%	126.3%	126.3%

Figure 5.18: Experiment 1 instability vs. noise showing the HSMMRT and HMMR most stable to noise.

tively) compared to the other models when the eye position data was corrupted with Gaussian random noise with a standard deviation of 100 pixels (Figure 5.19). The HSMMRT performed marginally better than the HSMMR due to the state duration PDF representing the attention durations of the psycholinguistic study task as opposed to the duration of fixation as a characteristic of general eye movement, independent of task.

The HMM, HMMR, HSMM, and HSMMT had similar accuracy levels, ranging from 57% to 63%. The HMM and HMMR performed marginally better than the HSMM and HSMMT, demonstrating that the models using a semi-Markov chain required the Richter distribution as part of the observation PDF to outperform the models using a Markov chain.

The ordinal rank of all models did not change in terms of accuracy as noise was added. All HMM and HSMM variants beat the baseline 'NN' method, whose accuracy at the maximum noise level was 28%. The reduction in accuracy tended to be constant for all HMM and HSMM variants, whereas the 'NN' model's decline was more dramatic due to it ignoring FOVA durations of less than 100ms which became more common as noise was added to the data.

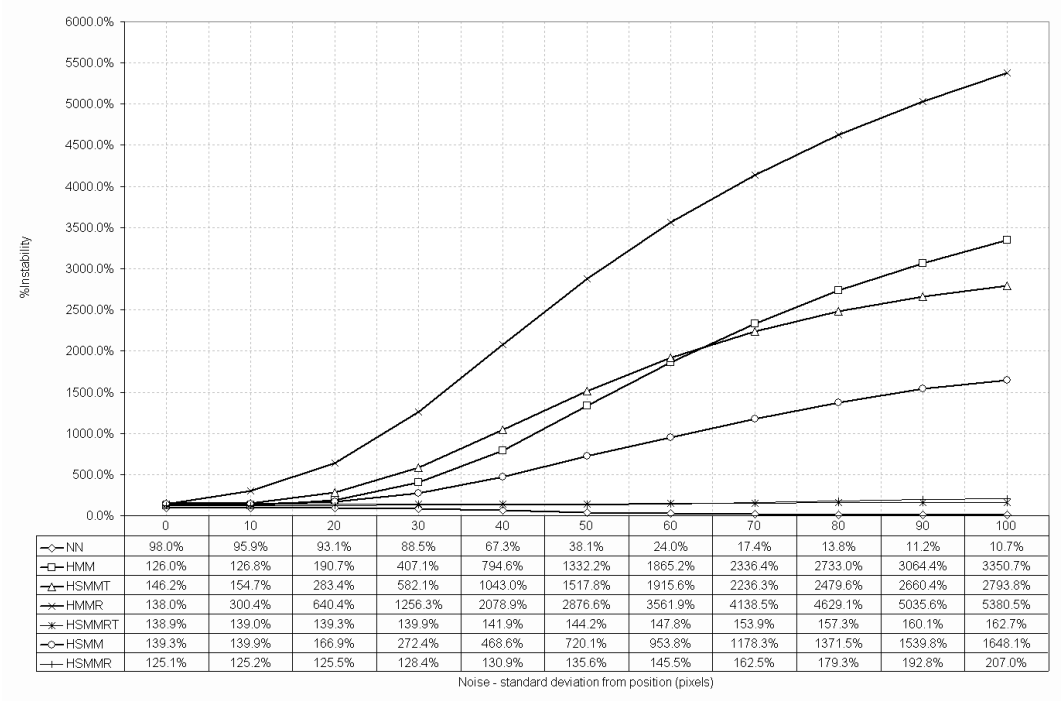


	0	10	20	30	40	50	60	70	80	90	100
NN	98.0%	96.5%	93.3%	86.4%	66.2%	44.6%	33.8%	29.3%	28.0%	27.5%	27.8%
HMM	95.3%	95.2%	94.4%	91.7%	88.0%	83.0%	78.2%	73.6%	69.5%	66.0%	62.8%
HSMMT	95.0%	94.9%	93.4%	89.3%	83.7%	77.7%	72.2%	67.4%	63.3%	60.1%	57.2%
HMMR	95.2%	94.6%	93.0%	89.8%	85.4%	80.6%	76.1%	71.9%	67.9%	64.4%	61.4%
HSMMRT	95.3%	95.3%	95.2%	95.1%	94.7%	94.0%	93.3%	91.9%	90.8%	89.6%	88.7%
HSMM	94.7%	94.6%	93.2%	89.8%	85.2%	80.0%	74.9%	69.9%	65.6%	61.5%	58.0%
HSMMR	95.2%	95.1%	94.9%	94.4%	93.7%	92.7%	91.4%	89.0%	86.8%	84.7%	82.9%

Figure 5.19: Experiment 2 accuracy vs. noise showing HMM with explicit state duration PDFs incorporating Richter distribution most resistant to added noise.

The Instability results complemented the Accuracy results. The Instability of the HSMMRT and HSMMR at the maximum noise level was 163% and 207% respectively (i.e. 63% and 107% more FOVA classifications were made compared to the gold standard respectively). The other models were more unstable, the HSMM less so with an Instability of 1648%, whereas the HMMR fared worse with 5380%. This was expected, as the HSMM would make longer state durations more likely resulting in less state transitions. The HMMR was the most unstable model, demonstrating that relaxing a model's classification dependency on the observation PDF using a Richter distribution must be counterbalanced by the model's other strengths, such as the state duration PDF in the HSMMR and HSMMRT. The baseline model had an Instability measurement of 9% due to short FOVA durations being ignored and thus fewer classifications being made. Figure 5.20 shows the instability results.

Experiment 2 showed a demonstrable benefit in using the semi-Markov chain for FOVA classification over the Markov chain, although the benefit was observed only in the HSMMR and HSMMRT where the observation PDFs dominance in Viterbi decoding was relaxed.



	0	10	20	30	40	50	60	70	80	90	100
NN	98.0%	95.9%	93.1%	88.5%	67.3%	38.1%	24.0%	17.4%	13.8%	11.2%	10.7%
HMM	126.0%	126.8%	190.7%	407.1%	794.6%	1332.2%	1865.2%	2336.4%	2733.0%	3064.4%	3350.7%
HSMMT	146.2%	154.7%	283.4%	582.1%	1043.0%	1517.8%	1915.6%	2236.3%	2479.6%	2660.4%	2793.8%
HMMR	138.0%	300.4%	640.4%	1256.3%	2078.9%	2876.6%	3561.9%	4138.5%	4629.1%	5035.6%	5380.5%
HSMMRT	138.9%	139.0%	139.3%	139.9%	141.9%	144.2%	147.8%	153.9%	157.3%	160.1%	162.7%
HSMM	139.3%	139.9%	166.9%	272.4%	468.6%	720.1%	953.8%	1178.3%	1371.5%	1539.8%	1648.1%
HSMMR	125.1%	125.2%	125.5%	128.4%	130.9%	135.6%	145.5%	162.5%	179.3%	192.8%	207.0%

Figure 5.20: Experiment 2 instability vs. noise showing HMMs with explicit state duration PDF with Richter distribution most stable to added noise.

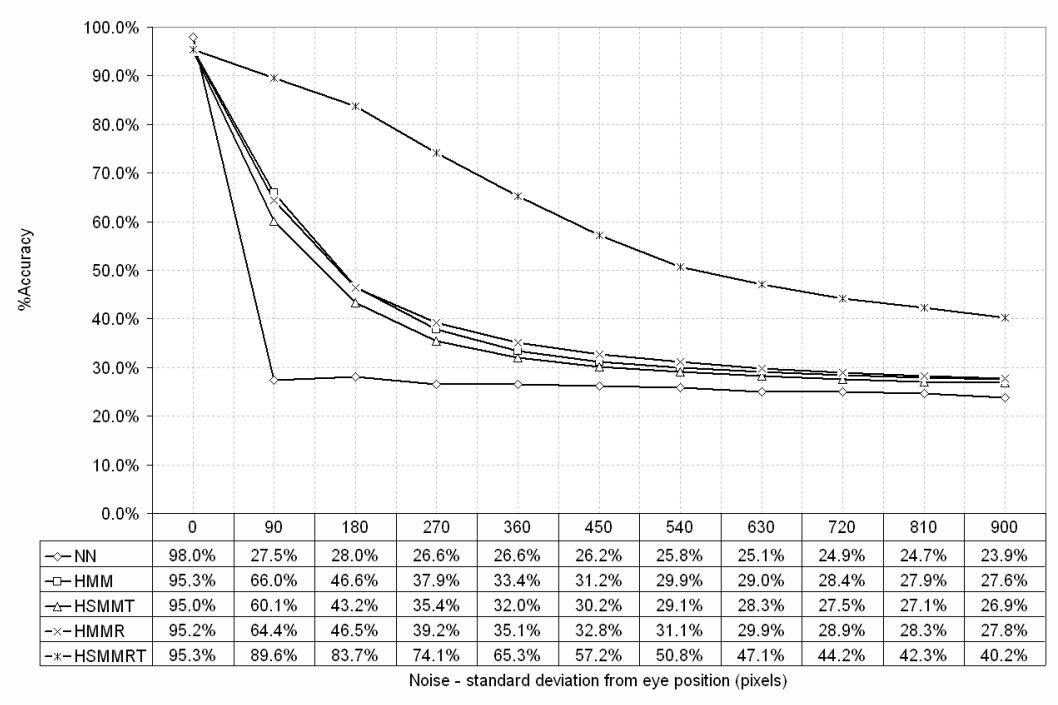
5.7.8 Experiment 3 Results

(Repeat of Experiment 2 using additional noise)

The findings of experiment 3 supported those of Experiment 2, showing the HSMMRT as the most accurate (Figure 5.21) and stable (Figure 5.22) model. The HMMR was the least stable and accurate. The models with task-independent duration PDFs (HSMM and HSMMR) were omitted from the experiment since these were shown to perform worse than the task-dependant models (HSMMT and HSMMRT) in Experiment 2. The HMM, HMMR and HSMMT performed similarly.

5.7.9 Discussion

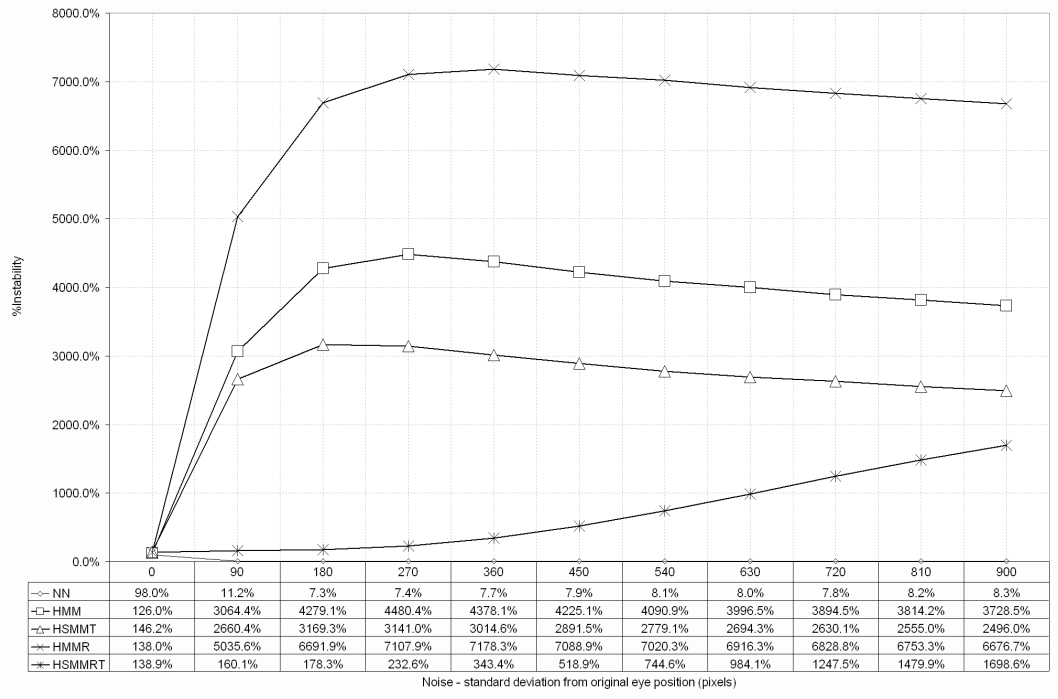
With the session-specific models (Experiment 1), all model variants displayed improved performance against the baseline 'NN' model when noise was added, showing the Markov or semi-Markov chain had some benefit. When single models were trained and evaluated on multiple sessions (Experiment 2) the HSMMR and HSMMRT models were more resilient to high levels of noise, showing a large improvement in accuracy and stability compared to all other HMM-based methods. An explicit state dura-



	0	90	180	270	360	450	540	630	720	810	900
NN	98.0%	27.5%	28.0%	26.6%	26.6%	26.2%	25.8%	25.1%	24.9%	24.7%	23.9%
HMM	95.3%	66.0%	46.6%	37.9%	33.4%	31.2%	29.9%	29.0%	28.4%	27.9%	27.6%
HSMMT	95.0%	60.1%	43.2%	35.4%	32.0%	30.2%	29.1%	28.3%	27.5%	27.1%	26.9%
HMMR	95.2%	64.4%	46.5%	39.2%	35.1%	32.8%	31.1%	29.9%	28.9%	28.3%	27.8%
HSMMRT	95.3%	89.6%	83.7%	74.1%	65.3%	57.2%	50.8%	47.1%	44.2%	42.3%	40.2%

Noise - standard deviation from eye position (pixels)

Figure 5.21: Experiment 2 accuracy vs. noise repeated for additional noise, showing HSMMRT outperforming all other HMM models.



	0	90	180	270	360	450	540	630	720	810	900
NN	98.0%	11.2%	7.3%	7.4%	7.7%	7.9%	8.1%	8.0%	7.8%	8.2%	8.3%
HMM	126.0%	3064.4%	4279.1%	4480.4%	4378.1%	4225.1%	4090.9%	3996.5%	3894.5%	3814.2%	3728.5%
HSMMT	146.2%	2660.4%	3169.3%	3141.0%	3014.6%	2891.5%	2779.1%	2694.3%	2630.1%	2555.0%	2496.0%
HMMR	138.0%	5035.6%	6691.9%	7107.9%	7178.3%	7088.9%	7020.3%	6916.3%	6828.8%	6753.3%	6676.7%
HSMMRT	138.9%	160.1%	178.3%	232.6%	343.4%	518.9%	744.6%	984.1%	1247.5%	1479.9%	1698.6%

Noise - standard deviation from original eye position (pixels)

Figure 5.22: Experiment 2 instability vs. noise repeated for additional noise showing HSMMRT outperforming all other models.

tion PDF and compensating for noise by using a Richter distribution for observation PDF increased the performance (i.e. HSMMR and HSMMRT), but utilizing one or the other on their own (i.e. HMMR or HSMM) showed no improvement over using the HMM. In additional noise (Experiment 3) the performance of the HSMMRT degraded less and more gracefully than other models, supporting the findings from the previous experiments.

The main finding therefore was that the semi-Markov chain shows benefits over using the Markov chain but only if the observation PDFs dominance during Viterbi decoding can be relaxed. The aim of proving the value of the Markov and semi-Markov chain was achieved by comparing the models' performances against a single baseline model. Further experiments could compare these models against a range of baseline models that do not use Markov or semi-Markov chains in order to refute the null hypothesis.

The evaluation measures in this experiment penalised small misalignments between the FOVA sequence in the ground truth and the models because they considered each eye position classification separately rather than considering successive eye positions classified as the same FOVA as one instance of a FOVA. Consequently, the accuracy level for any model with no noise in the eye position data was never 100%, even though the actual FOVA sequence was identical to the ground truth. Future experiments could utilise an error rate measure equivalent to WER[22] in speech recognition which considers the number of insertions, deletions and substitutions in the FOVA sequence as a measure of performance.

The novel contribution in this section is the use of semi-Markov chain to make noise-robust interpretations regarding where someone looks by treating saccades as random noise. This approach assumes that the FOVA is continuous in that it does not break for saccades. A continuous indication of attention may be more congruent with Posner's attention spotlight metaphor discussed in §2.2.3. To increase the utility of using HMM and HSMM-based FOVA classifiers, future studies should address non-static visual scenes where objects must be recognised and tracked with observation PDFs updated accordingly.

Gaussian probability distributions represented observation and state duration PDFs, and it is likely that asymmetric distributions, such as gamma, may yield better performance for state duration PDFs. Another extension would be to use the first order derivative of eye position (velocity) in addition to eye position in the

[22] §4.2

observation vector to increase the model's ability to deal with saccades.

To successfully utilise these models, the sequences of objects viewed for a task must be predictable so that the state transitions can be reliably estimated. The Psycholinguistic dataset is a controlled experiment where the task explicitly requires participants to look at objects in sequence. Less controlled tasks (e.g. the eye/speech corpus task) have greater variation in expected FOVA sequences. The value of the Markov or semi-Markov chain is lessened in cases where the FOVA sequence is random, encoding only a-priori estimations of average attention durations.

5.7.10 Summary

A HSMM which has an explicit state duration PDF representing task-constrained visual attention (HSMMRT) is a more stable and accurate way to uncover the FOVA from noisy eye tracker data than using standard HMM with an inherent state duration PDFs. However this gain in performance was only observed when the observation distribution PDF's dominance on classification decision in the model was relaxed by adding an additional Gaussian component to the observation distribution PDF - the Richter distribution. When a standard HMM had the Richter distribution observation PDF, it weakened the model's robustness and produced variable results.

5.8 Computationally efficient implementations of explicit state distributions

The HSMM implementations for the eye-type and FOVA classifiers were pure in the sense that any state duration PDF in an ergodic HSMM could be modeled explicitly by specifying the distribution form. While suitable for these experiments, the solution is impractical for many applications due to the computational time required for HSMM Viterbi decoding. More pragmatic solutions exist that use HMMs and more expedient Viterbi decoding.

Tying sequences of states in a left to right HMM topology produces a negative binomial distribution for the overall duration distribution, which is discrete version of the gamma distribution [Bil02]. Thus the FOVA classifier HMM state representing a single visual focus could be replaced with left to right multiple state-tied HMMs.

Alternatively, the HMM can utilise dynamic state transition probabilities, where state transitions are defined as a function of state occupancy. This enables the state duration PDF to deviate from its geometric form [Dju02].

5.9 Auto-discovery of visual foci using clustering

To detect and locate the visual foci, the visual scene can be examined using image-processing techniques to discover areas of potential interest. Alternatively, eye movements made when people look at a given visual scene can be analysed. With the latter in mind, clustering algorithms were applied to the eye/speech corpus data to determine the number and locations of visual foci in the corpus maps.

There has been some research into using data-driven clustering algorithms to discover visual foci from eye position data by the use of the most common clustering algorithm, K-Means clustering [PS00] and derivatives [SD04]. Rather than extend previous research by considering a myriad of other clustering algorithms, as a preliminary study this section describes the application of basic K-means clustering to the eye/speech corpus. Similarity measures on clustering results were used to determine the optimum number of clusters, which in turn gave an indication of the number of states required for a FOVA classifier used in the eye/speech corpus.

5.9.1 Previous studies

Identifying the visual foci in complex scenes from fixation sequences has attracted previous research with applications in usability research and image compression, as well as research into scene perception. To determine visual foci, mapping fixation sequences onto the visual scene to determine areas viewed by the eye tracker wearer is the simplest and oldest method [MM67]. Modulating the visual scene with a frequency histogram of the fixation distribution is a common method [Lat88] [Woo02] and one used already in this chapter (Figure 5.25).

Determining the number of states in a HMM or HSMM-based FOVA classifier requires prior knowledge of the visual scene. In the previous FOVA classification experiments[23], estimating the number and locations of visual foci was trivial, as the Psycholinguistic study dataset comprises images of 4 objects and one state per object. The eye/speech corpus is a more complex visual scene. The potential visual foci are the objects on the map and the route around these objects. Although the number of states corresponding to objects can be confidently stated from inspection of the image, the route through the map cannot as participants were observed to fixate anywhere along it with intersections along the route (i.e. when it crossed over itself) attracting the most fixations.

[23] §5.7

5.9.2 K-means clustering

K-means clustering has potential for discovering the number and locations of visual foci. It consists of an initial random allocation of points to clusters, followed by repeated allocation and update steps until the allocation of points to clusters becomes stable (i.e. exhibits minimal or no change). K-means clustering attempts to minimise the average distance between a data point and its nearest cluster centre.

In its simplest form, the k-means function $i(t)$ allocates each of T N-dimensional points in a set, $x = \{x_1, \ldots, x_t, x_T\}$, to one of K clusters. $\mu = \{\mu_1, \ldots, \mu_k, \mu_K\}$, according to a distance measure between cluster centroids and points, $d(\mu_k, x_t)$[24]:

$$i(t) \quad = \quad \arg\min_k d(\mu_k, x_t) \tag{5.13}$$

After allocation, the cluster centroids, μ_k, are updated based on the set of points allocated to them:

$$\mu_k \quad = \quad \frac{\sum_{i(t)=k} x_t}{T_k} \tag{5.14}$$

Where T_k is the number of points allocated to cluster k.

When applying k-means clustering, one must consider the number of clusters, distance measure and methods for dealing with clusters with low numbers of points allocated to them. Even if there is no clustering in a set of data, k-means clustering will still cluster it. Due to initial random allocation of points to centroids, it is standard to run the algorithm multiple times then select the best set of clusters based on, for example, the centroid set with the lowest average distance between points and cluster centroids. Different strategies exist for dealing with clusters allocated a minimum number of points. Two common methods are to reallocate a random number of points to the centroid (the method used in this experiment), or remove the centroid altogether.

5.9.3 Measuring the optimum number of clusters

Similarity measures are an approach for determining the optimum number of clusters, K, to apply to the data. The average 'similarity measure', $s(x)$, across all points yields a value indicating the clustering effectiveness:

[24]The generalised distance measure between 2 N-dimensional points is often used, $(\sum_{n=1}^{N} |x_{1,n} + x_{2,n}|^m)^{\frac{1}{m}}$. The case where $m = 2$ is referred to as the Euclidean distance.

$$s(x) = \frac{\sum_{t=1}^{T} s(x_t)}{T} \tag{5.15}$$

Where $s(x_t)$ is the similarity measure for a data point:

$$s(x_t) = \frac{r(x_t) - l(x_t)}{l(x_t)} \tag{5.16}$$

Where $l(x_t)$ is the average distance from point x_t to all other points in its allocated (i.e. nearest) cluster and $r(x_t)$ is the average distance from point x_t to all other points allocated to the next nearest cluster:

$$l(x_t) = \frac{\sum_{i(j)=i(t), j \neq t} d(x_t, x_j)}{T_{i(t)}} \tag{5.17}$$

$$r(x_t) = \frac{\sum_{i(j)=c(t)} d(x_t, x_j)}{T_{c(t)}} \tag{5.18}$$

$T_{i(t)}$ and $T_{c(t)}$ are the number of points allocated clusters indicated by $i(t)$ and $c(t)$ respectively. $c(t)$ is the next nearest cluster to x_t:

$$c(t) = \arg \min_{n, n \neq i(t)} d(\mu_n, x_t) \tag{5.19}$$

It is desirable to achieve a high similarity measure where the average distance between points belonging to the same cluster is low, relative to the average distance between points in one cluster and points in the closest neighbouring cluster. When applying K-means clustering to eye movement position data, selecting the number of clusters which yields the highest similarity measure is a method to automatically discover the number of areas of interest in the visual field.

5.9.4 Eye/speech corpus foci discovery

To demonstrate whether visual foci in the case of complex scenes can be learnt from fixations in complex visual scenes, the k-means clustering algorithm was applied to eye positions indicated by fixation events for session M3G2F1 in the eye/speech corpus. The number of clusters was varied and similarity measures obtained. K-means was run 50 times for each value of K and the run with the highest value for $s(x)$ was used.

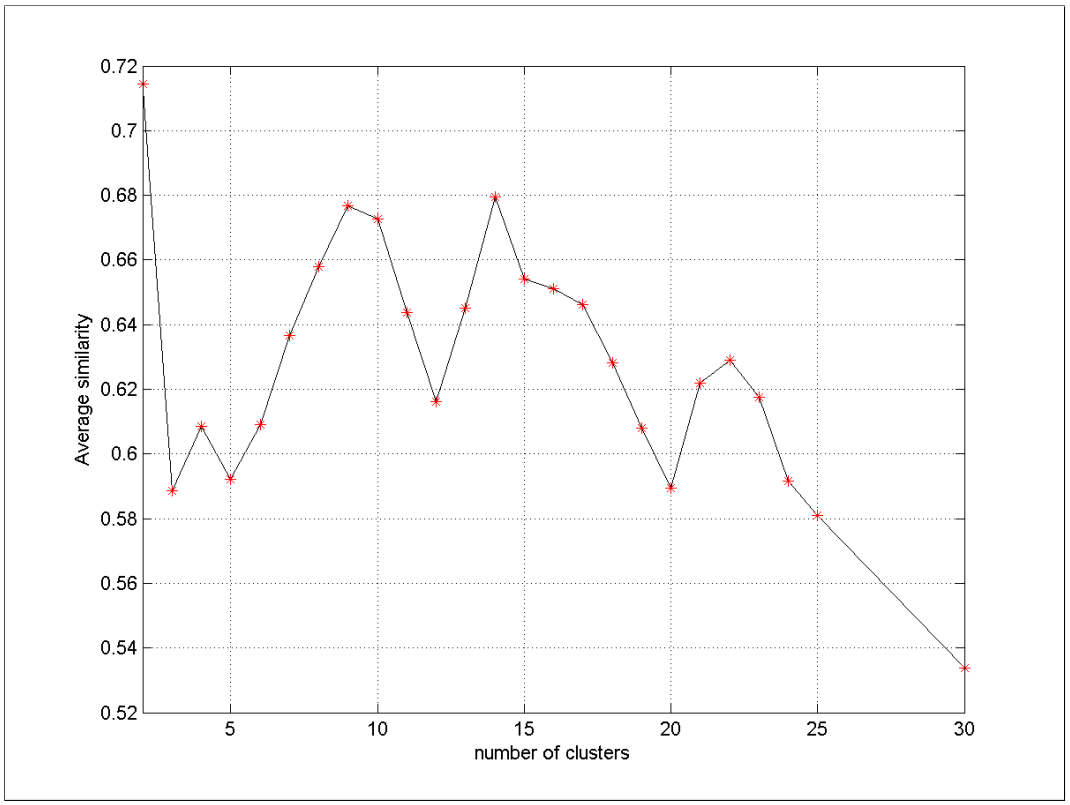


Figure 5.23: Similarity measures against increasing the number of clusters in the eye/speech corpus. High values of similarity correspond to a better fit for clusters.

Figure 5.23 shows that as K was increased from 12, $s(x)$ peaked at $K = 14$ with $s(x) = 0.68$ then fell until $K = 20$ with $s(x) = 0.59$, before a slight rise at $K = 22$ ($s(x) = 0.63$), then a terminal decline in $s(x)$ onwards. Given that the number of objects in the visual scene was 12, it was reasonable to assume that $K >= 12$.

The results of clustering FOVA in the eye/speech corpus demonstrated a difficulty in identifying a specific number of visual foci (K). The likely reason for this is that the FOVA does not necessarily correspond to the visual focus being projected onto the centre of the fovea - i.e. the eye position. $K = 12$ resulted in a lower $s(x)$ than higher values of K indicating the presence of route features as visual foci. The decline in similarity for $K > 22$ gives an indication of an upper bound estimate for K. Figure 5.24 shows that cluster centroids did not correspond exactly to either object locations or route intersections for $K = 15$. Other values of K showed similar results.

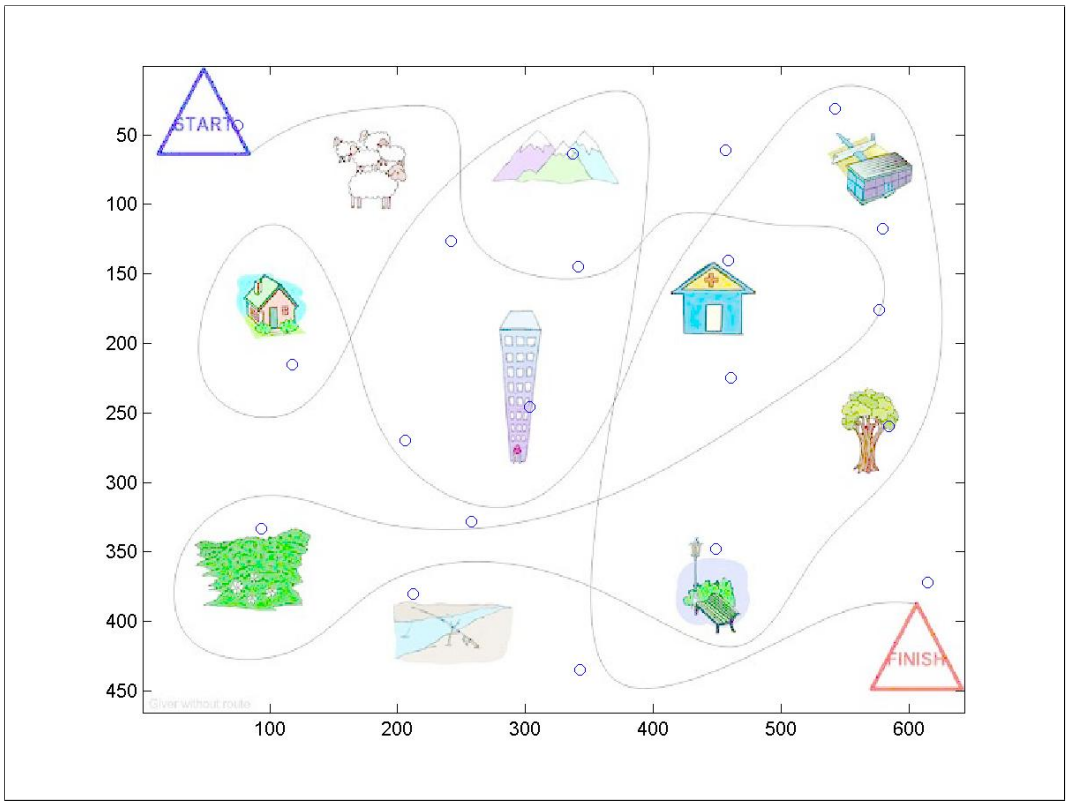


Figure 5.24: K-Means estimated centroids of interest in a session of the eye/speech corpus. Centres of centroids are small circles.

5.9.5 Discussion

The results for automatic discovery of locations and number of visual foci using k-means clustering on the eye/speech corpus eye movement data showed no clear indication of the optimum value of K; this was due to fixations along the route in addition to objects. From the observation of frequency histograms, two types of clusters were evident: spherical clusters for fixations on objects and route intersections, and linear clusters for fixating along the route. K-means clustering accounted for only the spherical clusters. Further work could consider clustering using both spherical and other shaped clusters (e.g. linear) by using a suitable distance measure.

5.10 Behaviour recognition using eye movement

As discussed previously[25], FOVA shifts are due to both stimulus driven control and knowledge driven control [Hen03].

[25] §2.2.3

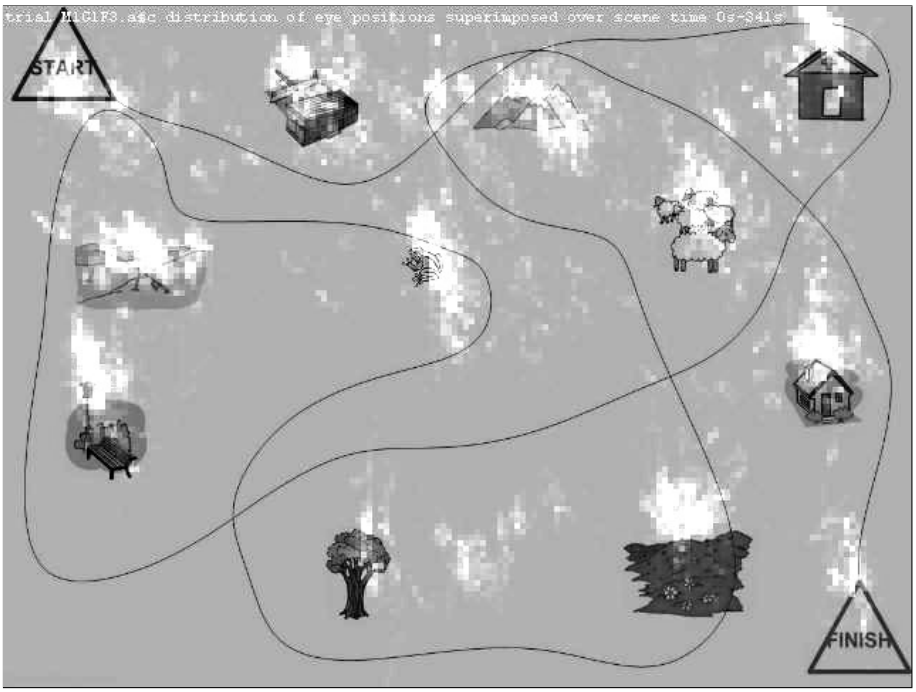


Figure 5.25: Example of one of the maps displayed to the participants, showing objects and route, superimposed with an indication of the distribution of the participant's gaze direction. The distribution of eye movement is indicated by brightness. This figure shows eye movements while the participant described objects.

In terms of knowledge driven control, two user goals (behaviours) were identified from participant's speech in the eye/speech corpus during the corpus quality evaluation[26]- describing landmarks on a map and describing the route around the landmarks. From an analysis of the distribution of eye positions across the visual scene (see Figure 5.25 and Figure 5.26), it was clear that the two behaviours were also identifiable from the eye-positions on the map in addition to the speech and thus could be recognised by HMM-based FOVA classifiers.

A preliminary experiment was conducted to see whether these two user behaviours could be distinguished based on their FOVA sequences, by using two FOVA HMM classifiers[27], one for each behaviour.

In a previous study, HMMs were employed in a similar way to determine cognitive strategies in more constrained activities such as equation solving [Sal00].

[26] §3.4
[27] §5.7

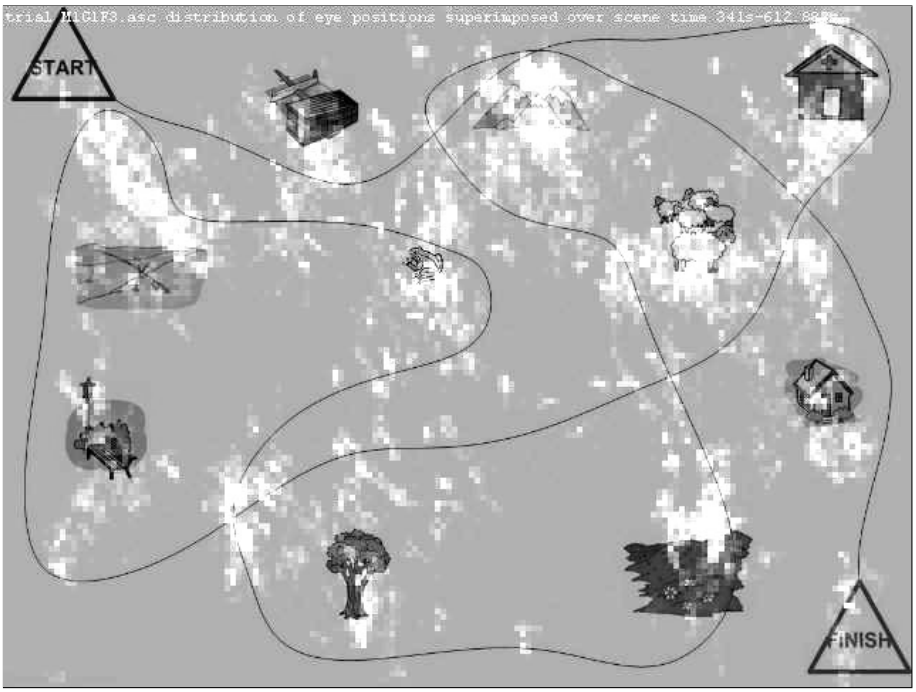


Figure 5.26: The same map as 5.25, but this time showing eye positions while the participant described the route. The difference between the two distributions of gaze direction is used to identify the two types of behaviour using focus of visual attention HMMs

5.10.1 Recognition system

Two sessions from the matched eye and speech data were used to evaluate behaviour recognition for training and testing respectively. Each session had the same participant's eyes tracked, for consistency. Two FOVA HMM's were trained using unsupervised Viterbi reestimation - a HMM for eye behaviour while the user was describing objects on the map, and a corresponding HMM for eye behaviour while the user was describing the route. Segments of speech corresponding to the two behaviours were identified from the transcriptions. Each HMM consisted of 16 states, with each state representing a visual focus. Sixteen states were chosen because this corresponds roughly to the number of objects on the map plus route intersections. This also concurs with the automatic discovery of foci discussed in the previous section[28]. Initially, the visual foci (observation PDFs) were distributed uniformly across the visual scene and state transition probabilities were uniform.

The test data was split into segments of a fixed duration, with segments over-

[28] §5.9

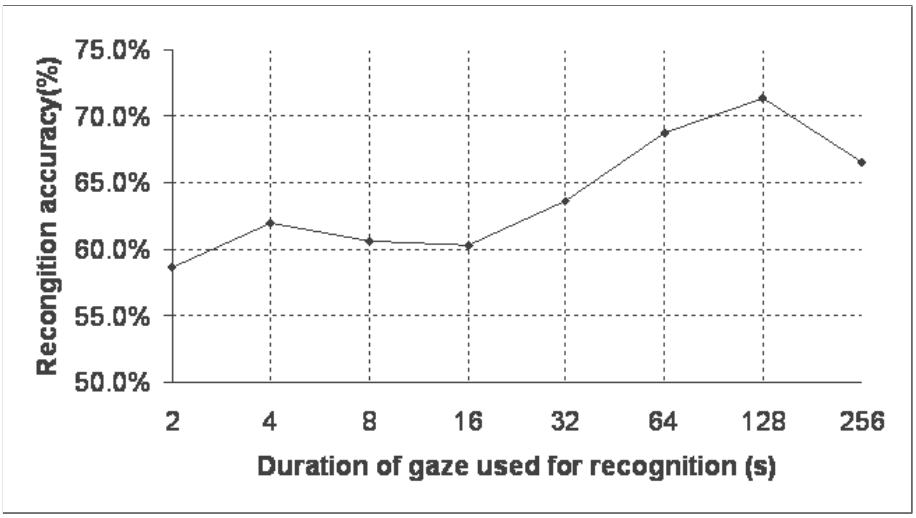


Figure 5.27: Eye movement behaviour recognition showing recognition accuracy as a function of the duration of the segment used in recognition. Two behaviours were identified from speech and recognised from gaze direction, namely describing landmarks and describing the route. Describing landmarks involves greater coupling between gaze direction and speech than describing the route. Longer durations increase accuracy. The fall-off in recognition performance for analysis segment durations greater than $64s$ was expected, since in this case the segment duration approached the total duration of the data

lapping by 50%. Each HMM's likelihood is obtained for each such segment, thus enabling the most likely behaviour to be identified at a period of half the segment duration. The recognition system[29] was tested for various segment durations. Test and training data were swapped and the experiment repeated.

5.10.2 Results

Figure 5.27 shows that recognition generally fell in the region of 60-70%. The system performed optimally using segment durations in the order of $40s$, with shorter segment durations yielding less accurate recognition. This demonstrates that FOVA HMMs may be used to determine user behaviour, although use of this single modality results in error-prone recognition.

The preliminary experiment on behaviour recognition demonstrated some utility in using FOVA HMMs. The discriminatory power of the model set was due to the differences in the models' observation PDFs rather than state transition probabilities. Figure 5.28 shows the distance of the eye to the nearest object over the course of a

[29]The recognition system was implemented in C#.

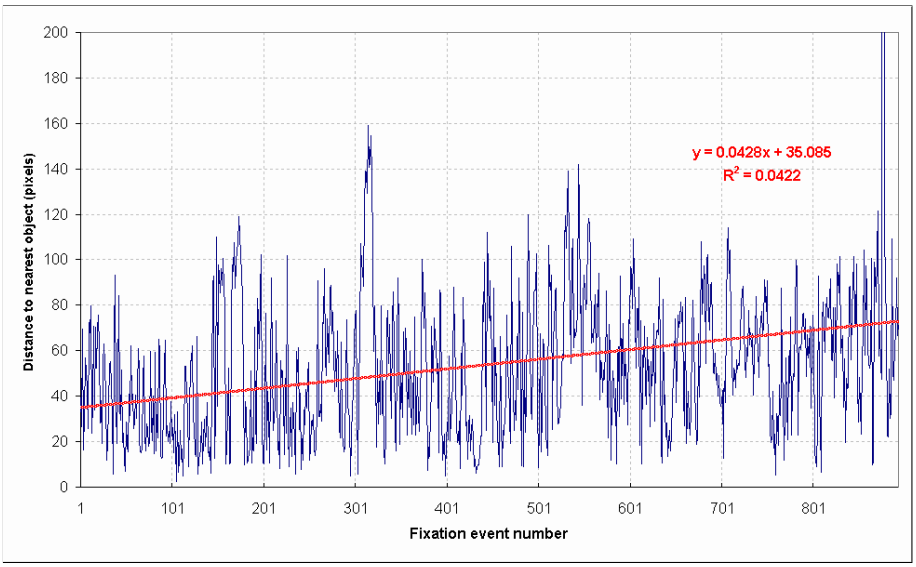


Figure 5.28: Distance of eye fixations to objects in an eye/speech corpus session over time. The trend line indicates distance from objects becomes greater as the session progresses, indicating fixations away from objects towards the route on the map.

session in the eye speech corpus. The linear trend line indicates that, over time, the eye fixates further away from the centre of objects. The most probable explanation for this is that the route was described in the latter half of the session, meaning that there were more fixations away from the objects. Increasing familiarity with the visual scene as the session progresses may also be a contributing factor.

Comparison of state transition probabilities between models showed only minor variation - all exit-state transition probabilities had similar values, suggesting that FOVA sequences exhibit a high degree of randomness. Similar observations from psychology research into scene perception have been made regarding the randomness of FOVA sequences [Ray98]. The possibility of FOVA shifts guided by knowledge-driven gaze control constrained by user behaviour was not demonstrated successfully in this experiment. Further study in this area could concentrate building a behaviour recognition system which also considers the speech in an attempt to improve performance.

5.11 Summary

This objective of this chapter was to determine what information could be extracted from eye movement data for integration into an ASR. HMM and HSMM variants

were investigated as potential information extractors and the novel application of the semi-Markov chain against unpublished eye movement corpora was explored.

The semi-Markov chain in the HSMM made extracting information more robust to noisy eye tracking data compared to utilising the Markov chain in the HMMs, demonstrating the importance of modelling state duration. Eye-movement type classification and FOVA sequence were uncovered. For classifying the FOVA, relaxation of the HSMM observation distributions influence on decoding was necessary to allow the model to treat saccades as random noise. Potential workarounds to make HSMM-based solutions computationally tractable were discussed in §5.8. FOVA sequences characterising user behaviour requires further work including collecting more data in order to draw substantive conclusions.

Eye tracking today has moved towards portable real-time eye tracking outside of controlled laboratory experiments. In this research, only the limited Smart Eye dataset was available as representative data of this type of eye tracking. Further studies of FOVA HMM's should use data that has neither a static visual scene nor user. This would require updating of observation PDFs by tracking visual foci. Target detection, recognition, and tracking, are image-processing topics which could inform updating FOVA HMMs in these cases.

The main motivation behind this chapter was to implement HMM-based formalisations of decoding eye movement. These formalisations are analogous to decoding speech and other modalities and could facilitate development of robust decoding schemes in multi-modal systems. To this end, the integration of eye movement and speech, first discussed in Chapter 2 is investigated in Chapter 6.

6 Integration

This chapter describes the experimental work for integrating eye movement into an ASR system. No studies have attempted integrating eye movement to recognise speech that is continuous, spontaneous, informal and conversational. Previous studies have narrowly focused on human-to-machine dialogues with small vocabularies (E.g. [SH97]).

The first stage in integration was to characterise the relationship between the eye movement and speech in the eye/speech corpus[1]. Next, the baseline ASR system[2] had eye movement information integrated into it[3]. Integration[4] experiments explored the implementation constraints required to assure ASR performance gains.

6.1 Relationship between eye movement and speech

This section verifies the relationships between eye movements and speech published in cognitive science research. It adds to previous work in this field by considering whether the relationships hold for the spontaneous conversational speech between two humans in the eye/speech corpus, rather than more constrained experiments typically involving one person following instructions. Verifying the relationships is an important prerequisite for integrating eye movement into ASR systems.

As discussed in §2.2.2 and §2.2.4, cognitive psychology suggests 3 main relations:

[1] Chapter 3 described the eye/speech corpus.
[2] Chapter 4 described the baseline ASR system.
[3] Chapter 5 described techniques for extracting eye movement information.
[4] Chapter 2 described the theoretical framework for multimodal integration.

1. The proximity of visual attention to a visual focus indicates an increased chance of a spoken utterance relating to that focus.

2. Looking at an object precedes its utterance - the 'eye-voice span' - typically between 430ms [Ray98] and 902ms [GB00]

3. The eye must look at an object long enough for the brain to complete cognitive processing for word production. This cognitive processing time is thus a lower-bound estimation for the 'eye-voice span'.

To verify these relationships, FOVA-level and eye-position-level analyses were carried out on the eye/speech corpus data. The FOVA-level analysis assigned fixations, obtained using the eye tracker's threshold-based detection algorithms, to the nearest visual focus. §5.7 described the basic FOVA-classification method. The eye-position-level analysis considered each sample of eye position.

The FOVA-level analysis of eye movement and speech used fixations as a unit of analysis. From a temporal perspective this draws a better analogy with words from speech, compared to periodic 4ms samples of eye position. Unfortunately, the FOVA-level analysis has two major drawbacks: There is a risk in neither detecting fixations nor assigning them to the correct visual focus (landmark) due to the randomness of eye movement and the close proximity (in relation to the eye tracker's spatial accuracy) of landmarks in the eye/speech corpus maps. To offset these drawbacks, a computationally intensive eye-position-level analysis complemented the FOVA-level analysis, using the eye position samples as a unit of analysis. The number of fixation events is much smaller than the number of samples of eye position because a fixation is made up from a sequence of eye movement position samples - e.g. session M3G2F1 contained 1280 fixation events from 156502 samples of eye position.

6.1.1 Linguistic references to map landmarks

Measuring the relationship between eye movement and speech required the identification of the fragments of participants' speech that referred to map landmarks. There were two different levels of linguistic reference used:

- Keyword detection. Each landmark has a set of words that participants used to describe it. These words were identified from the transcriptions and were listed in §3.4.7 Table 3.3.

- Linguistic Coding. The transcriptions were parsed to identify which landmark was being spoken about. This explained later in §6.1.3.2.

The FOVA-level analysis used the keyword detection level. The eye-position-level analysis used both levels.

6.1.2 FOVA-level analysis

To classify the FOVA, fixations were detected using the eye tracker's inbuilt fixation detection algorithm[5] and assigned to the nearest visual focus (i.e. map landmark) in the visual field using the Euclidean distance.

The eye and speech data were parsed considering the synchronisation issues[6]. Keyword occurrences were detected. Over time this enabled the instruction giver's FOVA sequence to be visualised with the occurrence of both participants naming landmarks. Figures 6.1, 6.2, 6.3 and 6.4 show this visualisation over different temporal windows. The squares and triangles in the figures indicate when the instruction giver and follower respectively named the landmarks (i.e. spoke a keyword), with each of the 12 landmarks referred to by a reference number on the y axis. The line on the graph shows the instruction giver's FOVA.

Figure 6.1 covers the entire 630s session. The graph is crowded but included here because it shows that when participants named a landmark, the instruction giver's FOVA was usually on the same landmark. This provided a subjective verification that visual attention indicates an increased chance of a spoken utterance relating to the attended focus.

Inspection of the other graphs show that the FOVA precedes related speech, i.e. looking at a landmark precedes naming it. Figure 6.2 shows the first 200s of the session during which the instruction giver is describing the landmarks to the instruction follower. From 0-40s the participant looks at all objects in quick succession while comprehending the map. From 40s, the participant becomes familiar with the map and the average FOVA durations lengthen indicating that the landmarks are described in detail (e.g. between 140s and 160s the tree object is being described).

Between 350s-450s (Figure 6.3) the instruction giver was describing the route rather than the landmarks. The route was generally described in terms of its spatial relationship with the nearest landmark (e.g. 'if you follow the route around the

[5] §5.6
[6] §3.3.7

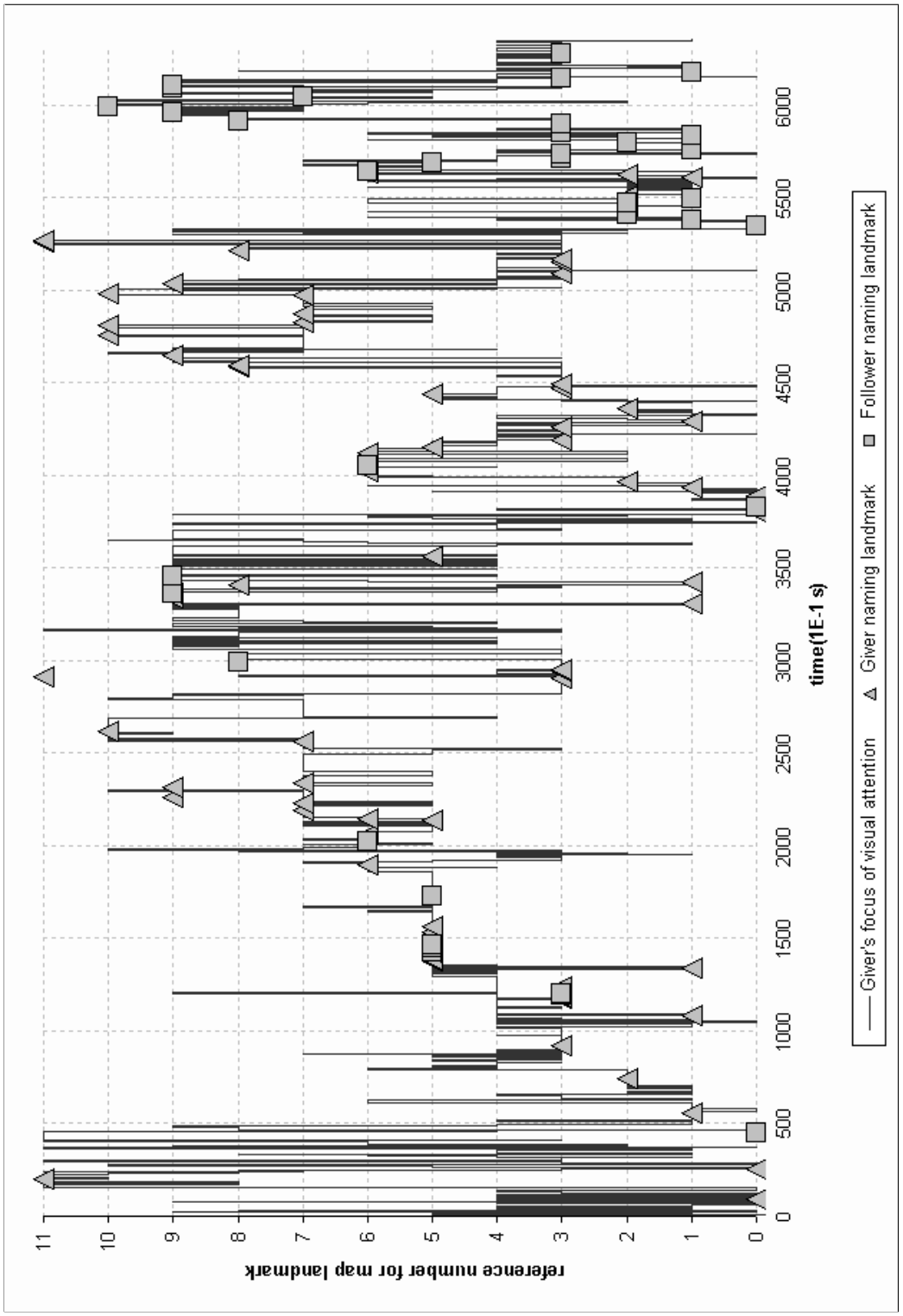


Figure 6.1: Eye/speech corpus session m3g2f1, showing FOVA sequence for the instruction giver. Symbols indicate when participants named landmarks.

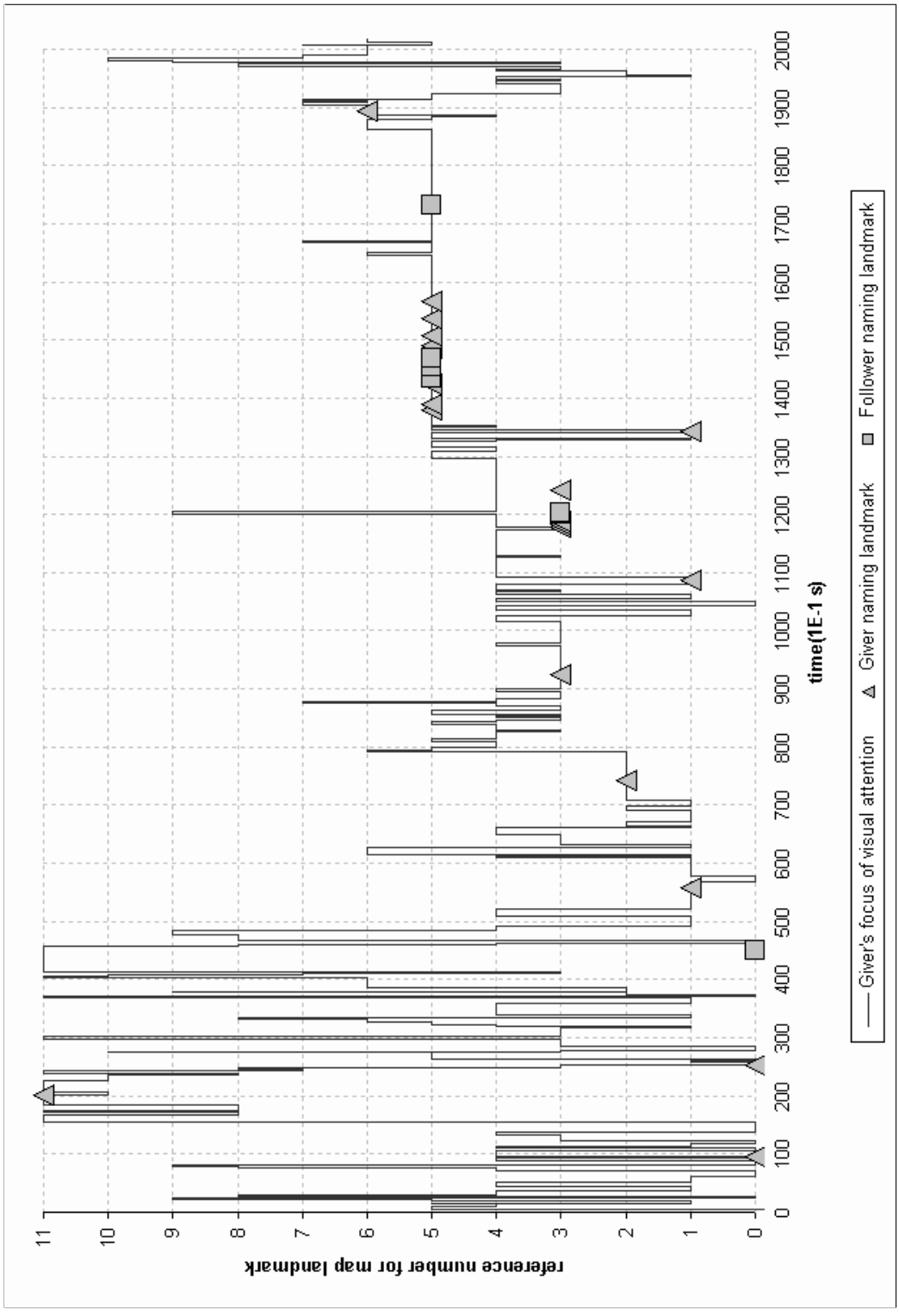


Figure 6.2: The FOVA sequence for the instruction giver with symbols indicating when participants named landmarks for the first 200s of the session.

tree...') and consequently the majority of landmark naming while describing the route coincided with visual attention on the same landmark. Instances where the FOVA and naming the landmark did not correlate (E.g. object reference 5 at 355s) were due to FOVA classification errors that motivated the eye-position-level analysis described in §6.1.3.

Figure 6.4 shows the final 100s of the session. During this time the instruction follower's speech preceded the instruction giver's related FOVA because the instruction follower was verifying their reproduced map - i.e. the instruction follower was guiding the instructions giver's visual attention. The relationship between instruction givers' eye movements and instruction followers' speech is not explored further in this study nor used in the integration experiments. Graphs from the other sessions showed similar results.

To see whether the 'eye-voice span' was evident in the correlation, the eye movement data was time-shifted in relation to the speech in an attempt to increase the correlation. This time-shifting the eye movement is an implementation of the transform function $p^*(W,t)$ described in §2.4.6 and used later in this chapter[7].

All sessions' eye-movement data was time-shifted and the correlation between the modalities measured. The correlation was measured as the ratio of the number of times landmarks were named that corresponded to looking at the landmark, against the total number of times the landmarks were named. The correlation between looking at a landmark and naming it was defined as the period of time from the word onset until either the end of word or a shift in visual attention to another landmark. The experiment was conducted for time shifts between -1.0s to +1.0s, in 50ms increments.

Table 6.1 shows improved correlation between the modalities when delaying the eye movement. The highest correlations all lie within the 300-800ms range. Figure 6.5 shows a typical session's correlation for various time-shifts where the correlation peaked when the eye movement was delayed by 300ms. Further eye movement delays reduced the correlation. Delaying the speech resulted in a fall in correlation, verifying that visual attention precedes related speech, not vice-versa.

The FOVA-level analysis of the eye-speech corpus verified that a participant looking an object increased the chance of him or her naming the object. The 'eye-voice' span was evident, ranging from 300-700ms. The lower bound of this range was less than the lowest time published (460ms [Ray98]). As neither the map nor landmarks

[7] §6.2.

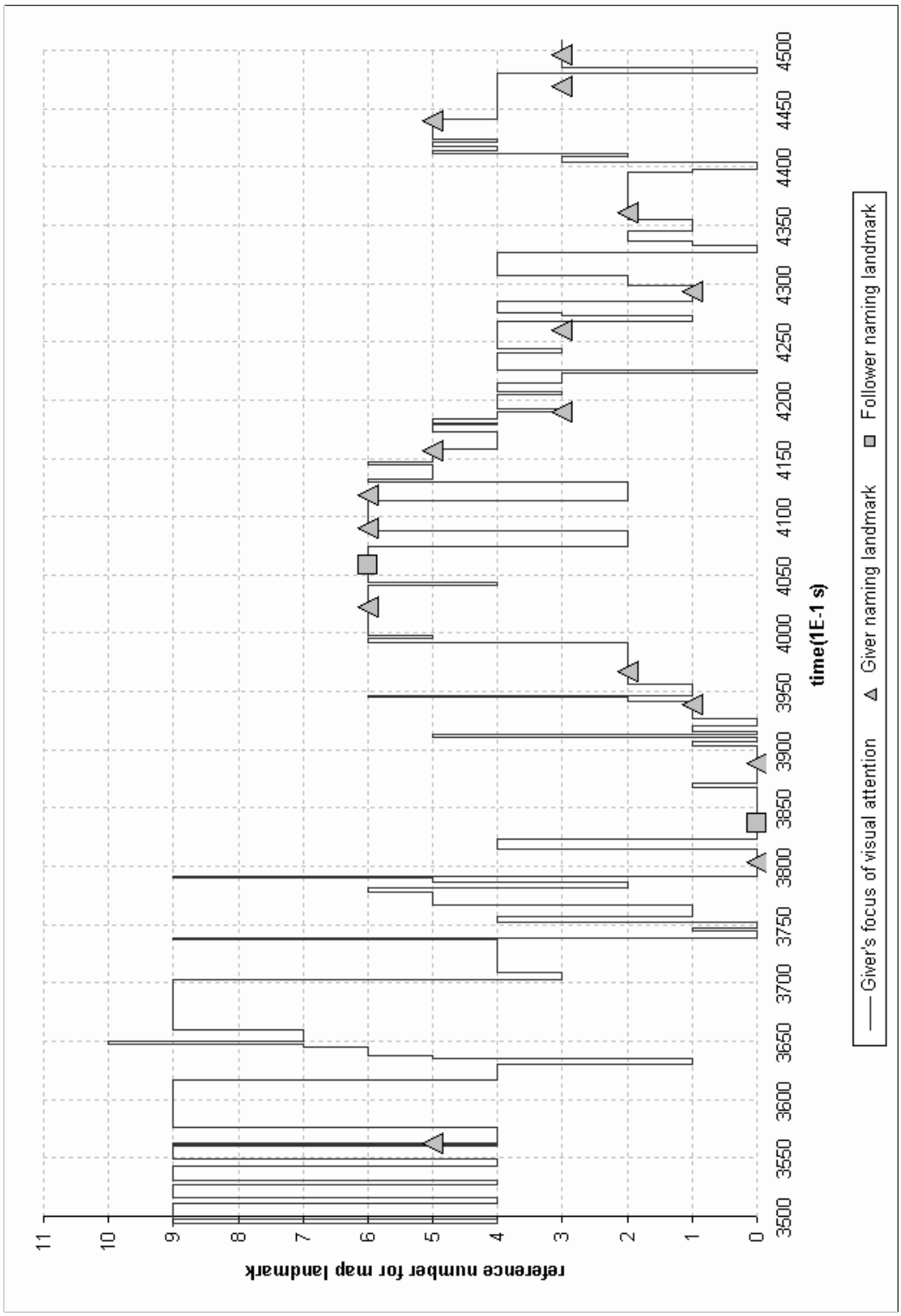


Figure 6.3: The FOVA sequence for the instruction giver with symbols indicating when participants named landmarks during 350s-450s.

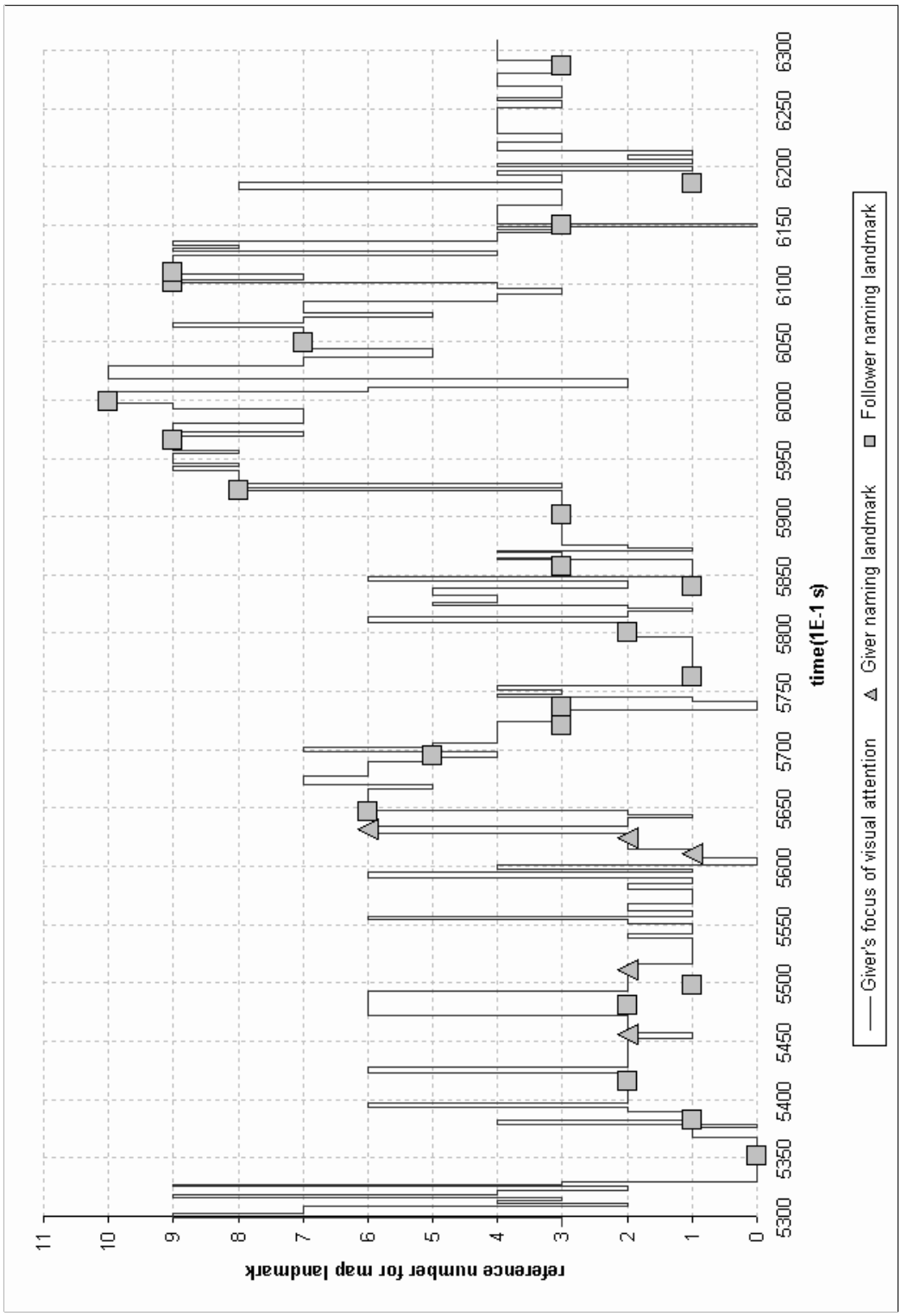


Figure 6.4: The FOVA sequence for the instruction giver with symbols indicating when participants named landmarks during the final 100s.

Session	named landmarks	% FOVA matching named landmark (no time shift)	% FOVA matching named landmark (optimum time shift)	optimum time shift (ms)
m1g1f3	99	57	63	-400
m1g3f2	91	39	45	-700
m1g3f1	103	59	78	-800
m2g1f2	95	32	38	-300
m2g2f3	57	46	51	-300
m3g2f1	79	48	52	-350
m3g2f3	70	33	36	-650

Table 6.1: Effect of shifting eye movement with respect to speech in relation to joint occurrence of the FOVA and naming the landmark.

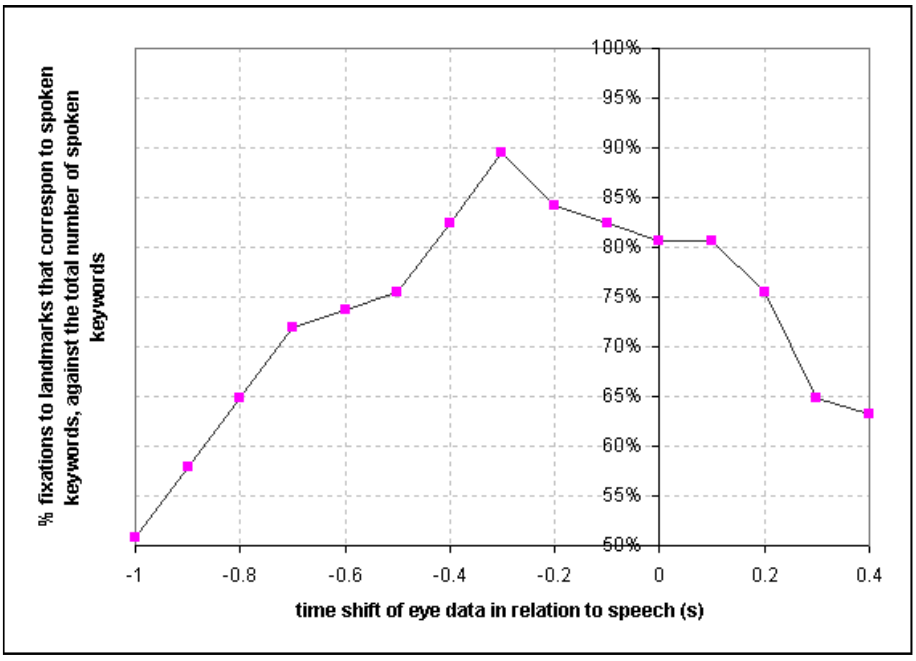


Figure 6.5: The effect of time shifting on correlation for session m2g2f3. A 300ms delay in eye movement data results in the highest correlation.

changed during the session, a likely explanation for the lower times is that the instruction giver had become familiar with the map over the course of the session and was using short-term memory to recall landmarks rather than having to look at them before referring to them.

6.1.3 Eye-position-level analysis

The FOVA-level analysis required the eye movement to be classified as a FOVA sequence. This carried a risk of misclassifying the FOVA due to the eye tracker misdetecting fixations, and the classification algorithm not assigning fixations to the correct visual focus (landmark). These potential errors were due to the spatial and calibration accuracy of the eye tracker[8]. To supplement the FOVA-level analysis, the eye-position-level analysis considered temporal windows of eye positions, requiring extraction of neither FOVA nor fixation sequence from the eye movement data.

The measured distance between each periodic sample of eye position data and each landmark removed the need to classify the eye movement sequence in terms of fixations or FOVAs. Temporal windows (i.e. sequences) of eye positions were considered, with the corresponding average distance to each landmark for each window becoming the unit of analysis. Due to the heavy computational requirements of this approach, only three sessions - m1g3f2, m2g2f3, and m3g2f1, were analysed as a representative subset of all seven sessions.

The analysis used both keyword detection and linguistic coding to identify speech fragments which related to the landmarks. The results of these two methods were compared.

6.1.3.1 Keyword detection

The mean Euclidean distance of the eye position to each map landmark was calculated and normalised against the distance of the eye position to all other landmarks while the user named a landmark (i.e. spoke a keyword). The temporal length of the eye-position window was thus initially equal to the duration of the keyword utterance. The start and end-times for keywords were recovered from the time-aligned transcriptions[9].

The temporal window of eye positions was varied in two ways:

[8] §3.3.6.1
[9] §4.10.1

- *Varying the eye position window length :* The window length of eye positions was varied to include eye positions up to 60s before the keyword utterance, and up to 60s after the utterance.

- *Time-shifting the eye position window :* The original window length of eye position was retained but shifted in time relative to the speech. Time-shifts between -1 and 1s were tried.

For *Varying the eye position window length*, Figure 6.6 shows the mean distance between the instruction giver's eye position relative to the map and the map landmarks, for periods up to 60s before and 60s after a keyword utterance for session m2g2f3. The x-axis indicates the window size over which eye position distance from the landmark was calculated. For example at 0s, the mean position of the eye was calculated for eye positions that occur only while the giver explicitly mentioned the landmark (i.e. spoke a keyword). At -10s on the x-axis the mean position was calculated for positions while the giver mentioned the landmark and all eye positions up to 10s *prior* to mentioning the landmark. Similarly, at +10s the mean position was calculated for positions while the giver explicitly mentioned the landmark, and up to 10s *after*, and so on. The thicker line on the graph indicates the mean distance for all landmarks. The other two sessions in the analysis showed similar results.

The results showed that the eye position was nearer to the landmark when the instruction giver explicitly mentioned the landmark. The bell-shaped curve in the graph shows the average distance for all landmarks suggesting a potential for deriving a normally distributed, probabilistic measure of the correlation between keyword utterance and eye position proximity to the related landmark.

For *Time-shifting the eye position window*, Figure 6.7 shows the mean distance of the eye position relative from the landmarks at various time shifts relative to the time of the keyword utterance for session m2g2f3. The x-axis indicates the time shift in the eye data relative to the speech data. E.g. At 0s the mean position of the eye is determined for eye positions occurring when the participant says a keyword - i.e. no time shift. At -1.0s the eye positions used to calculate the mean occur 1s *before* the participant explicitly mentions the landmark, and so on.

The time corresponding to the smallest mean distance indicates the highest correlation. The eye position is closest to a landmark, on average, 500ms prior to naming it, a similar time to the 'eye-voice span' in published findings. Similar results were obtained for the 2 other sessions.

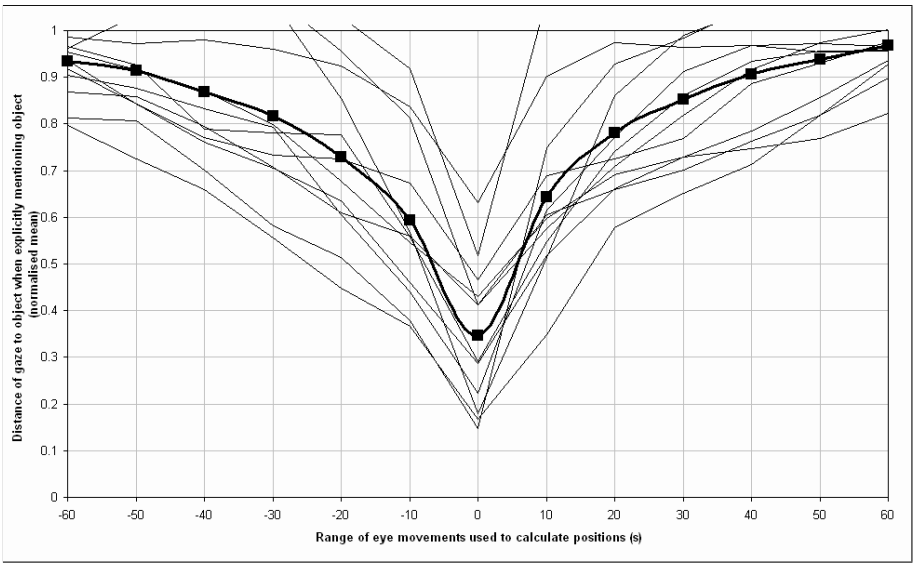


Figure 6.6: Mean distance of eye position relative to map object (landmark) when the subject explicitly mentions that object for session m2g2f3. Each thin line indicates a single object, and the thick line indicates the average of all objects. The x-axis shows mean gaze distance before and after the utterance, indicated by negative and positive values respectively. Correlation between eye direction and speech is evident from reduced distance about the utterance time (0s).

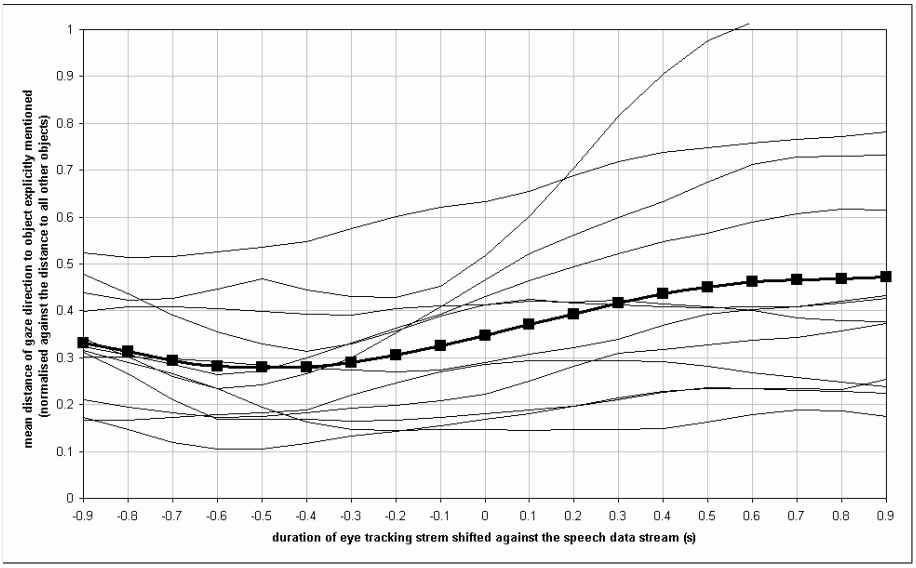


Figure 6.7: Mean distance of eye position relative to an object (landmark) when explicitly mentioning the object for session m2g2f3. Each thin line indicates a single landmark, thick line average of all landmarks. The x-axis shows the time shift the in eye data relative to the speech data. The graph indicates that gaze direction is closest to landmark 500ms on average prior to related speech.

6.1.3.2 Linguistic coding

Measuring the correlation between eye movement and spoken word in the FOVA and eye-position analysis with keyword detection relied on keywords. If the speaker implicitly mentioned the landmark, i.e. by an unspecified noun e.g. 'it' or 'the thing', the utterance was not considered. Likewise, predications to keywords e.g., 'the top left corner is the start...' for the keyword 'start', were not accounted for either. If these unspecified nouns and predications were included, then the correlation between visual attention and the participant talking *about* the landmark could be measured. Linguistic coding was carried out to see whether measuring the correlation between visual attention and the *subject* of speech is the same as measuring the correlation between visual attention and the occurrence of keywords in speech.

To correlate implicit and explicit spoken references to landmarks and any predications against the eye movement data, Professor Antje Meyer from the School of Psychology at the University of Birmingham conducted a linguistic parse of the transcriptions in June 2004. The parse involved coding each transcription fragment by allocating it to particular landmark to indicate the subject of the speech.

For the analysis, there were two types of coding fragments defined - narrow and broad. Narrow coding fragments were defined as the noun (specified or unspecified) that corresponded to a map landmark, together with any preceding adjectives and/or determiners. Disfluencies (i.e. 'eh', 'etc.') that were within the noun-phrase were included.

Broad coding fragments were defined as the noun (specified or unspecified) corresponding to the landmark name, plus predications following or preceding the noun which referred to the landmark. Landmark nouns were the keywords identified in §3.4.7 Table 3.3, e.g. 'frog', 'plane', 'house' etc. Transcription fragments relating to no landmark were left uncoded.

Narrow coding took priority over broad coding for situations where participants referred to multiple landmarks. E.g. For the utterance 'between the start and the mountains is the aeroplane hangar', a broad coding would consider the whole utterance to be talking about the aeroplane hangar. Narrow coding would consider 'the start', 'the mountains', and 'the aeroplane hangar' to be coded separately and assigned to each corresponding landmark. The decision to prioritise narrow over broad coding was based on current research suggesting that lexical search processes rather than thematic structure guide the allocation of visual attention [MvdMB04].

The eye-position-level analysis for linguistic coding was the same as that for key-

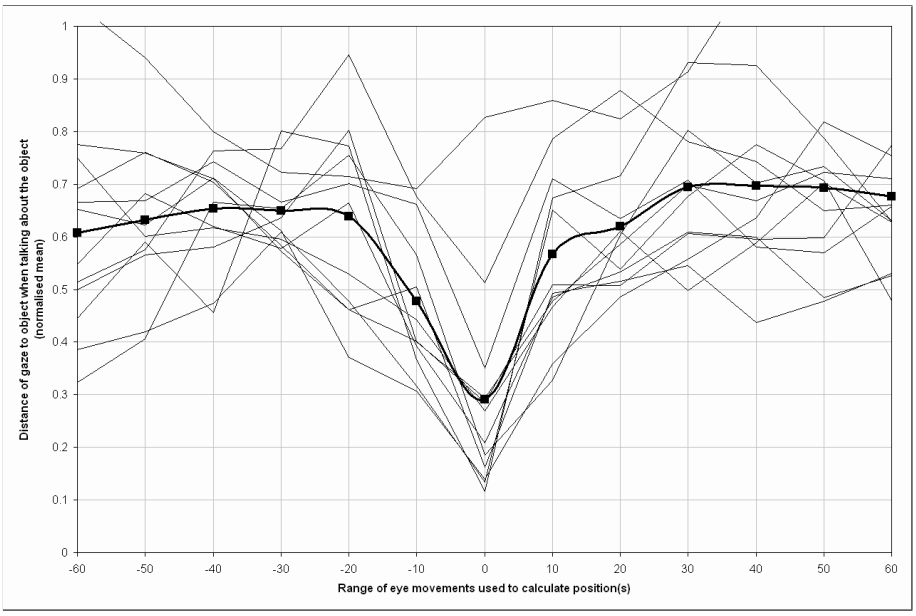


Figure 6.8: Mean distance of eye position relative to object when the subject talks about the landmarks for session m2g2f3. Each thin line indicates a single object, and the thick line indicates the average of all objects. The x-axis shows mean eye position distance before and after the utterance, indicated by negative and positive values respectively. Correlation between eye position and speech is evident from reduced distance.

word detection. Temporal windows of eye positions were varied by changing the window length and time shifting. The results (Figures 6.8 and 6.9) were compared against the equivalent results for keyword detection (Figures 6.6 and 6.7) and similar characteristics were observed. A 300ms shift in eye movement in relation to speech yielded maximum correlation between the eye movement and speech, compared to 500ms for the keyword detection. The linguistic coding results show a better improvement in terms of the normalised mean distance of eye position to a landmark while the user is talking about it - i.e. Figure 6.8 shows a normalised distance of 0.3 at 0s, as opposed to 0.35 in Figure 6.6. The next section discusses the results.

6.1.4 Discussion of correlation analysis

The correlation analysis successfully verified that, using the eye/speech corpus as representative data, the three established relationships between eye movement and speech were observed in spontaneous conversational speech between two humans. To recap, these relationships were:

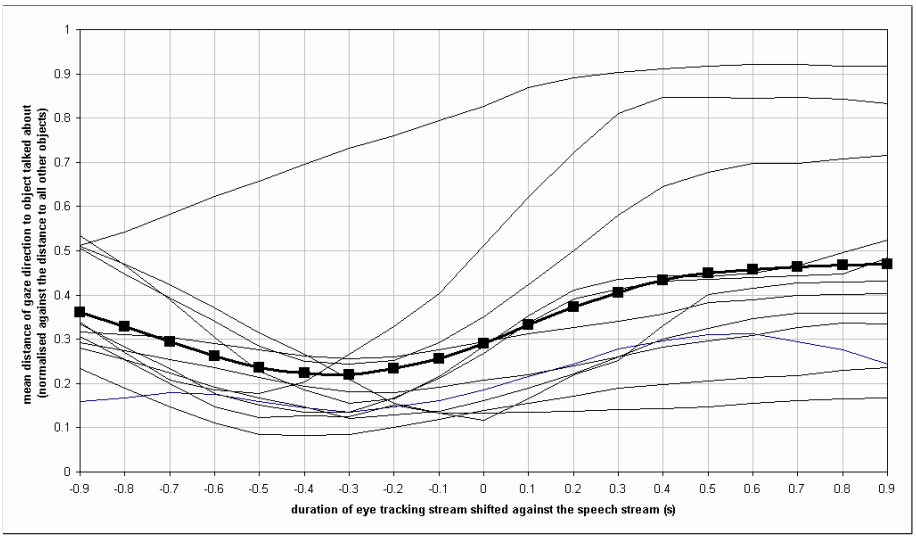


Figure 6.9: Mean distance of gaze relative to landmark when the subject talks about the landmarks for session m2g2f3. Each thin line indicates a single landmark, thick line average of all landmarks. The x-axis shows the time shift the in gaze data relative to the speech data. The graph indicates that eye position is closest to landmark 300ms on average prior to related speech.

1. The proximity of visual attention to a visual focus indicates an increased chance of a spoken utterance relating to that focus.

2. Looking at an object precedes its utterance - the 'eye-voice span' - typically between 430ms [Ray98] and 902ms [GB00]

3. The eye must look at an object long enough for the brain to complete cognitive processing for word production. This cognitive processing time is thus a lower-bound estimation for the 'eye-voice span'.

FOVA-level and eye-position level analyses were undertaken. The FOVA-level analysis only partially demonstrated relationship 1 because the FOVA was assigned deterministically to the closest landmark, and consequently did not address the issue of proximity. Despite this shortcoming, the results shown in Figure 6.1 to Figure 6.4 show a high correlation between the participant's visual attention and their naming the associated landmark.

The eye-position-level analysis demonstrated relationship 1 better than the FOVA-level analysis because it measured the proximity of the eye position to potential foci and removed the risk of FOVA misclassification. The characteristic bell-shaped curve (Figure 6.6) showed that the proximity of a person's eye position to a landmark gives

an indication of the probability of the person naming it. The bell-shaped curve suggests that the PDF of eye position proximity to a visual focus given the occurrence of related speech may be normally distributed.

Both analyses demonstrated relationships 2 and 3. A time delay in eye movement relative to speech of between 0.3s and 0.8s showed improved correlation between the two. These are in line with published findings[10] [Ray98] [MSL98] [MML01] [GS04].

Using keywords to identify speech related to visual foci constrained the analyses to measuring the correlation between naming landmarks and visual attention. The linguistic coding of the transcriptions extended the eye-position-level analysis to examine the relationship between the subject of speech and visual attention. A shorter optimum eye-voice span was observed with the linguistic coding (compare Figure 6.7 with Figure 6.9) and the likely explanation for this is the coding of speech prior to the landmark being named (i.e. predications). Longer eye position window lengths in the linguistic coding is the likely explanation for the lower average normalised distance from eye position to landmark (compare Figure 6.6 with Figure 6.8 at $0s$).

A useful extension to this work would be to use conversational speech and eye movement data other than the eye/speech corpus. For the FOVA-level analysis, using noise robust, probabilistic decoding methods (such as HMMs or HSMMs)[11] to uncover the FOVA is possible but such models were not used because the eye/speech corpus eye data was relatively free from noise. The motivation behind the eye-position-level analysis was to address the shortcomings of the FOVA-level analysis. In considering all eye movement data - including saccadic eye movements, the correlation was measured using eye data that did not indicate visual attention. Considering only eye movement data corresponding to fixations in the eye-position-analysis and taking into account variance as well as the mean distance of the eye position to landmark could improve upon the eye-position-level analysis.

Although these extensions were not persued, the relationship between eye movement and speech for spontaneous conversational speech was adequately verified for the purposes of this study. With eye/speech correlation characterised, this chapter moves on to integration of eye movements into the Baseline ASR system.

[10] §2.2.4
[11] §5.7

6.2 Experiment variables and performance expectations

The integration experiments evaluated implementations of the theoretical multimodal feature-fusion decoding architecture given in Chapter 2. The experiments helped to determine whether the integration of eye movements into an ASR system was beneficial for the task of recognising spontaneous conversational British speech. Eye movements have been previously integrated into ASR systems for constrained, isolated word recognition only [SH97] [ZIGM04].

This section sets the scene for the experiments in §6.3 by introducing the important experiment design variables and setting ASR system performance expectations. Performance measures are discussed in §6.2.1. §6.2.2 summarises the integration theory proposed in Chapter 2. §6.2.3 to §6.2.8 gives the implementation scheme and provides an upper-bound estimation for ASR system performance. §6.2.9 details specific integration variants.

6.2.1 Performance measures

It was suggested in §2.4.7 that the overlap in semantic representations between what is being looked at to what is being said could be defined by associating each landmark with keywords, and that this association forms a modality-independent semantic class which represents a person's intent[12]. In §4.2.1, two ASR performance measures for ASR were defined - the Word Error Rate (WER) and the Figure Of Merit (FOM). FOM was described as a useful complement to WER because it bases its estimate of performance on the ability to detect a subset of the system's vocabulary, e.g. words pertinent to the application.

In the case of a gaze-contingent ASR system, FOM can be used to measure the detection of words associated with visual attention and, accordingly, the recognition of the modality-independent semantic class. FOM therefore serves as the primary measure of interest in the integration experiments to assess integration performance.

6.2.2 Integration framework recap

Before progressing further, it is worth summarising the integration scheme outlined in Chapter 2 so that it may be related to the practical realisation. The joint optimisation

[12] §1.5

problem for decoding the visual attention (v) and word sequence (w) from the eye movement (e) and speech (y) is:

$$p(\widehat{v}, \widehat{w}|e, y) \quad \propto \quad \max_{v,w} p(w, v)p(e|v)p(y|w) \tag{2.18}$$

Where $\widehat{v}$ and $\widehat{w}$ are the most probable visual attention and word sequence respectively. This expression is simplified by assuming that the word sequence is dependent on the visual attention and not vice-versa:

$$p(\widehat{v}, \widehat{w}|e, y) \quad = \quad \max_{w} p(w|\widehat{v})p(y|w) \tag{2.46}$$

Where $\widehat{v}$ is the most probable focus of visual attention sequence:

$$\widehat{v} \quad = \quad \arg\max_{v} p(e|v)p(v) \tag{2.47}$$

The rationale for this - temporal precedence and the relative confidence in decoding each modality, was discussed in §2.4.6.

The probabilities that make up the probability distribution $p(w, v)$ are determined by considering whether the word w^{ς} is part of the keyword set k^{σ}. If it is not then $P(w_t|v_t) = P(w_t)$, i.e. visual attention has no effect on word sequence. If w^{ς} is part of the keyword set k^{σ} then the probability of the word given the visual attention at time t, $P(w_t|v_t)$, is a function, g, of the mutual information between word and visual attention, $i(v_t, w_t)$:

$$P(w_t = w^{\varsigma}|v_t = v^{\sigma}) \quad = \quad \begin{cases} P(w_t = w^{\varsigma}) & \text{if } w^{\varsigma} \notin k^{\sigma} \\ g(i(v_t = v^{\sigma}; w_t = k^{\sigma}), P(w_t = w^{\varsigma})) & \text{if } w^{\varsigma} \in k^{\sigma} \end{cases} \tag{2.48}$$

Where v_t and w_t are the potential FOVA and words respectively at time t. Each value of v^{σ} represents a landmark, each value of w^{ς} is a word and each value of k^{σ} is a set of words (keywords) representing each landmark.

The eye movement is shifted by time τ (the 'eye-voice span') relative to the speech to increase the correlation beween modalities. This makes the mutual information, $i(v_{t+\tau}, w_t)$:

$$i(v_{t+\tau} = v^{\sigma}; w_t = k^{\sigma}) = P(v_{t+\tau} = v^{\sigma}, w_t = k^{\sigma}) \log \frac{P(v_{t+\tau} = v^{\sigma}, w_t = k^{\sigma})}{P(v_{t+\tau} = v^{\sigma})P(w_t = k^{\sigma})} \tag{2.44}$$

Visual attention influences ASR by updating the language model probability distribution $p^*(W)$ during Viterbi decoding:

$$p^*(W,t) \quad = \quad f(p(W),\widehat{v},t) \tag{2.37}$$

Where f is a function of the original language model probability distribution $p(W)$ at time t, and the most probable visual attention sequence $\widehat{v}$.

For the experiments, to avoid non-recoverable error propagation due to visual attention misclassification function f was realised using probabilistic measures of visual attention in two variants[13].

6.2.3 Implementation

As described in the previous section, the integration framework required the updating of the language model probability distribution during Viterbi decoding using the integration function $f(p(W),\widehat{v},t)$. The implementation had two main features:

1. N-Best list rescoring - Rescoring the N-Best lists generated by the Baseline ASR system to change word sequence probabilities in response to the eye movement.

2. Landmark-specific language models - Language models that contained high probabilities for words associated with specific landmarks. They were used to determine the N-Best list rescores.

§6.2.4 and §6.2.6 describe and evaluate these implementation features in order to obtain an upper-bound estimation of the ASR performance improvements that could be expected.

6.2.4 N-Best list rescoring

An ASR system outputs the most likely word sequence for a given utterance using Viterbi decoding. The N-Best algorithm [CS89] extends Viterbi decoding to generate the N most likely word sequences, together with their probabilities. The 'N-Best list' presents these sequences in order of their probability. The N-Best list may be subsequently reordered by rescoring the word sequence probabilities in response to, for example, eye movement information. This is a standard approach in ASR research that obviates the need to implement a customised Viterbi decoder. Such a decoder, in addition to being a major software development effort, could potentially make the direct comparisons between the Baseline ASR system and other research lab systems, given in §4.7.1, unsound.

[13] f_2 and f_3. See §6.2.9 for implementation forms.

In the integration experiments, N-Best list rescoring served as an approximation to Viterbi decoding with dynamic language model probabilities. For the specific word w_t at time t in the word sequence w, of length T, the probability of the word sequence (i.e. entry) in the N-Nest list, $P(w)$, was updated to $P^*(w)$ on a word-by-word basis:

$$P^*(w) = P(w) \prod_{t=1}^{T} \frac{P^*(W_i = w_t | W_j = w_{t-1})}{P(W_i = w_t | W_j = w_{t-1})} \tag{6.1}$$

Where $P(W_i|W_j)$ is the baseline language model probability and $P^*(W_i|W_j)$ is the language model probability determined from the landmark-specific language models described in §6.2.6 and the eye movement. For the integration experiments this computation was carried out in the log domain.

The N-Best lists were generated using HTK [ea02], which relies on a 'token passing' algorithm for Viterbi decoding. The token passing algorithm uses multiple tokens (i.e. forward probabilities) to generate multiple hypotheses. The greater the number of tokens, the more accurate the list[14].

Table 6.2 shows how the number of tokens increased the accuracy (WER) of the baseline ASR performance on eye/speech corpus data. The WER for the N-Best list is defined as the WER for the highest scoring entry (word sequence) in the list. The reason for the increase in acuracy was because there were more accurate entries present in the N-Best list. Increasing the number of tokens beyond 50 proved computationally inefficient for generating large (100+) lists. However, generating large lists with a small number of tokens resulted in inaccurate results - i.e. a 500 N-Best list using 5 tokens was less accurate than a 100-Best list using 20 tokens.

Consequently, as a trade-off between computational efficiency and accuracy, 50 tokens were used to generate 250-Best lists. The baseline ASR was run on the speech component of the eye/speech corpus, generating 250-Best lists for each of the instruction giver's speech segments. The 7 high quality eye movement sessions were used[15]. The 250-Best lists contained word sequence, word times, and word probabilities.

6.2.5 Evaluation of N-Best list rescoring

Rescoring the N-Best list carries the risk that the most probable word sequence that the customised Viterbi decoder would produce is not in the list. If aiming for

[14]More details on token passing algorithm can be found in [ea02].
[15]§3.4.8

Number of tokens	Number of levels in the N-Best List (N)						
	1	10	100	200	300	400	500
5	46.3	34.3	27.3	26.2	25.2	24.9	24.8
10	46.3	33.9	25.2	23.8	23.0	22.3	22.1
20	46.3	33.9	24.6	22.9	x	x	x
50	46.3	33.9	24.3	22.7	x	x	x
100	46.3	33.9	24.3	x	x	x	x

Table 6.2: %WER for eye/speech corpus session m1g1f3 considering all alternatives in the N-Best list with varying list size and number of tokens. 'x' indicates decoding was computationally inefficient. Decoding a session was considered inefficient if not completed within 24 hours on a PC with processing speed in the 3GHz-class. Typical decoding times for tractable sessions were much shorter - no more than 1 hour in most cases. The giver's speech in session m1g1f3 consisted of 223 speech segments and 1324 words.

increased recognition performance, the correct word sequence (i.e. that given in the transcriptions) should be in the list, since rescoring a list so that the correct sequence is the most probable corresponds to 0% WER.

The 250-Best lists generated by the Baseline ASR system when recognising speech in the eye/speech corpus data were analysed to determine their potential to supply the correct word sequences. The probability for the correct word sequence for each utterance was compared against all word sequence probabilities in the N-Best list for that utterance. These probabilities were deduced by running the system with speech segment-specific language models[16] for each utterance.

The N-Best list was deemed suitable for rescoring if the correct word sequence was present within it. If the correct word sequence also contained any keywords associated with landmarks, then this indicated the potential to improve speech recognition performance by rescoring the N-Best list when integrating eye movements. The N-Best list was judged as having an unrealisable potential for performance improvement by rescoring if any word sequence above the correct word sequence in the N-Best list also contained any keyword associated with landmarks. This potential was considered unrealisable because integrating eye movement would not move the correct sequence in the N-Best list above word sequences that contained the same keywords which were above the correct sequence in the N-Best list prior to rescoring.

Table 6.3 shows the potential benefit to be gained by integrating eye movements

[16]Speech segment-specific language models were previously described in §4.10.1 when they were used to generate time-aligned transcriptions for the eye/speech corpus.

		Segment counts				
		Location in N-Best List			Integration Potential	
Session	no. of segments	1st	2nd to 250th	Not in list	Unrealisable	Realisable
m1g1f3	223	39	96	88	28	9
m1g3f1	199	54	68	77	22	5
m1g3f2	197	35	69	93	18	7
m2g1f2	302	49	106	147	15	3
m2g2f3	108	29	44	35	14	1
m3g2f1	150	27	50	73	11	2
m3g2f3	166	38	59	69	19	0
μ	192	20.9	36.7	42.4	9.9	1.9
σ	61	4.7	3.7	6.1	2.9	1.5

Table 6.3: N-Best list potential for rescoring the correct result due to integrating eye movements.

and speech into ASR for the eye/speech corpus. Only 27 utterance segments could be rescored so that the correct word sequence became the most probable, out of 1345 utterance segments. This signifies an approximate 2% potential decrease in WER when integrating eye movements compared to using the baseline ASR system.

The N-Best list evaluation showed that the hypothetical decrease in WER is small. This decrease is unlikely to be realised by integrating eye movements because, as shown in §6.1, the correlation between the eye position and landmark naming is not 100%. This supports the assertion that WER is not the best indication of performance improvement for integrating eye movements into ASR[17]. Keyword-spotting based measures (E.g. FOM) are more appropriate because they indicate the performance of the system to recognise the modality-independant semantic class. In this respect, §6.2.6 considers keyword-spotting measures in the evaluation of the second feature of the implementation: Landmark-specific language models.

6.2.6 Landmark-specific language models

To integrate eye movements into ASR, the N-Best list was rescored by changing the word probabilities. To determine the rescored value of word probabilities, a separate language model for each landmark was generated by modifying the baseline back-off bigram language model. These 'landmark-specific language models' had higher probabilities for words associated with the landmarks.

[17] §6.2.1

Recalling §4.11, the baseline bigram language model was constructed using frequentist estimates of bigrams based on their occurrence in the BNC [Bur00] and HCRC map task corpora [ABB+91]. Back-off weights [Kat87] were used for the robust estimation of unseen bigrams.

The modification of the Baseline language model to create a landmark-specific language model was realised by shifting probability mass away from unigrams and bigrams that did not involve landmark keywords, towards unigrams and bigrams that did.

Let W^{from} be the set of words in the language model to shift probability mass from, and W^{to} to be the set of words (i.e. the keywords associated with map landmarks) in the language model to which the shifted probability mass is redistributed. Let m be the proportion of mass shifted from each word. New unigram probabilities, $P_{shift}(W_i)$, were calculated from $P(W_i)$:

$$P_{shift}(W_i) = \begin{cases} (1-m)P(W_i) & \text{if } W_i \in W^{from} \\ P(W_i) + m\frac{P(W_i)}{\sum_{W_t \in W^{to}} P(W_t)} \sum_{W_f \in W^{from}} P(W_f) & \text{if } W_i \in W^{to} \end{cases} \tag{6.2}$$

Likewise, for the bigram probabilities $P_{shift}(W_i|W_j)$:

$$P_{shift}(W_i|W_j) = \begin{cases} (1-m)P(W_i|W_j) & \text{if } W_i \in W^{from} \\ P(W_i|W_j) + m\frac{P(W_i|W_j)}{\sum_{W_t \in W^{to}} P(W_t|W_j)} \sum_{W_f \in W^{from}} P(W_f|W_j) & \text{if } W_i \in W^{to} \end{cases} \tag{6.3}$$

These are seen bigrams in the baseline language model, not those estimated using back-off weights and unigrams.

Two variants for shifting probability mass were used. One shifted probability mass from all other words regardless of their association with landmarks. The results of applying this method will be hereafter referred to as the 'all-mass' language model. The second variant shifted probability mass away only from words associated with the other landmarks on the map. The result of applying this method from now on will be referred to as the 'compete-mass' language model.

The motivation behind using two variants for landmark-specific language models was to see whether using a gaze-contigent ASR that shifted probability mass away from words that were associated with other landmarks but not the common, often

short length words used in conversational speech (e.g. 'it', 'and' and so-forth), would increase the recognition of the latter.

For both variants, 99% of the available probability mass was shifted to words associated with the landmark and redistributed according to the relative language model probabilities of the receiving words. Not all (100%) of the probability mass was shifted in order to avoid a shrunken and unrealistic vocabulary.

6.2.7 Evaluation of landmark-specific language models

The landmark-specific language models were evaluated to see what ASR performance differences the 'all-mass' and 'compete-mass' variants could yield. 'All-mass' and 'compete-mass' landmark-specific ASR systems were built for each landmark, using the associated language model and the same acoustic model as the Baseline ASR system. Each landmark-specific ASR system was used to recognise all of the Instruction giver's speech in the eye/speech corpus. This included the speech data from the sessions rejected for bad eye movement, since this evaluation did not utilise the eye movement data.

The landmark-specific ASR systems' performances were compared against the baseline ASR system. Performance measures used were WER and keyword-spotting performance in terms of the True Positives (TP) and False Alarms (FA) counts. TP and FA are the two component measures of FOM. The counts indicate the number of times keywords are correctly and falsely detected respectively[18]. The performance of each landmark-specific ASR system was expected to be worse than the baseline ASR system due to it being used to recognise all of the speech in the eye/speech corpus session rather than only those segments corresponding to when the speaker looked at the specific landmark. For this reason, the FOM measure was not of interest because the FA count was expected to increase significantly, resulting in a low FOM. However, it was also anticipated that the landmark specific ASR systems would better recognise words corresponding to their landmarks and show an increase in TP. The TP count was therefore of interest rather than FOM in demonstrating the landmark-specific ASR systems' ability to improve keyword detection.

Table 6.4 shows the performance of the landmark-specific ASR systems for all sessions in eye/speech corpus map set 2 (a total a 6 sessions from 3 speakers), against the baseline ASR system. The average WER in the 'all-mass' landmark-specific ASR systems was higher (55.4%) than the baseline ASR system's WER of 41.7%. The

[18] §4.2.1

Obj. no.	no. keywords	Baseline ASR WER=41.77%		'all-mass' ASR systems			'compete-mass' ASR systems		
		%TP	%FA	%WER	%TP	%FA	%WER	%TP	%FA
0	36	77.8	16.7	55.30	94.4	1111.1	41.05	88.9	22.2
1	41	56.1	9.8	53.99	87.8	487.8	40.85	82.9	24.4
2	46	76.1	17.4	54.61	82.6	473.9	41.32	80.4	15.2
3	37	86.5	10.8	55.62	89.2	1394.6	40.98	94.6	27.0
4	49	83.7	10.2	55.80	91.8	1036.7	41.17	87.8	12.2
5	32	78.1	37.5	55.83	84.4	1743.8	41.33	81.3	90.6
6	55	83.6	20.0	55.76	92.7	820.0	41.34	87.3	36.4
7	31	74.2	32.3	55.48	77.4	1529.0	41.22	77.4	51.6
8	20	75.0	15.0	55.55	90.0	2160.0	41.27	85.0	40.0
9	41	68.3	26.8	56.89	78.0	1673.2	41.42	73.2	43.9
10	91	75.8	20.9	55.64	79.1	536.3	40.89	76.9	34.1
11	15	86.7	20.0	54.20	93.3	920.0	41.29	93.3	40.0
μ	41.2	76.8	19.8	55.4	86.7	1157.2	41.2	84.1	36.5
σ	19.4	8.5	8.7	0.8	6.2	547.5	0.2	6.6	20.7

Table 6.4: The performance of landmark-specific ASR systems against the baseline ASR system for all sessions in eye/speech corpus map set 2. Each row represents a landmark. Column 'no. keywords' shows the number of times words associated with each landmark were spoken. The next 2 columns give the Baseline ASR keyword spotting performance in terms of %TP and %FA. The 'all-mass' and 'compete-mass' systems' performance make up the remaining columns. The landmark-specific ASR system performance results on each row are those from the associated landmark-specific ASR systems.

WER in the 'compete-mass' landmark-specific ASR system was similar (41.2%) to the baseline ASR system's WER.

In terms of detecting keywords, landmark-specific ASR systems improved over the baseline models' ASR system performance - i.e. The average percentage of keywords spotted correctly (TP count) improved from 76.8% to 86.7% and 84.1% for the 'all-mass' and 'compete-mass' variants respectively. The FA count however also increased, significantly in the case if the 'all-mass' model but less so, for the 'compete-mass' model - i.e. The average percentage rise in the FA count rose from 19.8% to 1157.2% and 36.5% for the 'all-mass' and 'compete-mass' variants respectively. Other map sets yielded similar results.

These results lowered expectations of large performance gains over the baseline. Although the 'all-mass' variants increased the detection of keywords, they had a high false alarm rate and consequently a high WER. The 'compete-mass' showed more

promise - there was no rise in WER and keywords detection rates and false alarms increased by similar amounts.

The baseline ASR system keyword spotting performance was high - i.e. a TP of 76.8%. A plausible reason for this is there is less chance of the keywords (landmark names) being transcribed in error, compared to common words associated with disfluencies. It was also observed that words associated with the visual field and related to the task were less likely to be associated with disfluencies, as the participants take more care in their pronunciation of these than of common-use words.

The landmark-specific language model ASR system keyword spotting performance was expected to improve with language models biased towards those keywords; however, the higher FA rates for keywords were due to the landmark-specific models not being used selectively by integration of the eye movement. The aim of the eye movement integration experiments is to reduce the increase in FA but to realise the gain in TP observed in this evaluation.

To summarise, this evaluation shows that the 'compete-mass' language model approach for rescoring yields a potential performance improvement, as an ASR system built using the language model recognised 7.3% more keywords compared to the baseline, whilst maintaining a similar WER due to only a small increase (16.5%) in false alarms. The 'all-mass' language model recognised 9.9% more keywords compared to the baseline but high false alarm rates pushed up WER by 14.2%.

6.2.8 Previous work

The landmark-specific language models are specific examples of context-specific language models, involving shifting probability mass due a defined context. The contextual cue in a gaze-contingent ASR system is the focus of visual attention from eye movement. A previous study using this contextual cue achieves a 7% increase in the number of city names recognised by a gaze-contingent ASR whilst a user issued commands to a journey planning system that displayed a geographical map to the user [SH97]. In more heavily constrained conditions, acoustically similar words (Japanese words for various colours) were correctly recognised by a gaze-contingent ASR. In both these systems the speech was command-driven rather than the spontaneous, conversational speech used in this study.

Other contextual cues for redistributing probability mass in language models have been presented. Rather than use eye tracking, the focus of visual attention has been inferred from previous words in a word sequence with weight given to the most

recent. In [RM05], as part of a spoken language understanding system ('fuse'), language model probability mass was redistributed towards sets of words that referred to objects (i.e. colour, shape, position) in a visual scene using previous words in the word sequence together with knowledge of the visual scene (i.e. image segmentation based on colour). In [CR97] [CR98] the British National Corpus[19] was used to build a bigram language model similar to the one used in the baseline ASR system in this study. The authors attempted to reduce the perplexity[20] of word sequences by redistributing probability mass according to words which had occurred in recent speech.

6.2.9 Forms of the integration function f

The aim of the integration experiments was to determine which methods for integrating eye movement can achieve ASR system performance improvement. In §6.2.5 and §6.2.7 the upper-bound estimation of increasing FOM (by increasing TP and reducing FA counts) whilst maintaining WER was set for the ASR performance gain achievable. To realise the performance improvement, the most probable FOVA sequence, $\widehat{v}$, must supply all of the information necessary to selectively use the landmark-specific language models in the N-Best list rescoring process.

To recap, integration of eye movement into the ASR is via the integration function f, which modifies the language model probability distribution $p(W)$ to $p^*(W)$ at the onset of each word depending on the most probable FOVA at that time (t):

$$p^*(W,t) \quad = \quad f(p(W),\widehat{v},t) \tag{2.37}$$

Function f was implemented in three forms - f_1, f_2 and f_3:

1. $f_1(p(W),\widehat{v},t)$- *Deterministic FOVA assignment with selection of a landmark-specific language model.*

2. $f_2(p(W),\widehat{v},t)$- *Weighted FOVA assignment with combination of the landmark-specific language models.*

3. $f_3(p(W),\widehat{v},t)$- *Probabilistic FOVA assignment with combination of a landmark-specific language model and the baseline language model.*

[19] §4.3.2
[20] §4.2.2

The different forms enabled the following questions to be answered, concerning increasing ASR system performance:

- Should only the focus being viewed be considered during integration (f_1) or should all foci be considered (f_2)?
- Is there any benefit in using the spatial proximity of visual attention to a visual focus to indicate the increased chance of a spoken utterance relating to the visual focus (f_3)?
- Given than an increase in the correlation between the modalities was empirically demonstrated in §6.1, does temporally shifting the eye movement stream in relation to the speech to account for the eye-voice span lead to improved ASR performance using either f_1, f_2, or f_3 ?

For $f_1(p(W), \widehat{v}, t)$- *Deterministic FOVA assignment with selection of a landmark-specific language model*, the closest landmark, σ to the visual attention that temporally corresponded to the word onset at time t was selected for rescoring:

$$\sigma = \arg\min_{\varsigma} D(v_{t+\tau}, v^{\varsigma}) \tag{6.4}$$

Where $D(v_{t+\tau}, v^{\varsigma})$ is the Euclidean distance between eye position and landmark ς for the visual attention $v_{t+\tau}$ at time $t+\tau$. The associated landmark-specific language model, $p_{\sigma}(W)$, was used to rescore the word:

$$f_1(p(W), \widehat{v}), t) = p_{\sigma}(W_t) \tag{6.5}$$

Note that the function f_1 allowed for the eye movement stream to be temporally shifted by time τ in relation to the speech in order to increase the correlation between modalities and reduce the risk of FOVA misclassification due to the eye-voice span. The function is a deterministic assignment and FOVA misclassification would result in an incorrect selection of landmark-specific language model. There are three further possible causes for FOVA misclassification: eye tracker accuracy; the close proximity of landmarks and the use of saccadic eye movement to select the language model if the word onset corresponds to the participant making a saccade.

The f_2 function, $f_2(p(W), \widehat{v}, t)$- *Weighted FOVA assignment with combination of the landmark-specific language models*, aimed to address the risk of FOVA misclassification in f_1. Rather than use only one landmark-specific language model, a

combined language model was produced from all landmark-specific language models using weighted interpolation:

$$f_2(p(W), \widehat{v}, t) = \sum_{\sigma} c_\sigma p_\sigma(W) \tag{6.6}$$

Where c_σ is the weight given to the landmark σ's language model, $p_\sigma(W)$. The weights corresponded to the relative Euclidean distance of the eye position to each landmark at time $t + \tau$ normalised against the sum of distances to all landmarks:

$$c_\sigma = \frac{D_\sigma(v_{t+\tau}, v^\sigma)}{\sum_{\epsilon=1}^{E} D_\epsilon(v_{t+\tau}, v^\epsilon)} \tag{6.7}$$

Where E is the total number of landmarks.

The functions f_1 and f_2 carry the assumption that because a landmark is being viewed, the associated landmark-specific language model, or combination of models, should be used in the ASR system. The presupposition discussed in §1.5 stated that 'the semantic relationship between what is said and what is looked at exists regardless of whether or not the speech turns out, in practice, to be related to the visual focus.'. Clearly, if speech turns out not to be related to the landmark then the baseline language model should be used in preference to any landmark-specific variety. Furthermore, a probabilistic measure which related the temporal and spatial proximity of the eye position to a landmark to an increased chance of utterance relating to the landmark was empirically demonstrated in §6.1.3.

To support this presupposition, the third function, $f_3(p(W), \widehat{v}, t)$- *Probabilistic FOVA assignment with combination of a landmark-specific language model and the baseline language model*, aims to address the preference of using the baseline language model over the landmark-specific variants by using the spatial proximity of the visual attention to the nearest landmark to measure the probability that the speech is related to the landmark. With this information, the baseline language model and a landmark-specific language model are combined using weighted interpolation:

$$f_3(p(W), \widehat{v}, t) = c_\sigma p_\sigma(W) + (1 - c_\sigma)p(W) \tag{6.8}$$

Where the weight c_σ is the same as that given for f_2 in expression 6.7. σ corresponds to the nearest landmark to the visual attention, and is the same as that given for f_1 in expression 6.4.

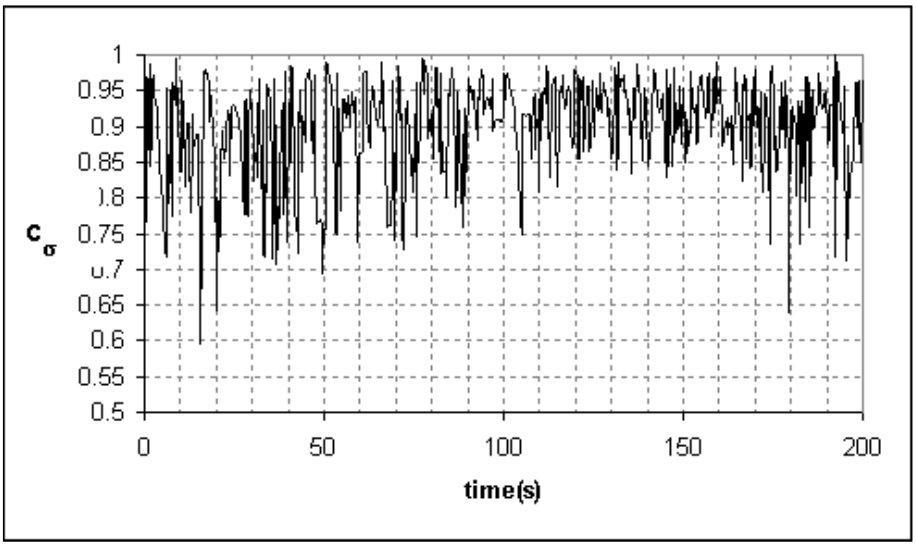


Figure 6.10: The variation in the probabilistic value of confidence in FOVA assignments for session m1g1f3. The spatial proximity of FOVAs to a landmark is weighted against the spatial proximity of the FOVAs to all other landmarks.

c_σ served as a probabilistic measure of the confidence in the FOVA assignment. This probabilistic expression assumed that if all landmarks were equidistant to the FOVA position, then the probability of a FOVA being assigned to each landmark would be $1/E$. The expression thus tends towards 1 as the relative distance of the FOVA position to a focus gets smaller. Another way of understanding f_3 is that it enables the ASR system back-off towards the baseline language model from the landmark-specific language model the further the FOVA is from the closest landmark. To demonstrate what this means in practice, Figure 6.10 shows how c_σ varied over an entire session in the eye/speech corpus. The weights are seen to vary between values of 0.6 and 1.

With the implementation of these three functions, some notable assumptions were made. The use of the relative Euclidean distance for calculating the interpolation weights for the language models in f_2 and f_3 serves as demonstrative of the approach rather than an optimised method. In f_3, the assumption that the spatial proximity of visual attention to a focus leads to an increased chance of naming the landmark was justified from the correlation experiment in §6.1.3, which showed the distance of visual attention to a landmark decreased, on average, if the participant talked about the landmark. The bell-shaped curve in Figure 6.6 suggested a probabilistic measure for proximity, but it should be noted that proximity in Figure 6.6 is temporal, rather than the spatial proximity that is assumed for f_3.

6.2.10 Summary of the implementation approach

The implementation approach of rescoring N-Best lists using landmark-specific language models was evaluated and described in sections 6.2.4 to §6.2.7. Improvements in keyword spotting accuracy using compete-mass landmark-specific models showed the greatest promise. Expectations for the integration experiments showed that the use of the 'all-mass' language models could produce higher WER than the baseline and that the 'compete-mass' language model use could produce similar WER to the baseline. Both variants should detect more keywords - up to 10% more, however false alarms may mitigate any overall improvement, leading to a lower FOM and higher WER. The aim in ASR system performance impovement therefore was to maintain WER and increase FOM.

§6.2.9 proposed three integration functions to define how the eye movement would be incorporated into an ASR system in order to give the desired improvement in performance. Comparing the effect of these functions on ASR performance would also assist in answering specific questions regarding successful implementation.

6.3 Experiments

The design variables defined were:

- The size of the N-Best list. This was fixed for all experiments at 250.
- The use of either the 'all-mass' or the 'compete-mass' landmark-specific language models.
- The integration function used (f_1, f_2 or f_3).
- The shift in eye movement in relation to speech (τ).

These design variables realise 12 integration experiments, which are summarised in Table 6.5. The shift in eye movement in relation to speech was either no shift or 500ms, the latter being an empirical estimation of the average duration of the eye-voice span[21]. Experiments were grouped into pairs with each pair's integration differing only by the time shift.

A 250-Best list was generated for all of the Instruction giver's speech segments. Each speech segment typically contained several words. For each integration experiment, each word sequence in the 250-Best lists was rescored using the chosen

[21] §6.1

Experiment Number	Language model		Integration Function			τ
	'all-mass'	'compete-mass'	f_1	f_2	f_3	(ms)
1	✓		✓			0
2	✓		✓			500
3	✓			✓		0
4	✓			✓		500
5		✓	✓			0
6		✓	✓			500
7		✓		✓		0
8		✓		✓		500
9	✓				✓	0
10	✓				✓	500
11		✓			✓	0
12		✓			✓	500

Table 6.5: The twelve integration experiments' design variables for language model use, integration function, and the time shift between modalities. The N-Best list length was fixed at 250 for all experiments.

integration function and the WER and FOM were measured and compared to the baseline. Each integration experiment was performed on the seven eye/speech corpus sessions that had high-quality eye data. §6.3.1 to §6.3.6 give the experiment results and further details. §6.4 collates the results for all experiments with analysis.

6.3.1 Experiments 1 and 2: Deterministic FOVA assignment with selection of an 'all-mass' landmark-specific language model

For the first two experiments, integration was performed using function f_1, by selecting a single landmark-specific language model at each word onset according to the FOVA at the word onset time. New word sequence probabilities were calculated from the original word sequence probabilities using the ratio of the baseline language model's bigram probability to the 'all-mass' landmark-specific language model probability. Experiment 2 repeated experiment 1 but shifted the eye movement 500ms prior to the onset of the word.

Table 6.6 shows the recognition results for experiments 1 and 2 in terms of WER and FOM. The results showed an increase in WER for all sessions compared to the baseline, with FOM compared to the baseline increasing in some sessions and decreasing in others.

Session	% WER			% FOM		
	Baseline	Rescored no shift	Rescored 500ms shift	Baseline	Rescored no shift	Rescored 500ms shift
m1g1f3	46.3	52.1	52.3	57.0	65.2	55.7
m1g3f1	45.2	47.8	48.0	70.7	55.8	60.9
m1g3f2	52.5	52.1	52.3	46.6	50.0	50.0
m2g1f2	64.7	69.4	69.2	35.9	32.5	43.4
m2g2f3	43.3	47.9	47.5	57.9	84.2	26.3
m3g2f1	54.5	54.1	54.2	64.3	55.1	54.7
m3g2f3	56.5	59.7	59.5	65.1	63.2	29.7
μ	51.8	54.7	54.7	56.8	56.7	53.0
σ	7.5	7.6	7.6	12.0	16.9	18.0

Table 6.6: Integration experiments 1 and 2 compared to baseline. 'no shift' refers to experiment 1, '500ms shift' refers to experiment 2.

For experiment 1, session m1g1f3 showed an increase in WER of 5.8% and an increase in FOM of 8.2%, showing that while overall recognition performance decreased when adding eye movement, the keyword spotting accuracy for words associated with visual attention increased. Sessions m1g3f2 and m2g2f3 showed similar results. Session m1g3f1 showed a 2.6% increase in WER and a 14.9% fall in FOM, showing that whilst integrating the eye movement helped to recognise keywords, it increased the number of keywords falsely recognised (i.e. false alarms), cancelling out the benefit of integrating eye movements. Sessions m2g1f2, m3g2f1, and m3g2f3 showed similar results. Overall (using the sample mean and standard error), WER increased by 5.8% and the FOM remained the same.

For experiment 2, shifting the eye movement by 500ms in relation to the speech resulted in minor changes in the WER compared to the no-shift case. The FOM however varied considerably. For example, session m1g1f2 the FOM decreased by 1.4% compared to the baseline, whereas in experiment 1 the FOM increased by 8.2%. In session m2g2f3 the FOM decreased by 31.6%, whereas in experiment 2 it had shown an increase of 26.3%. Overall, WER increased by 5.8% and the FOM decreased by 4.1%.

Considering both experiments, it can be concluded that using $f_1(p(W), \widehat{v}, t)$- *Deterministic FOVA assignment with selection of a landmark-specific language model*, with the 'all-mass' language model variants resulted in an increase in WER and fluctuations in FOM. The fluctuation in FOM across sessions was due to the high false recognition of keywords, echoing the evaluation of landmark-specific language model

Session	% WER			% FOM		
	Baseline	Rescored no shift	Rescored 500ms shift	Baseline	Rescored no shift	Rescored 500ms shift
m1g1f3	46.3	56.7	56.7	57.0	56.6	56.6
m1g3f1	45.2	51.7	51.5	70.7	72.7	66.8
m1g3f2	52.5	54.9	54.4	46.6	52.0	52.7
m2g1f2	64.7	73.2	73.3	35.9	32.8	30.7
m2g2f3	43.3	52.3	52.3	57.9	71.9	71.9
m3g2f1	54.5	55.2	55.2	64.3	52.8	52.8
m3g2f3	56.5	61.8	61.8	65.1	63.5	63.5
μ	51.8	58.0	57.9	56.8	57.5	56.4
σ	7.5	7.5	7.6	12.0	13.8	13.5

Table 6.7: Integration experiments 3 and 4 compared to baseline. 'no shift' refers to experiment 3, '500ms shift' refers to experiment 4.

in §6.2.7.

6.3.2 Experiments 3 and 4: Weighted FOVA assignment with combination of 'all mass' landmark-specific language models

Experiments 3 and 4 aimed to reduce the false recognition of keywords found in experiments 1 and 2, caused by using the wrong landmark-specific model during N-Best list rescoring. They attempted this by employing integration function $f_2(p(W), \widehat{v}, t)$-*Weighted FOVA assignment with combination of the landmark-specific language models*instead of f_1 to allow for potential FOVA misclassification. To rescore each word, a language model was generated from all landmark-specific language models using weighted interpolation, the weights derived from the relative Euclidean distance of the eye position to each landmark at the word onset. Experiment 4 repeated experiment 3 but shifted the eye movement 500ms prior to the onset of the word.

Table 6.7 shows the results of integration. WER was observed to increase for all sessions, ranging from an increase from 0.7% (m3g2f1) to 10.4% (m1g1f3) whilst the FOM showed both increases and decreases. For experiment 3, session m1g1f3 showed an increase in WER of 10.4% over the baseline and a negligible decrease in FOM of 0.4%, showing that whilst the overall recognition performance decreased, the keyword spotting accuracy for words associated with visual attention remained the same. Decreases in FOM were also observed in session m2g1f2 (3.1%), m3g2f1 (11.5%) and

m3g2f3 (1.6%), whereas increases were observed in m1g3f1 (2.0%), m1g3f2 (5.4%), and m2g2f3 (14.0%). The average increase in WER was 12.3% and the average increase in FOM was 1.3%.

For experiment 4, shifting the eye movement by 500ms in relation to the speech resulted in minor changes in the WER and FOM compared to the no-shift case. This contrasted with experiments 1 and 2 where the FOM showed a greater change, indicating that there was less sensitivity in the rescoring process to the shift. This was an expected result because the N-Best list rescoring process using the integration function f_2 made use of every landmark-specific language model (combined using weighted interpolation) for modifying each word probability. This meant the eye movement had less effect in rescoring because the language model probability for *all* words associated with landmarks was increased for every word rescore, compared to f_1 where the language model probability for words associated with only one landmark was increased for each word rescore. Shifting the eye movement demonstrated this as a weakness.

Compared to experiments 1 and 2, experiments 3 and 4 showed larger increases in WER for all sessions and no overall trend for FOM. This suggests that using the weighted FOVA assignment reduced the recognition accuracy and did not improve keyword spotting accuracy. This also suggests that FOVA-misclassification (the motivation for integration function f_2) was not the main cause for the high false alarms rates in experiments 1 and 2.

6.3.3 Experiments 5 and 6: Deterministic FOVA assignment with selection of a 'compete-mass' landmark-specific language model

The results from experiments 3 and 4 showed that switching from a deterministic FOVA assignment to a weighted FOVA assignment did not yield better performance using the 'all-mass' language models. Experiments 5 and 6 reverted to the deterministic FOVA assignment (integration function f_1) from the first two experiments and use the 'compete-mass' landmark-specific language models rather than the 'all-mass' variants to try and achieve the performance expectations of maintaining WER and increasing FOM. Experiments 5 and 6 thus repeated experiments 1 and 2 using the 'compete mass' landmark specific language models instead of the 'all-mass'.

Table 6.8 shows the recognition results for experiments 5 and 6 in terms of WER and FOM. For all sessions, the WER showed a minor increase over the baseline.

Session	% WER			% FOM		
	Baseline	Rescored no shift	Rescored 500ms shift	Baseline	Rescored no shift	Rescored 500ms shift
m1g1f3	46.3	47.2	47.4	57.0	60.6	59.6
m1g3f1	45.2	45.3	44.9	70.7	65.9	69.7
m1g3f2	52.5	53.2	53.0	46.6	58.5	54.3
m2g1f2	64.7	64.8	64.7	35.9	41.4	42.5
m2g2f3	43.3	43.3	43.1	57.9	80.7	82.6
m3g2f1	54.5	55.3	55.2	64.3	56.5	58.0
m3g2f3	56.5	56.8	56.7	65.1	68.9	61.1
μ	51.8	52.3	52.1	56.8	61.8	61.1
σ	7.5	7.6	7.6	12.0	12.1	12.5

Table 6.8: Integration experiments 5 and 6 compared to baseline. 'no shift' refers to experiment 5, '500ms shift' refers to experiment 6.

For session m1g1f3 the WER increased 0.9% against the baseline and other sessions had similar results. The FOM showed an improvement over the baseline in 5 of the 7 sessions, with improvements ranging from 3.6% to 22.8%. Falls in FOM were observed in sessions m1g3f1 (4.8%) and m3g2f1 (7.8%). Overall, WER increased by 0.8% and the FOM by 10.5%.

For experiment 6, shifting the eye movement in relation to the speech resulted in minor variations in WER (-0.4% to $+0.2\%$) and FOM (-8.8% to 3.4%).

Compared to the first four experiments, which used the 'all-mass' landmark-specific language models, using the 'compete-mass' language model yielded better performance, indicating that redistributing the probability mass between words associated with visual attention, as opposed to redistributing the mass between all words in the vocabulary, is a better strategy. WER was maintained showing only a small increase over the baseline. The majority of sessions showed an increase in keyword spotting accuracy and thus met the performance expectations set out in §6.2.7.

6.3.4 Experiments 7 and 8: Weighted FOVA assignment with combination of 'compete-mass' landmark-specific language models

The success of experiments 6 and 7 in using the 'compete-mass' language models and integration function f_1 motivated experiments 7 and 8, where the 'compete-mass' language models were used with integration function f_2. Experiments 3 and 4 had

Session	% WER			% FOM		
	Baseline	Rescored no shift	Rescored 500ms shift	Baseline	Rescored no shift	Rescored 500ms shift
m1g1f3	46.3	54.5	54.4	57.0	63.6	64.7
m1g3f1	45.2	50.7	50.0	70.7	63.0	64.0
m1g3f2	52.5	52.6	52.6	46.6	62.8	62.8
m2g1f2	64.7	71.7	71.7	35.9	32.6	32.6
m2g2f3	43.3	47.3	47.5	57.9	79.0	79.0
m3g2f1	54.5	55.8	54.7	64.3	61.6	63.1
m3g2f3	56.5	60.3	60.3	65.1	64.6	64.6
μ	51.8	56.1	55.9	56.8	61.0	61.5
σ	7.5	8.0	8.0	12.0	13.9	14.0

Table 6.9: Integration experiments 7 and 8 compared to baseline. 'no shift' refers to experiment 7, '500ms shift' refers to experiment 8.

shown that using f_2 with the 'all-mass' language models gave the worst results, so these were repeated using the 'compete mass' landmark specific language models to see whether some improvement could be made as it did for experiments using f_1.

Table 6.9 shows the recognition results for experiments 7 and 8 in terms of WER and FOM. For experiment 7, WER increased in every session compared to the baseline, with the increases ranging from 0.1% to 8.2%. Both increases and decreases in FOM were observed, with 4 sessions decreasing (the decreases ranging from 0.5% to 7.7%) and 3 sessions increasing (the increases ranging from 6.6% to 21.1%). Compared to experiment 3, the average increase in WER was lower and the FOM performance comparable, suggesting again that the 'compete-mass' model worked better than the 'all-mass', however using integration function f_2 was still observed to be inferior to integration function f_3.

When the eye movement was shifted in relation to the speech in experiment 8, there was some minor variation in the results as in experiment 4, demonstrating again that the weighted assignment of the FOVA resulted in the eye movement not influencing the integration process sufficiently.

Session	% WER			% FOM		
	Baseline	Rescored no shift	Rescored 500ms shift	Baseline	Rescored no shift	Rescored 500ms shift
m1g1f3	46.3	52.8	52.7	57.0	65.2	61.0
m1g3f1	45.2	48.1	48.1	70.7	55.8	66.8
m1g3f2	52.5	52.1	52.5	46.6	42.0	41.0
m2g1f2	64.7	70.1	69.9	35.9	33.6	44.4
m2g2f3	43.3	47.7	47.3	57.9	84.2	86.0
m3g2f1	54.5	53.9	54.0	64.3	60.6	54.7
m3g2f3	56.5	59.5	59.6	65.1	63.2	29.7
μ	51.8	54.9	54.9	56.8	57.8	54.8
σ	7.5	7.8	7.8	12.0	16.5	18.6

Table 6.10: Integration experiments 9 and 10 compared to baseline. 'no shift' refers to experiment 9, '500ms shift' refers to experiment 10.

6.3.5 Experiments 9 and 10: Probabilistic FOVA assignment with selection of an 'all-mass' landmark-specific language model

The first 8 experiments established that a deterministic FOVA assignment (function f_1) and 'compete-mass' language model resulted in the best results. The final 4 experiments used the final integration function, $f_3(p(W), \widehat{v}, t)$- *Probabilistic FOVA assignment with combination of a landmark-specific language model and the baseline language model*, which used a measure of confidence in the deterministic FOVA assignment to back-off towards the baseline language model from the landmark-specific language model during word rescoring. f_3 was motivated by the thesis presupposition that looking at an object does not guarantee that words spoken will relate to it, only an increased chance of a related utterance[22].

Experiment 9 used the 'all mass' language models and experiment 10 repeated it with shifted eye movement. Table 6.10 shows the results for experiments 9 and 10. For experiment 9, integrating eye movement using f_3 and the 'all-mass model' resulted in increases in WER compared to the baseline for all sessions, the increases ranging from 0.1% to 8.2%. The FOM showed an increase against the baseline for 4 of the sessions, with increases ranging from 6.6% to 21.1%, whereas the FOM decreases ranged from 0.5% to 7.7%. For experiment 10, the effect of shifting the eye movement on WER was negligible and the FOM varied considerably. For example,

[22] §1.5

Session	% WER			% FOM		
	Baseline	Rescored no shift	Rescored 500ms shift	Baseline	Rescored no shift	Rescored 500ms shift
m1g1f3	46.3	47.2	47.3	57.0	60.6	59.6
m1g3f1	45.2	46.1	45.6	70.7	65.9	69.8
m1g3f2	52.5	52.9	53.3	46.6	58.5	54.3
m2g1f2	64.7	64.8	64.8	35.9	41.4	42.5
m2g2f3	43.3	42.9	42.6	57.9	57.0	82.5
m3g2f1	54.5	54.2	54.0	64.3	56.5	58.0
m3g2f3	56.5	57.1	56.7	65.1	67.2	61.1
μ	51.8	52.2	52.0	56.8	58.2	61.1
σ	7.5	7.5	7.6	12.0	8.5	12.5

Table 6.11: Integration experiments 11 and 12 compared to baseline. 'no shift' refers to experiment 11, '500ms shift' refers to experiment 12.

session m2g1f2 the FOM increased by 8.5% compared to the baseline, whereas in experiment 1 the FOM decreased by 2.3%. The results for experiments 9 and 10 were similar to experiments 1 and 2 (which differed in using f_1 instead of f_3) suggesting that the use of a probabilistic FOVA assignment over the deterministic assignment made little difference.

6.3.6 Experiments 11 and 12: Probabilistic FOVA assignment with selection of a 'compete-mass' landmark-specific language model

Experiments 11 and 12 used the integration function $f_3(p(W), \widehat{v}, t)$- *Probabilistic FOVA assignment with combination of a landmark-specific language model and the baseline language model.* Experiment 11 used the 'compete-mass' language models and experiment 12 repeated it with shifted eye movement. Table 6.11 shows the results for experiments 11 and 12. In experiment 11, WER was observed to be similar to the baseline model, with changes ranging from -0.6% to 0.9%. The FOM improved in 4 out of 7 sessions, with the changes ranging from -7.8% to 11.9%. The results for experiments 11 and 12 were similar to those for experiments 5 and 6 (which differed in using f_1 instead of f_3) suggesting, as they did for experiments 9 and 10, that the use of a probabilistic FOVA assignment over the deterministic assignment made little difference.

6.4 Analysis

The specific questions to be answered by the integration experiments were stated in §6.2.9. These were:

- Should only the focus being viewed be considered during integration (f_1) or should all foci be considered (f_2)?
- Is there any benefit in using the spatial proximity of visual attention to a visual focus to indicate the increased chance of a spoken utterance relating to the visual focus (f_3)?
- Given than an increase in the correlation between the modalities was empirically demonstrated in §6.1, does temporally shifting the eye movement stream in relation to the speech to account for the eye-voice span lead to improved ASR performance using either f_1, f_2, or f_3 ?

The experiment results were presented in §6.3.1 to §6.3.6, along with a preliminary analysis that found:

- The integration function f_1 was preferred over f_2.
- There was no benefit shown in using the spatial proximity of visual attention to a visual focus to indicate an increased chance of speech relating to that focus (f_3) over a deterministic assignment (f_1).
- Shifting the eye movement stream did not yield great differences in WER, however a large change in FOM, either positive or negative, gave an indication of the influence that eye movement was having on the ASR process.

The broader question for these experiments was:

- Can integration lead to ASR system performance gains in terms of maintaining WER and increasing FOM ?

The performance gains were demonstrated when using integration functions f_1 and f_3 with the 'compete-mass' model. This section collates the results for all experiments to provide further analysis and discussion of these findings.

6.4.1 Result visualisation

Figure 6.11 and Figure 6.12 graphically illustrate the change in WER and FOM in the form of a boxplot[23] for all sessions. The boxplot shows the distribution of results

[23]Boxplot is a feature of the Matlab statistical toolbox [Mat02]

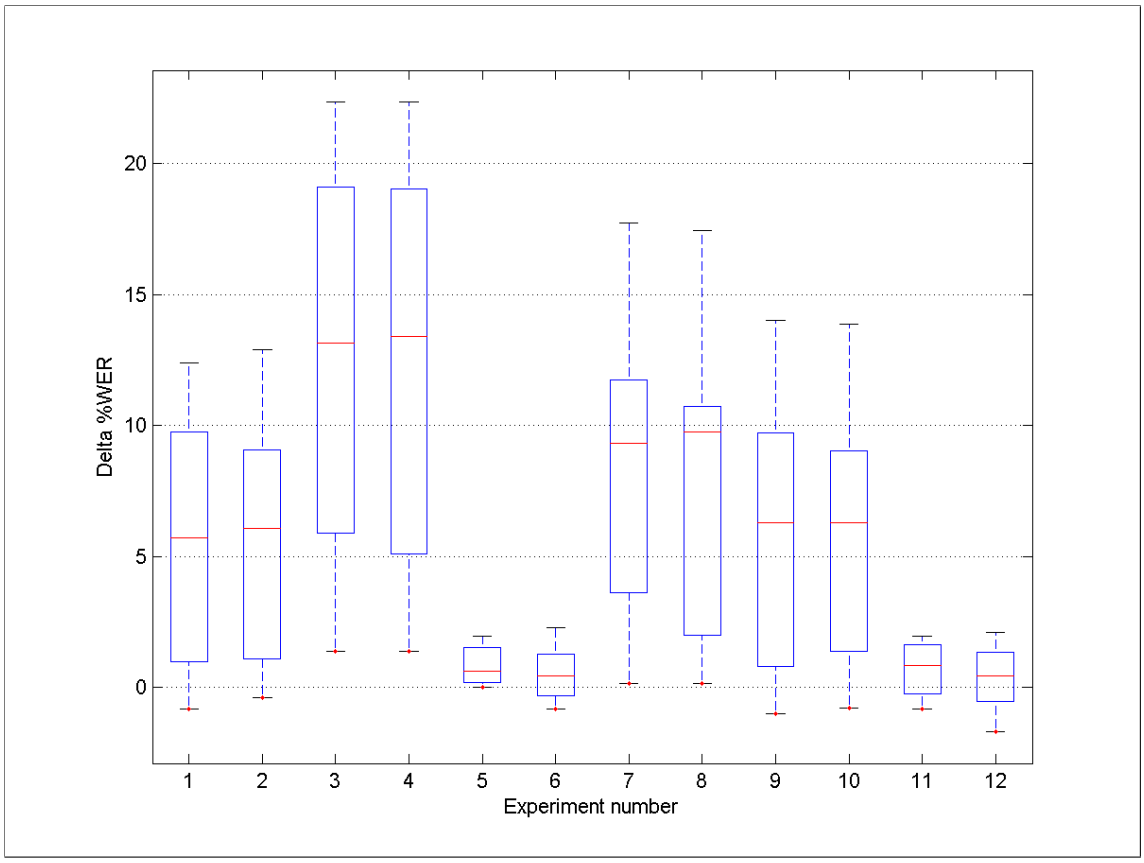


Figure 6.11: Boxplot showing $\Delta\%WER$ for integration experiments against the baseline ASR. The boxplot shows the distribution of results across all sessions as a box for each integration experiment. The top, middle, and bottom horizontal lines in each box indicate the upper, median, and lower quartile of the distribution. The vertical lines ('whiskers') extend from the box upwards and downwards by 1.5 times the interquartile range. Outliers are represented by crosses beyond the whiskers [Mat02].

across all sessions as a box for each integration experiment. The top, middle, and bottom horizontal lines in each box indicate the upper, median, and lower quartile of the distribution. The vertical lines ('whiskers') extend from the box upwards and downwards by 1.5 times the interquartile range. Outliers are represented by crosses beyond the whiskers [Mat02].E.g. Figure 6.12 experiment 9 had one outlying session.

For the $\Delta\%WER$ boxplot (Figure 6.11) the increase in WER is shown by the position of the box above the 0 line on the y-axis. This boxplot visually confirms the experiment results - that experiments 5, 6, 11 and 12 had lower increases in %WER compared to the baseline. Likewise, the $\Delta\%FOM$ boxplot, Figure 6.12, visually confirmed increases in FOM for these experiments and decreases for all others.

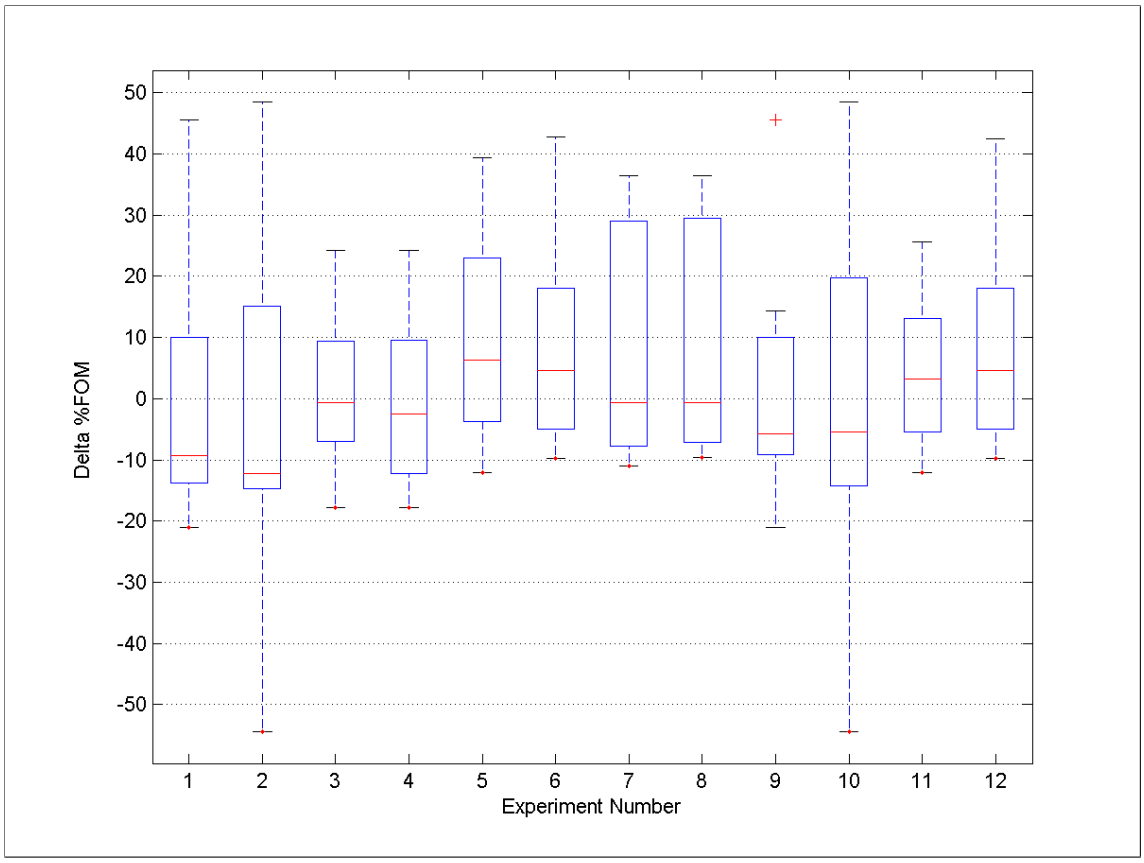


Figure 6.12: Boxplot showing $\Delta\%FOM$ for integration experiments against the baseline ASR. See the caption below Figure 6.11 for an explanation of the boxplot.

The boxplots also show the effect of shifting the eye movement in relation to the speech. The change in box size and position between the experiment pairs (e.g. experiment 1 and 2, 3 and 4 etc.) indicates the performance difference. Although variations were evident, no conclusions could be drawn from the boxplot other than the fact that there was some variation.

6.4.2 Integration behaviour

Table 6.12 summarises the results for the integration experiments in terms of the difference ($\Delta\%$) between the average performance, against the baseline ASR performance. To reveal more insights on the keyword spotting behaviour, the table includes the keyword spotting measures for True Positives (TP) and False Alarms (FA), in addition to the FOM.

In the four best performing experiments (5, 6, 11 and 12), WER was maintained

Experiment	$\Delta\%WER$		$\Delta\%TP$		$\Delta\%FA$		$\Delta\%FOM$	
	μ	σ	μ	σ	μ	σ	μ	σ
1	5.8	5.1	8.4	2.6	57.6	47.8	0.0	23.0
2	5.8	4.9	9.2	3.2	59.0	50.4	-4.1	32.2
3	12.3	7.8	7.9	4.2	151.6	106.7	1.3	13.6
4	12.1	7.9	7.7	4.3	151.6	106.7	-0.6	14.8
5	0.8	0.8	-5.8	5.8	-22.4	24.1	10.5	18.0
6	0.6	1.1	-4.2	6.9	-18.6	21.6	9.3	18.2
7	8.5	5.9	-6.4	5.8	-21.1	22.7	8.3	20.0
8	8.0	6.2	-5.9	5.9	-22.9	19.9	9.1	19.6
9	6.1	5.5	8.8	2.5	49.3	43.7	2.0	21.9
10	6.0	5.2	9.2	3.7	50.7	45.7	-1.1	32.5
11	0.7	1.1	-6.2	6.1	-20.6	23.3	4.3	13.0
12	0.4	1.3	-4.4	6.9	-16.8	20.3	9.3	18.1

Table 6.12: Summary of all integration experiments showing the change in ASR system performance against the baseline ASR system. The average change is presented as the sample mean (μ) and standard error of the sample mean (σ). Samples are the individual eye/speech corpus session results.

and the FOM improved upon, in relation to the baseline. The $\Delta\%FA$ ranged from -16.8% to -22.4% and the $\Delta\%TP$ ranged from -4.2% to -6.2%, showing that the overall increase in FOM was due to a fall in *both* the FA and TP compared to the baseline. This fall in TP meant that integrating the eye movement, on average, did not lead to the ASR system improving the recognition of words associated with the visual attention. What the integration did do however, was to lower the word probabilities for keywords associated with all landmarks other than the one being looked at, causing the reduction in TP and FA. WER was maintained (i.e. only a minor increase) because the word probabilities for the words not associated with landmarks were increased, relative to the words associated with all landmarks other than the landmark being viewed at the time of integration.

A reasonable prior assumption regarding the integration was that the TP rate should rise because the integration ensures that looking at landmark increases the chances of the ASR system recognising the landmark keywords. Mitigating this assumption however, is the fact that the Baseline ASR system performed well enough to recognise occurrence of the majority of keywords without having to use eye movement[24]. Further mitigation arises from the fact that visual attention and speech were

[24]See §6.2.7 Table 6.4

not 100% correlated[25]. Shifting the eye movement by 500ms supports this further mitigation as lower falls in TP were observed when increasing the correlation between the modalities (e.g. -4.2% for experiment 6 compared to -5.8% for experiment 5).

Experiments 7 and 8 also used the 'compete-mass' landmark-specific language models, however the performance improvement was not realised. The change in FOM over the baseline was positive (8.3% and 9.1% respectively) and similar to those in experiments 5, 6, 11 and 12 but the WER was not maintained against the baseline, increasing instead on average by 8.5% (experiment 7) and 8.0% (experiment 8). From this observation, using weighted combination of landmark-specific language models in rescoring meant that the probabilities for words associated with landmarks not being looked at were reduced less, leading to an increase in WER.

For integration using the 'all-mass' language models (experiments 1, 2, 3, 4, 9 and 10), WER, TP and FA showed increases and the FOM decreased against the baseline. This performance was worse than experiments that used the 'compete-mass' language models because the 'all-mass' models shifted probability mass from the non-landmark specific words, leading to large increases in WER.

6.4.3 Discussion

The main finding for the integration experiments was that integrating the eye movement into an ASR system helped to recognise spontaneous, conversational speech but it did not help to recognise words associated with the visual attention. Instead, the eye movement integration helped the ASR system to reduce the incorrect recognition of words associated with the visual field but not associated with the FOVA at the time of integration. Thus, WER was maintained and keyword spotting accuracy increased. Accounting for the 'eye-voice span' increased the correlation between the modalities although the increase was not enough to reflect in an increase in performance.

Various extensions could be made to this work. The implementation approach of rescoring the N-Best lists using word probabilities derived from landmark-specific language models and information about the eye position relied on the list supplying the correct entry, as though the ASR system had a language model that changed its word probabilities during Viterbi decoding in response to the visual attention. Computational resource capabilities enabled the use of 250-Best lists. On inspection of these lists it could be seen that the entries did not differ considerably, e.g., the same word sequence could arise in multiple entries differing only by the silences between words,

[25] See §6.1.2

word boundary times, and probability. The version of the HTK Viterbi decoder used (v3.1) does not support decoding dynamic language probabilities. An enhancement to the integration experiments would be to repeat them using a Viterbi decoder with this capability. This capability could be developed by either modifying the HTK code or writing a new decoder with support for HTK file formats. Alternatively, rather than rescoring N-Best lists, the space-time trellis[26] (also referred to as the N-Best lattice in the literature) used to generate the N-Best list could be rescored. This approach was not possible due to memory limitations in HTK regarding generating N-Best lattices with multiple tokens.

The benchmarks for the Baseline ASR system compared favourably with other research-grade systems from the literature[27]. As well as making the results in this thesis credible, the value that the integrating eye movement gave to the decoding process was not overestimated; a poorer performing baseline ASR system may have yielded greater performance improvements. To maintain the credibility of the Baseline ASR system but make the recognition task harder, the vocabulary of the recogniser could be increased and/or noisy speech data recognised.

Adding noise to the relatively noise-free speech data in the eye/speech corpus and repeating the experiments would be one approach however this would be flawed because of the different speaking patterns people adopt in noisy environments - e.g. the Lombard reflex [Jun93]. A better enhancement would be to use speech data recorded in noisy environments and/or use a remote microphone.

The Baseline ASR system was limited to the vocabulary of the eye/speech corpus. Increasing this vocabulary would result in lower FOM scores, because the ASR system would recognise words that are acoustically similar to keywords instead of the keywords. This would promote a situation where integrating the eye movement via landmark-specific language models would change the word probabilities so that the keywords were recognised instead. However, since the N-Best list contained repeated word sequences, the length of the list would have to be far greater than the computationally restricted 250 used in the experiments, to ensure that the list contained word sequences including the keywords. Using a bigger vocabulary therefore would require the Viterbi decoding with dynamic updating of language model probabilities over the N-Best list rescoring approach.

The integration experiments were carried out using the eye-speech corpus. The

[26] §2.4.3.2
[27] §4.7.1

number of landmarks on the map (12) was excessive given the estimated spatial accuracy of the EyeLink 1 eye tracker which tracked the eye within 23 pixels on a 640x480 pixel visual field at a sample rate of 75 kHz[28]. Repeating these experiments using eye tracking data covering a larger visual field and faster sample rate would be possible with current (2006) equipment, and would provide measurements that are more accurate and enable the use of a dynamic visual field with a limit to the number of visual foci present in it at any time.

The landmark-specific language models shifted 99% of the probability mass to words associated with a specific landmark from all other words. This figure was arbitrary and a better approach would have been to learn the amount of probability mass to shift. A possible approach would be to deduce this from the N-Best list by calculating the values by which word probability scores would need to be adjusted to provide the correct output.

Seven eye/speech corpus sessions (samples) were used in the integration experiments. In the results, their individual performance metrics (WER and FOM) for each session were presented. The analysis summarised these using the sample mean, μ, and the standard error of the mean, σ. The small sample size meant that a normal distribution could not be assumed and consequently σ was not given consideration during the analysis - only the sample means were compared. Comparison of differences between the sample means between each integration experiment and the baseline could have been enhanced by calculating the confidence intervals for the difference between sample means and performing hypothesis testing (i.e. the 2 sample t-test) to determine where WER and FOM increased, decreased or did not change against the baseline.

6.5 Summary

In this chapter, the correlation between eye movement and speech was characterised, using FOVA-level and eye-position level analyses. The 'eye-voice span' was confirmed. A linguistic analysis of the eye/speech corpus transcriptions verified that estimating whether a person is speaking about a visual focus is approximated by spotting explicit keywords related to the visual focus.

Integration of eye movements into ASR was achieved by rescoring N-Best lists using landmark-specific language models, as an approximation to updating word prob-

[28]Eye tracking accuracy and calibration issues were discussed in §3.3.6.1.

abilities during Viterbi decoding. This approach was evaluated to set performance expectations. Twelve ASR system variants were implemented and their performance compared to the baseline ASR system. Analysis of the results revealed that performance expectations were met by systems which reduced language model probabilities for words associated with visual foci other than the current focus of visual attention. Future enhancements to this work were discussed.

7 Conclusion

In Chapter 1 the following questions were posed in relation to integrating eye movement information into ASR systems:

- How do the modalities semantically relate?
- How can multimodal decoding be realised?
- What are the benefits?

A formal framework for multimodal decoding was proposed and used to determine how eye movement information could be integrated into an ASR system. To implement and test the system, a research-grade baseline ASR system was developed and a corpus of eye movement and speech data collected. Information was extracted from eye movement and noise-resilient methods for doing so evaluated. The relationship between eye movement and speech was characterised. Finally, Gaze-contingent ASR systems were developed to evaluate integration behaviour and speech recognition performance.

Throughout this thesis, consideration has been given to discussing the findings of the results, relating them to previous work and commenting on the method and suggesting enhancements. This chapter focuses on the main contributions of this study with respect to the original research questions posed and suggests directions for future research.

7.1 Contributions

To achieve multimodal decoding, this study addressed the following issues:

- The resolution of the temporal asynchrony between modalities.
- The resolution of the semantic asymmetry between modalities.

- The preference of early over late integration of modalities to avoid error propagation.
- A computationally efficient implementation.

This study addressed these issues using the novel application of integrating eye movements into an ASR system to recognise spontaneous, conversational speech. Such a goal envisaged multimodal interfaces that go well beyond today's simple command and control structure which require carefully spoken commands. This section summarises this study and its main findings in terms of the contributions made.

7.1.1 A formalised framework for combining modalities

To realise multimodal decoding and address the above issues, a formal probabilistic multimodal decoding framework was proposed using probability calculus. For time-series data, factorial Hidden Markov models were proposed to decode temporally asynchronous and semantically asymmetrical (i.e. loosely coupled) modalities. Factorisation and subsequent combinatorial explosion of the state spaces was avoided by defining the semantic relationship between the modalities and assuming that at any one time one modality helps to decode another, and not vice-versa. A speculative technique of applying transform functions to the individual streams to strengthen both the semantic relationships and synchrony was proposed. This technique uses the information-theoretic mutual information measure to discover how the modalities relate temporally and semantically. The multimodal decoding framework was applied to the specific application of integrating eye movement information into an ASR system to improve recognition performance. The transform function for eye movement was proposed as a temporal shift of the eye movement stream to account for the delay between looking at a visual focus and naming it.

7.1.2 The eye/speech corpus

To test gaze-contingent ASR systems, a novel corpus of matched eye movement and speech data was collected for a 2-person, visually-oriented, collaborative task. The task was inspired by the HCRC Map task. The participants were native British speakers with various accents. The speech recorded was spontaneous and conversational in nature. The hardware and software platform developed enabled modalities to be synchronously recorded. Time-aligned transcriptions of the speech were generated.

Experiment results verified the constraints on eye movement due to language production, for 2-person dialogues in less-constrained situations, using the eye/speech corpus as an exemplar.

7.1.3 A Baseline ASR system for British conversational speech

A research-grade baseline ASR system trained on conversational British speech was developed using the popular HTK software tools. Acoustic models were trained using British English read speech from the WSJCAM0 corpus. The models were adapted to the eye/speech corpus using MLLR and MAP adaptation. The novel aspect of the system was the language model, which was trained using conversational speech from both BNC and HCRC Map Task corpora, the latter providing task-specific language to complement the non-task specific language of the former. This approach of using two corpora to train the language model yielded better performance when recognising speech from the eye/speech corpus. Perplexity and WER measures were used to demonstrate the performance improvement. System performance of the Baseline ASR was credible.

7.1.4 Eye movement information extraction using the hidden Markov model and hidden *semi*-Markov model

Before integrating eye movement information into ASR, the information had to be extracted from the data. The information extracted was the eye movement-type and the focus of visual attention. Hidden Markov models and Hidden *semi*-Markov models (HSMM) were used to classify the data. The HSMM is a variant of the HMM which explicitly models the state duration probability distribution. The HSMM's and HMM's utility was characterised for eye tracking data which was subject to noise, motivated by the quality of the eye tracking data collected in less controlled conditions than in the laboratory. Benefits of the HSMM over the HMM were demonstrated. The incorporation of the Richter distribution into the state output PDF of the HSMM enabled the control of the balance of influence between state duration and state output PDF during decoding. The approach in this study promotes the development of noise-robust decoding techniques of eye movement data using the HSMM.

Multiple HMMs were run in parallel to distinguish different visual attention sequences, the application being to distinguish between user eye behaviours. The ran-

domness of visual attention shifts during scene perception limited the utility of this approach and future studies should be designed with this in mind.

7.1.5 A Gaze-contingent ASR

With the visual attention extracted from eye movement and a Baseline ASR system developed, novel gaze-contingent ASR systems were implemented using integration methods developed using the formal framework for integrating modalities laid out earlier in this work. The integration function the gaze-contingent ASR performed involved shifting probability mass within language models with the shifts guided by the visual attention sequence.

Twelve gaze-contingent ASR systems were realised which differed in the shifting method. The performance of ASR systems was measured using the Figure of Merit (keyword spotting accuracy) in conjunction with the popular Word Error Rate. The expected performance improvements in recognition were met within the constraints of the implementation scheme of rescoring N-Best lists. It was shown that information about what was *not* the focus of visual attention at any one time was more useful to ASR performance as information about what was.

Since the semantic relationship between eye movement and speech could be defined by relating visual foci to specific words, the Figure Of Merit proved a useful performance measure in this multimodal decoding task. This suggested that sets of keywords could be used to define the semantic concept that relates words to visual objects and that keyword spotting performance measures are a better method for evaluating multimodal ASR performance than Word Error Rate.

7.1.6 Software Tools

An object-oriented software library for computing vector algebra, HMMs, HSMMs, Gaussian probability distributions, and K-means clustering was developed for this work. Appendix A provides more details.

7.2 Recommendations for future research

The theoretical and technical contributions to multimodal decoding presented in this thesis provide the following insights for future research:

7.2.1 Other modalities and joint optimisation

Gesture could be considered as a third modality. The focus of visual attention is a deictic reference to an object in the visual field. A deictic gesture is the signalling (e.g. by pointing a finger or raising a hand) towards a visual focus. Therefore, the integration experiments may be repeated using gesture in addition to eye movement to determine the object about which the speaker is talking. Extending this work to the three-modality case would raise interesting questions about the dependency between modalities in decoding that could not be explored here. The integration experiments used eye movement to improve ASR and not vice-versa and this was the right approach given the relative ease of recognising the focus of visual attention compared to recognising speech. Deictic gesture and gaze however are more evenly matched in terms of recognition difficulty and thus there would be more opportunity to use either to improve the other's decode.

7.2.2 Corpora

The integration experiments considered only the Instruction giver's speech for speech recognition task. The eye/speech corpus also provides an opportunity for automatic information extraction to explore whether the map can be learnt when monitoring the conversation between two people.

Resources available for this study enabled the eye/speech corpus to be collected and the number of sessions that contained high-quality eye movement data was small with the experiment design confined to tracking a participant eye in relation to a computer screen. Given more resource, a larger corpus tracking the eye movement in relation to the environment would be a useful in supporting the results in this study and incorporating other modalities such as gesture.

7.2.3 Practical applications

Intended application areas for this research include:

- Speech-centric multimodal recognition systems.
- Machine understanding and monitoring of human-to-human communication.
- Human-computer interface technologies.

7.2.4 Maximising mutual information and machine learning

The formal framework for combining modalities for recognition outlined an approach to improve the temporal and semantic coupling between modalities by maximising the mutual information between them using a coupling function. The mutual information approach however was not used in the experiments. Instead, the semantic coupling was defined by identifying words that related to visual foci by inspection of the transcriptions. The temporal coupling was defined by time shifting the eye movement data to account for language production. Further work in this area would be to use the MI approach to learn about the relationship between words and visual foci, and the temporal shifts - i.e. a machine learning approach.

Appendices

Appendix A: Software Tools

Implementation and Platforms

This thesis was written using the MiKTeX LaTeX system, and TexnixCenter editor under Microsoft Windows. The Memoir book class was used for a template.

The software developed in this study for was primarily written in C# programming language. The compiled binaries ran under Microsoft Windows using the .NET framework, and Linux using the Mono framework. Some simple textual parsing and file manipulation was done using Linux shell script and Perl. Matlab and Microsoft Excel were used to draw graphs. Speech recognition was performed using CUED HTK running under Linux.

A number of C# classes were written for analysis of eye movement and speech data. As this study progressed, many bespoke tools were refactored, resulting in a collection of classes that are reused by various applications. Of primary importance is the 'DataMining' and 'DataParser' class libraries.

DataMining class library

The DataMining class library computes HMMs, HSMMs, Gaussian PDFs, GMMs and the necessary vector algebra, in addition to K-Means clustering. Figure 1 shows the main classes in the library. The data mining library was extensively used for eye movement analysis in Chapter 5. This generic library can be used beyond this thesis for building a variety of HMM-based applications. It has been released as open source software under GPL/GNU.

DataParser class library

The DataParser class library parses files of eye and speech data, for use by the applications. Figure 1 shows the main classes in the library.

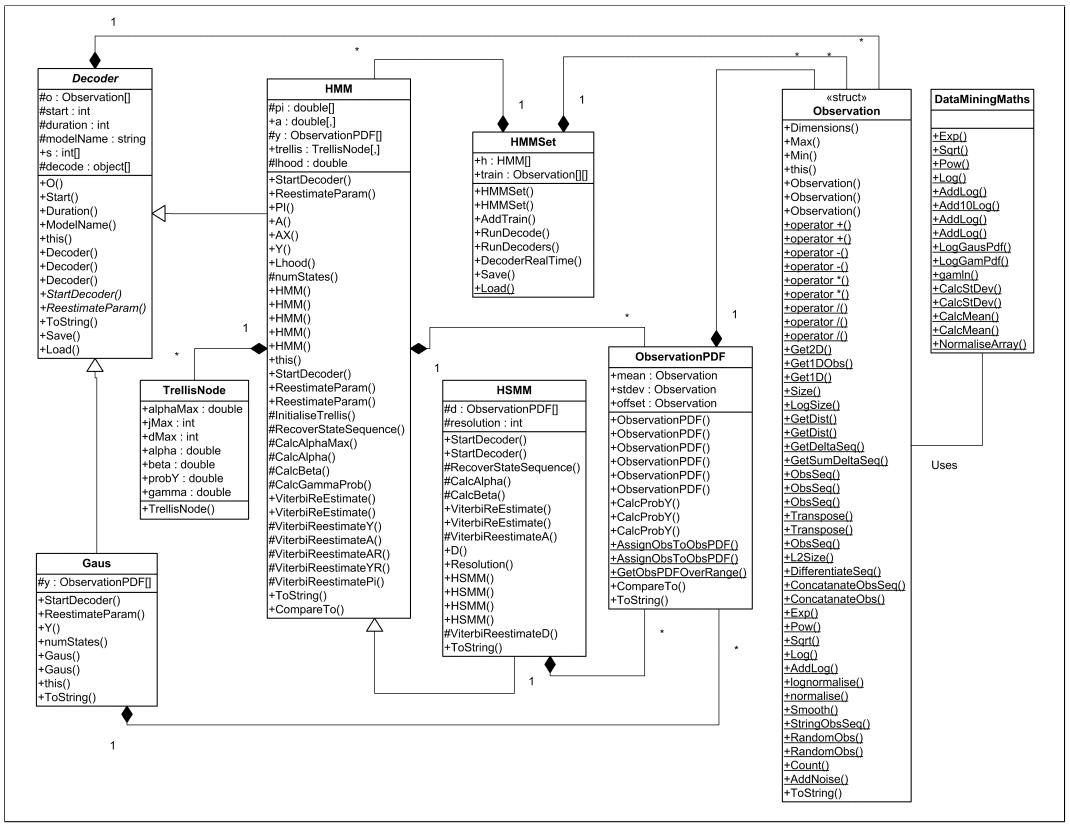


Figure 1: UML static class diagram showing a selection of the main classes in the DataMining class library. Observation represents an N-Dimensional vector. Gaussian PDFs are represented by the ObservationPDF class. HMM, HSMM and Gaussian (Nearest Neighbour) decoders are part of a inheritance hierarchy from an abstract Decoder class. Parallel decoding using multiple HMM's is implemented by a HMMSet class. Model parameters for HMMs may be saved to file in XML format.

Application layers

Applications make use of the class libraries described above. The main application areas were:

- Visualisation: A visualisation tool enables eye movement and speech to be viewed and analysed through a Graphical User Interface (GUI). The visualisation tool consists of a number of different methods for viewing data in the eye/speech corpus sessions, and running HMM and clustering algorithms. Figure 3 shows a screenshot of the visualisation tool.

- Eye Movement analysis

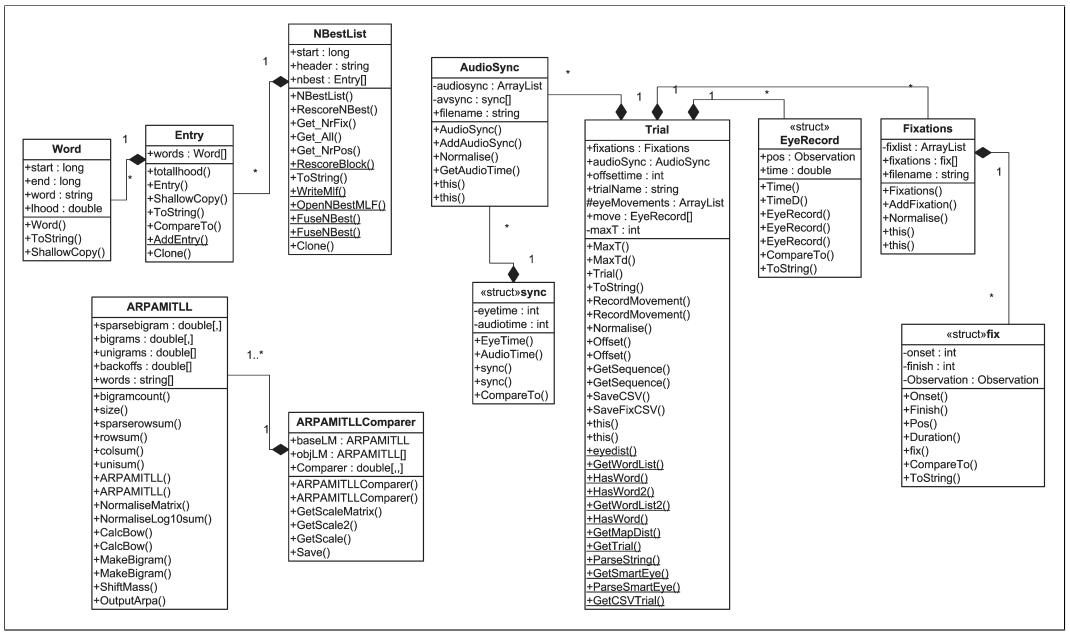


Figure 2: UML class diagram showing a selection of the main classes in the DataParser class library. Trial represents an Eye/Speech corpus session, composing of classes for synchronisation (AudioSync) and representing eye movement in terms of fixations (Fixations) and positions (EyeRecord). The N-Best List class is parses ASR output HTK from HTK allowing rescoring of lists. Bigram language models are represented by ARPAMITTL class. Language models can be compared using the ARPAMITLLComparer class.

- General parsing
- Language modelling
- Rescoring

Recording software acknowledgement

The collection of the eye/speech corpus data required custom sound recording software, which synchronised with the EyeLink eye tracker data collection. This software built in C# using DirectX 9.0 library. The audio recording routines, which gave access to sample counts and configuration options, were kindly supplied by Paul Dixon.

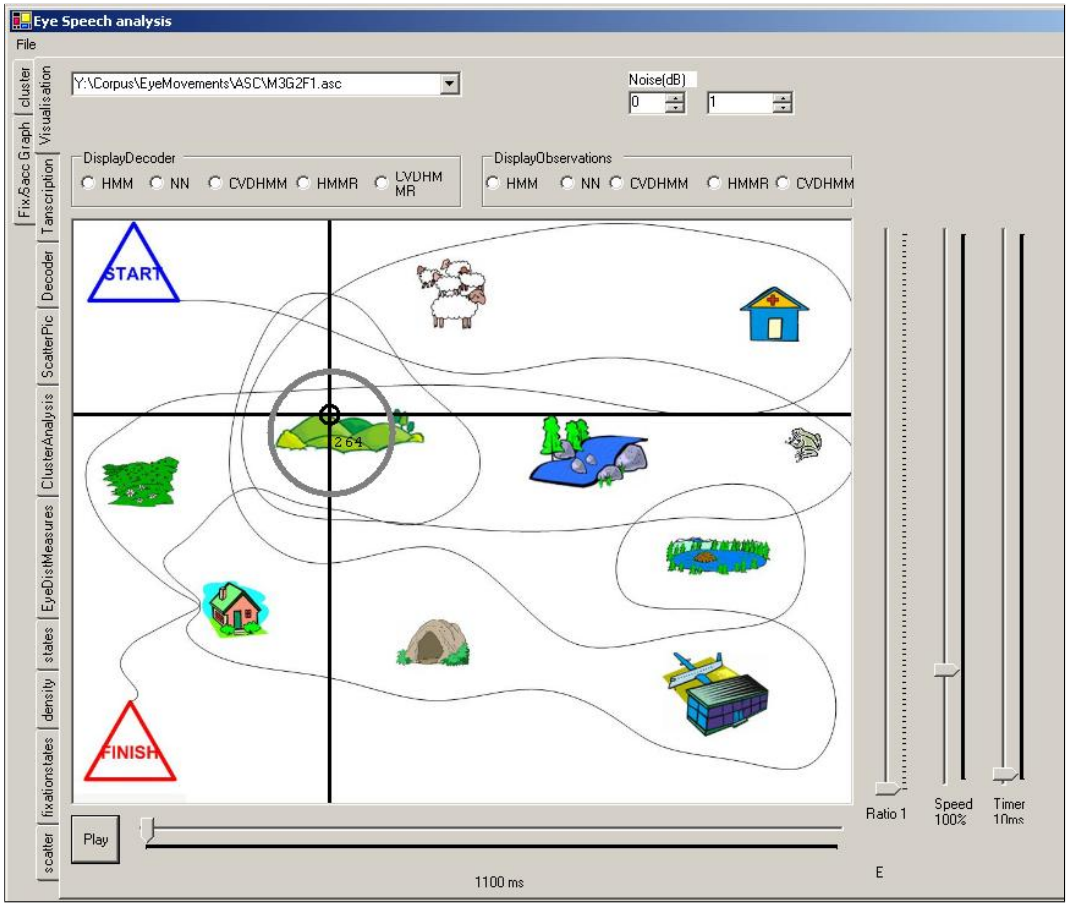


Figure 3: Visualisation tool built for analysis. The visualisation tool enabled eye movement sessions to be replayed along with audio and analysis with visual results.

Appendix B: Publications

Two conference publications are included in this section:

- Neil Cooke / Russell, Martin / Meyer Antje (2004): 'Poster Abstract: Evaluation of hidden Markov models robustness in uncovering focus of visual attention from noisy eye-tracker data.', In ETRA 2004, 56.
- Cooke, Neil / Russell, Martin (2005): 'Using the focus of visual attention to improve spontaneous speech recognition', In INTERSPEECH-2005, 1213-1216.

Poster abstract - Evaluation of hidden Markov models robustness in uncovering focus of visual attention from noisy eye-tracker data

Neil Cooke* Birmingham University
Martin Russell† Birmingham University
Antje Meyer‡ Birmingham University

1 Introduction

Eye position, captured via an eye tracker, can uncover the focus of visual attention by classifying eye movements into fixations, pursuit or saccades [Duchowski 2003], with the former two indicating foci of visual attention. Such classification requires all other variability in eye tracking data, from sensor error to other eye movements (such as microsaccades, nystagmus and drifts) to accounted for effectively.

The hidden Markov model provides a useful way of uncovering focus of visual attention from eye position when the user undertakes visually oriented tasks, allowing variability in eye tracking data to be modelled as a random variable.

2 Proposed Models

Two types of hidden Markov models are investigated- a standard hidden Markov model (HMM) and a hidden semi-Markov model (HSMM) [Russell and Moore 1985]. The HMM represents state duration inherently, decaying geometrically from the time of entering the state. The HSMM represents state duration explicitly, allowing more accurate representation of visual attention duration.

The visual field is represented by an ergodic HMM/HSMM with each hidden state representing a foci of visual attention. These foci are explicitly identified as regions of interest in the visual field. A state output PDF for each state describes the probable distribution of eye positions over an object while that object is the focus of visual attention.

The HMM and HSMM are implemented in two variants. The first uses a standard two-dimensional gaussian distribution as the observation PDF to represent the distribution of eye positions over an object. The second adds a second gaussian component to each dimension in the observation PDF with the same mean as the first component but with larger standard deviation (the Richter Distribution [Richter 1986]). This ensures less differentiation in the observation PDF between eye movement positions that are far away from the object compared to that of the standard gaussian PDF in the first variant.

*e-mail: cooken@eee-fs7.bham.ac.uk
†e-mail:m.j.russell@bham.ac.uk
‡e-mail:A.S.Meyer@bham.ac.uk

3 Evaluation

We evaluate performance of the HMM and HSMM in classifying focus of visual attention from eye tracking data from users undertaking a visually-oriented task. Increasing levels of random gaussian noise are added to the data. Performance degradation is measured in terms of:

Accuracy : proportion of states decoded correctly compared to a baseline non-HMM method with no added noise.

Instability : proportion of state transitions occurring compared to a baseline non-HMM method with no added noise.

4 Results

- Performance of all HMM and HSMM-based methods to added noise is better than the baseline non-HMM method.
- Best performance results from using an explicit state duration PDF (HSMM) with Richter distribution observation PDF.
- Weakened performance results from using the Richter distribution observation PDF in a HMM (as opposed to a HSMM).

5 Conclusion

Our findings show that a hidden semi-Markov model that has an explicit state duration PDF representing task-constrained visual attention is a more stable and accurate way to uncover the focus of visual attention from (simulated) noisy eye tracker data compared to using standard HMMs with inherent state duration PDF. This performance gain is only evident when the observation distribution PDF's dominance on classification is relaxed by adding an additional gaussian component to the observation distribution PDF to reduce discrimination between eye positions far away from the object.

Hidden semi-Markov models have promising uses in uncovering focus of visual attention from noisy eye tracking data. HMM-based formalisms of decoding eye movement analogous to decoding speech (and other modalities) may also facilitate development of robust decoding schemes in multi-modal systems.

References

DUCHOWSKI, A. T. 2003. Eye Tracking Methodology: Theory & Practice. Springer-Verlag, ch. 4, 43–51.

RICHTER, A. 1986. Modelling of continuous speech observations. In Advances in Speech Processing Conference, IBM Europe Institute.

RUSSELL, M. J., AND MOORE, R. K. 1985. Explicit modelling of state occupancy in hidden markov models for automatic speech recognition. In Proceedings International Conference on Acoustics Speech and Signal Processing (ICASSP 85), Tampa, 1.2.1–1.2.4.

Using the focus of visual attention to improve spontaneous speech recognition

Neil Cooke, Martin Russell

School of Electrical, Electronic and Computer Engineering
University of Birmingham, Birmingham, United Kingdom
n.j.cooke@bham.ac.uk

Abstract

We investigate recognition of spontaneous speech using the focus of visual attention as a secondary cue to speech. In our experiment we collected a corpus of eye and speech data where one participant describes a geographical map to another while having their eye movements tracked. Using this corpus we characterise the coupling between eye movement and speech. Speech recognition results are presented to demonstrate proof of concept for development of a bimodal ASR using focus of visual attention to drive a dynamic language model. Marginal improvement in WER is observed.

1. Introduction & Background

This study demonstrates automatic speech recognition in a multimodal context using gaze direction as a secondary modality to speech. Gaze direction provides the focus of a speaker's visual attention, indicating which objects in the speaker's field of view may be the subject of speech. The motivation thus is to exploit the relationship between these modalities to improve recognition performance, specifically for spontaneous speech between humans, as opposed to more constrained forms of dialogue between human and machine.

A multimodal recognition architecture is desirable as it affords an opportunity to resolve ambiguity, and various architectures are emerging which integrate semantically rich modalities with this aim[1]. Ambiguity in speech recognition is primarily caused by varied speaking style [2] and noise [3]. Likewise in eye movement ambiguity in recovering the focus of visual attention from gaze direction is due to a person's oculamotor imperfections and sensing errors in the eye tracker.

The integration of modalities is a data fusion problem. During the pattern recognition process modalities may be integrated at sensor, feature or decision level. The choice of fusion level is determined by the physical, temporal and semantic relationship between modalities and the multimodal pattern recognition process's target classification space(s).

The use of eye tracking to improve the the speech recognition process has previously been researched in terms of recovery of errors due to misnaming objects. For example in [4] participants named objects which had phonologically similar sounds on a VDU and the nearest object to the eye was used to resolve which object was being referred to. This work differs from this previous study as we are interested in natural dialogue between 2 people and eye movement is being used in its natural role as a passive modality rather than an active modality in a command driven interface.

Figure 1: Experimental setup showing participant wearing eye trackers(left) in acoustic booth giving instructions to instruction follower (right) situated outside the booth and out of sight.

2. Experiment

2.1. Participants and task

A corpus of eye and speech data was collected for spontaneous, visually-oriented task-constrained dialogue between two participants - an instruction "giver" and an instruction "follower". The giver described a map, presented on a VDU, containing objects (landmarks) and a route around them, to a follower. Neither participant could see the other and they communicated via microphones and headphones. The giver's gaze direction relative to the map on the VDU was recorded, as was the speech of both participants. See Figure 1. The task participants undertake is loosely based on the HCRC Map Task Corpus [5].

There were 3 map sets used in the experiment. Each Map set consists of 6 maps. Each map in the map set has the same objects (landmarks) as others in the set but shows a different route. The objects are illustrative to encourage participants to describe them in detail. Figure 2 shows an example map.

There were 9 participants who, working in groups of 3, formed 6 pairs of giver/followers with each participant in a group of 3 taking the role of giver 2 times. Each group of 3 used all 3 map sets, so that followers previously participating as givers (and vice versa) were not presented with a map they had seen before. In total 18 sessions are recorded each with eye movement and speech data from a different instruction giver.

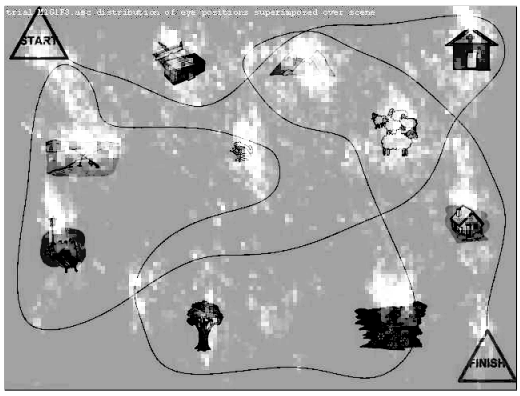


Figure 2: Map displayed to participant showing objects and route, superimposed with distribution of participant's eye positions with distribution of eye movement indicated by brightness

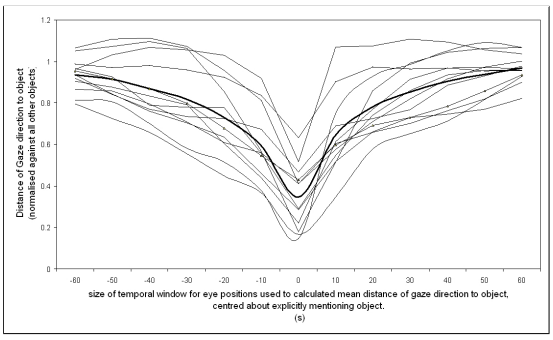


Figure 3: Mean distance of gaze relative to object when explictly mentioning object. Each thin line corresponds to a single object and the thick line corresponds to the average over all objects. The x-axis shows mean gaze distance before and after utterance indicated by negative and positive values respectively.

2.2. Apparatus

The experiment used an SR Research Ltd. "EyeLink 1" eye-tracking system to capture binocular gaze position relative to the screen within an accuracy of .01 degrees at 4ms sampling intervals. The eye tracker is owned by the Behavourial Brain Sciences centre in the School of Psychology at Birmingham University and its primary use is speech production research.

The giver's speech was captured using a high quality desk mounted microphone. The follower's was captured using a head-mounted microphone. Participants heard each other via headphones. Audio was recorded digitally in stereo (22.5Khz per channel) directly to a WAV-format audio file. The giver's and follower's speech was recorded to the right channel and left channel respectively. To ensure synchronisation between eye and speech data for reconstruction during analysis a software sound recorder was developed which uses a handshaking protocol enabling the audio software to insert periodic audio sample counts into the eye tracking data during capture. The handshaking protocol compensated for latencies between eye and sound capture softwares.

2.3. Data collection and processing

Session durations were typically in the order of 5 to 15 minutes. Whereas collecting speech data in controlled conditions is relatively reliable, eye tracking is fraught with difficulties. Previous research indicates that 10%-20% of participant's eyes cannot be tracked reliably [6] and evaluating the quality of the recordings must be done on a frame by frame by frame basis which can be time consuming and prone to error.

For this experiment two methods were employed to measure the quality of eye recordings. First, the raw eye and speech data was processed together to make compressed (MPEG-4) video sequences of sessions. The videos were observed and subjectively evaluated. Errors observed in eye movement recordings were typically degradation in calibration during the trial, noisy eye movements due to participants nystagmus (trembling of the eye), or poor pupil tracking in the eye tracker due to its sensitivity to ambient light levels. Corruption in eye data also occurred when participants inadvertently move the eye tracker while coughing, gesticulating and involuntarily touching their faces.

The second quality evaluation exercise was to visualise the distribution of eye positions about the image using analysis software. Figure 2 shows the concentration of eye positions about the objects on the map for a session. This gave a useful indication of quality and enabled correction of offsets in calibration. 4 sessions were recovered in this way.

Speech signal files were transcribed using the Transcriber software [7]. 2 passes of each session file were undertaken by different people.

The recalibration and visualisation of recorded eye data was done using custom software developed in C# .NET. This software is also used in the analysis of the relationship between direction of gaze and speech described in this paper.

Despite the problems in recording eye movement, out of 18 sessions undertaken 10 were deemed acceptable quality. For the subsequent analysis in this paper 3 sessions are used as these were identified as high quality early on in the research effort. The speech data collected consists of 35920 utterances from a 1116 word vocabulary.

3. Relationship between direction of gaze and speech

For speech recognition systems to eefectively use the direction of gaze, one must characterise the relationship between the two modalities. A reasonable hypothesis is that the proximity of gaze direction to an object indicates an increased chance of an utterance referring to an object being spoken. To prove this the mean distance of the gaze direction to each object is calculated and normalised against the distance of the gaze direction to all other objects while the user is explicitly mentioning a noun related to that object.

The eye positions used to calculate the mean distance of the eye to object are initially those for eye positions falling within the duration of the spoken noun pertaining to the object only. The mean distance of the eye to an object is then recalculated to include eye positions before the utterance, and also after the utterance. These windows of eye positions enable the temporal relationship between eye position relative to an object and its spoken utterance to be characterised.

Figure 3 shows the mean eye position relative to its spoken utterance for the giver's speech for one session. The x-axis

indicates the window of eye positions. For example at 0s the mean position of the eye is calculated for eye positions that occur while the giver explicitly mentions the object. At -10s on the x-axis the mean position is calculated for positions while the giver mentions the object and all eye positions up to 10s $prior$ to mentioning the object. Similarly, at +10s the mean position is calculated for positions while the giver explictly mentions the object, and up to 10s $after$ and so on. The mean distance for all objects is indicated by the thicker line on the graph. This analysis was undertaken on two other sessions and showed similar results for all.

Figure 3 show that when explicitly mentioning the object the gaze direction is nearer to the object than average. This is the correlation between modalities that can be exploited in a speech recognition system.

4. Bimodal ASR using focus of visual attention

4.1. The speech recognition model

A typical Hidden Markov Model (HMM) large vocabulary ASR is built using the HTK Toolkit. The ASR uses decision tree clustered tied-state triphone three-state HMMs with 4-mixture component gaussian observation distribution per state. Audio is parameterised using static and dynamic Mel-Frequency Cepstral Coefficients (MFCC) with Cepstral Mean Normalisation.

The HMMs were trained using designated phonetic transcriptions in training data in the UK English WSJCAM0 corpus [8]. The system was developed using a cluster of 7 Pentium 4 PCs running Grid Engine Distributed Source Management (DRM) software under Linux Red Hat 9.0.

The final system supports 23434 triphones made up from all triphones derived from the British English Example Pronunciation (BEEP) dictionary plus those observed in the data captured during this study. Recognition performance is 55.01% Word Error Rate(WER). This system incorporates a uniform language model with 12625 word vocabulary using the WSJCAM0 designated test data consisting of 48 speakers each reading 40 sentences from a 5000 word vocabulary.

4.1.1. Acoustic adaptation, language model and vocabulary

Acoustic model adaptation is applied to the WSJCAM0 models using Maximum Likelihood Linear Regression (MLLR). The HMMs are adapted to all speech data collected in the trial (corpus adaptation) and to individual speakers (speaker adaptation). Corpus acoustic adaptation gave improvement in WER whereas speaker adaption did not improve recognition due to data sparsity.

Two recognition systems are constructed from the models, one with a 1116 word vocabulary restricted to those words spoken during the experiment and using a unigram language model based on word frequency (henceforth referred to as "1K" system), and another with a 12911 word vocabulary consisting of all words used in the WSJCAM0 corpus for training and testing in addition to those in this experiment (henceforth referred to as "13K" system). The 13K system uses a uniform language model.

4.2. Integration of gaze direction into ASR

Language models in HMM-based ASR describe word transition probabilities. The development a gaze-contingent language model in ASR is a feature-level data fusion exercise resulting in dynamic word transition probabilities in the HMM.

To demonstrate proof of concept in developing such a system we reorder N-Best lists of ASR output.

The focus of visual attention is determined from the gaze direction by finding the nearest object to the gaze direction by measuring the Euclidean distance. Although this does not account for eye positions that do not indicate focus of visual attention, such as rapid eye movements between fixations (saccades), nor the noisy input due to eye tracker errors and eye (nystagmus), it is sufficient for this study as the EyeLink eye tracker output is relatively high quality.

The audio for each session used in this analysis is split into segments using silence in the giver's speech as indicated in the transcriptions. This results in a set of speech segments relating to the giver's speech only. Each segment is run through both the 1K and 13K vocabulary ASR systems and 100-best lists generated for each speech segment. Near optimum word insertion penalties and language model scale factors are determined empirically however this is not a priority and results should be interpreted as sub-optimal.

The 100-best lists are originally ordered based on their log likelihood. Word counts are calculated representing the number of words relating to the object which is the focus of visual attention during the speech segment. The words themselves are recovered from transcriptions. The word count is calculated for each level in the 100-best list. The list is then reordered using this word count as the primary sort key and the original log likelihood as a the secondary sort key. The list is also reordered using a word count relating to the entire map. This process is repeated for each speech segment. This provides 3 recognition results - speech ASR; speech and focus of visual attention ASR; speech and visual context ASR.

The motivation behind a speech + visual context ASR is to determine whether using eye movement is worthwhile over the more general case of priming an ASR with semantic information from the whole visual field, given that the latter is considerably easier to achieve in practice.

4.3. Recognition performance

Table 1: *WER and % segments rescored for 1K and 13K ASR systems*

1K ASR		
Modalities	Rescored	WER
Speech only	0%	72.66%
Speech + Focus of Visual Attention	25.23%	71.38%
Speech + Visual Context	39.64%	72.84%
13K ASR		
Modalities	Rescored	WER
Speech only	0%	91.56%
Speech + Focus of Visual Attention	15.32%	89.72%
Speech + Visual Context	27.03%	90.46%

Table 1 shows %WER and % of segments whose N-Best lists were rescored. As expected the visual focus of attention is rescored less often than visual context as it is more discriminating.

To compare the performance of ASR systems, the number of levels, n, that are used to determine the optimum WER is varied from 1 to 100 for both the 1K and 13K systems. For

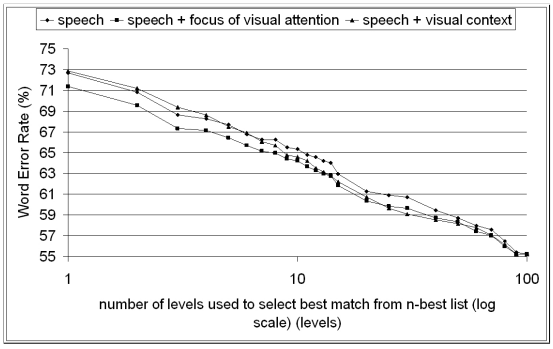


Figure 4: Comparison of ASR systems using 1116 word vocabuary showing focus of visual attention improving performance and visual context degrading performance

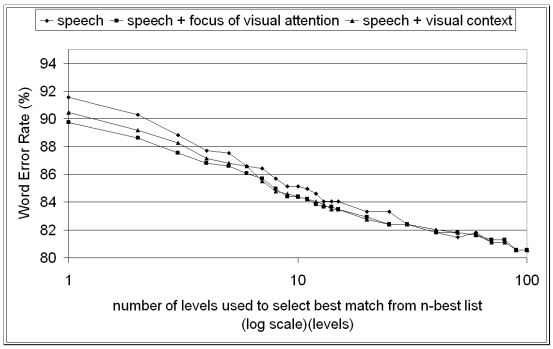


Figure 5: Comparison of ASR systems using 12911 word vocabuary showing focus of visual attention and visual context improving performance

example if $n = 10$ then the performance is based on the best result from the first 10 levels in the N-Best list.

For the 1K system shown in Figure 4 adding focus of visual attention is observed to reduce WER whereas the use of visual context in using a 1K vocabulary results in performance which is worse - the limited vocabulary increasing the chance of misclassification.

For the 13K system shown in Figure 5 adding focus of visual attention and visual context both improve recognition performance - demonstrating that recognition using visual context, while affecting performance compared to eye movement, is better than using speech alone.

Recognition improvement is marginal. It was found the nouns representing objects are normally recognised as participants spoke object names clearly. The ASR has more problem with shorter, general vocabulary words, for which the gaze direction (or visual context) provides no information about.

5. Conclusion & Future work

A modest improvement in ASR performance is observed when adding the focus of visual attention as a secondary modality. A new corpus of eye movement and speech has been collected containing passive eye movement and synchronised spontaneous speech between two people when participating in a visually-oriented task. It has been shown that eye movement is marginally better than visual context for priming language models in ASR, since the chances of misclassification are reduced - although for high vocabulary systems visual context may suffice.

From inspection of the N-Best lists it is observed that while recognition performance is improved, the improvement is marginal because participants tend to clearly speak the words associated with the objects in comparison with more general words occuring in everyday speech. Consequently the ASR has little or no problem recognising words associated with the objects leaving only small room for improvement by introducing focus of visual attention or visual context. This is not likely to be the case in noisy environments,

Whereas in this paper N-Best lists are rescored based on prior knowledge of which words are associated with which objects on the map, for future work this particular study in bimodal ASR will be completed by rescoring N-Best lists based on a data driven approach, deriving a probabilistic measure of word likelihood from N-gram language models which incorporate gaze direction information. Additionally audio input will be degraded to see whether using gaze as a secondary modality can increase robustness to noise. Finally, we envisage developing speech-centric recognition schemes incorporating gesture and gaze.

6. References

[1] Oviatt, S., "Breaking the Robustness Barrier: Recent Progress on the Design of Robust Multimodal Systems",Advances in Computers, 56 pp. 305-333,Elsevier 2002.

[2] Weintraub, M.,Taussig, K., Hunicke, K., and Snodgrass, A. "Effect of speaking style on LVCSR performance". Proc. of the Conversational Speech Recongition Workshop/DARP Hub-5E Evaluation. Morgan Kaufman, San Mateo, CA. 1997.

[3] Gong, Y.,"Speech recognition in noisy environments: a survey",Speech Communication, 16(3), pp. 261-291, 1995.

[4] Zhang, Q., Imamiya, A., Go, K., Mao, X., "Designing a Robust Speech and Gaze Multimodal System for Diverse Users", IEEE International conference on Information Reuse and Integration, 2003. 27–29 Oct. 2003 pp354–361 SIGGRAPH, 2004.

[5] Anderson, A. et al "The HCRC Map Task Corpus". Language and Speech, 34(4):351366. 1991.

[6] Jacob, J. K., Karn, K. S. "Commentary on Section 4. Eye tracking in human-computer interaction and usability research: Ready to deliver the promises." In Hyona J., Radach, R., Deubel, H.(Eds.), The Minds Eyes:Cognitive and Applied Aspects of Eye Movements. North-Holland. 2003.

[7] Barras, c., Geoffrois, Z., Wu, Z., Liberman, M., "Transcriber: development and use of a tool for assisting speech corpora production",Speech Communication, 33(1-2), pp. 5-22, January 2001.

[8] Robinson, T., Fransen J., Pye D., Foote J., Renals S., "WSJCAM0: A British English speech corpus for large vocabulary continuous speech recognition", Proc. IEEE ICASSP 95, 81-84, Detroit, 1995.

Summary and critique of INTERSPEECH-2005 conference publication

The Interspeech-2005 conference publication 'Using the focus of visual attention to improve spontaneous speech recognition' [CR05] was an initial attempt at integration. This section provides a summary and critique of this publication so that it can be read in the context of this thesis.

The baseline ASR system[1] was used although at the time of the experiment it was not fully developed. The acoustic adaptation to the eye/speech corpus used only MLLR and not MAP. A 1018 ('1k') word unigram language model containing the eye/speech corpus vocabulary, and a 12,812 ('13k') word uniform language model containing words from the WSJCAM0 training and evaluation set and eye/speech corpus vocabulary were used, instead of the (yet-to-be developed) BNC/MapTask-derived language model.

Visual attention was determined from the eye position by finding the nearest landmark to the eye position by measuring the Euclidean distance. Each Instruction giver's speech segment was recognised with the 1K and 13K vocabulary ASR systems. For each speech segment, 100-best lists were generated. The lists were rescored based on the number of occurrences of landmark keywords in the speech segment that related to the temporally corresponding FOVAs. Temporal correspondence was on a per-segment basis - e.g. If the participant looked at 3 different landmarks during an utterance, keywords for all 3 landmarks would be considered for the word count that determined the N-Best list reordering.

A small improvement in recognition accuracy was demonstrated. In the '1k' ASR system adding eye movement lowered WER from 72.7% to 71.4%, with 25% of the speech segments rescored resulting in a change in the most probable word sequence. Likewise, in the '13K' ASR system adding eye movement lowered WER from 91.6% to 89.72%, with 15% of the segments rescored. These results were far from conclusive, due to high WER.

The high WER in the initial prototype experiment motivated the development of an improved baseline ASR system that could be favourably benchmarked against other research-grade ASR systems. To improve performance a BNC/MapTask-derived bigram language model was implemented and MAP acoustic adaptation was undertaken. The vocabulary was also restricted to the 1018 vocabulary from the spoken

[1] Chapter 4

part of the eye/speech corpus. The improved baseline ASR system, together with benchmarks, was described in Chapter 4.

The method for integrating eye movement had some addressable flaws. The N-Best list was rescored based on word counts and not word probabilities. The temporal correspondence between eye movement and speech was considered on a per-segment basis rather than on a more desirable per-word basis. Addressing these shortcomings, the experiments in Chapter 6 adopted a more detailed integration scheme based on the integration theory in Chapter 2, using landmark-specific language models to change word probabilities. This enabled the rescoring the individual word probabilities using information from temporally corresponding eye movement.

Appendix C: Formula

Viterbi Reestimation

HMM

The reestimated value of state output PDF for state i, $\widetilde{b}_i(o)$, is calculated from observations in o belonging to state i in the viterbi decoded optimum state sequence, $\gamma_t(i)$:

$$\gamma_t(i) = \begin{cases} 1 & \text{if } \widehat{s} = i \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

$$\widetilde{b}_i(o) = f(\gamma_t(i), o) \quad (2)$$

Where f is a function which calculates the PDF based on the state sequence and observations E.g. a Gaussian PDF.

Reestimation of π, if desirable, is given by $\gamma_1(i)$:

$$\widetilde{\pi}_i = \begin{cases} 1 & \text{if } \gamma_1(i) = 1 \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

In a similar fashion, the uncovered state sequence allows A to be re-estimated by calculating the number of state transitions as a proportion of all transitions, T, in the state sequence[2]:

$$\psi_t(i,j) = \begin{cases} 1 & \text{if } \gamma_t(i) = 1 \text{ and } \gamma_{t+1}(j) = 1 \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

$$\widetilde{a}_{ij} = \frac{\sum_{t=1}^{T} \psi_t(i,j)}{T} \quad (5)$$

[2]Note that for a HMM $T = \sum_i \sum_j \sum_{t=1}^{T} \psi_t(i,j)$

HSMM

Reestimation expressions 1 thru 3 for HMM are the same for the HSMM.

A and $p_i(d)$ are reestimated by calculating the number of state transitions as a proportion of all transitions, T[3], in the state sequence:

$$\psi_t(i,j) = \begin{cases} 1 & \text{if } \gamma_t(i) = 1 \text{ and } \gamma_{t+1}(j) = 1 \\ 0 & \text{otherwise} \end{cases} \quad (6)$$

$$a_{ij} = \frac{\sum_{t=1}^{T} \psi_t(i,j)}{T} \text{for } i \neq j \quad (7)$$

$$p_i(d) = g(\psi_1(i,i), \ldots, \psi_T(i,i)) \quad (8)$$

Where g is a function which calculates the duration PDF based on the state transitions E.g. a Gaussian PDF.

[3]Note that $T = \sum_i \sum_j \sum_{t=1}^{T} \psi_t(i,j)$

Bibliography

[AB04] Smart Eye AB. Smart-eye white paper. Technical report, Smart Eye AB, september 2004.

[ABB+91] A H Anderson, M Bader, E G Bard, E H Boyle, G M Doherty, S C Garrod, S D Isard, J C Kowtko, J M McAllister, J Miller, C F Sotillo, H S Thompson, and R Weinert. The hcrc map task corpus. *Language and Speech*, 34(4):351–366, 1991.

[AH91] M.D.Bedworth A.J.R. Heading. Data fusion for object classification. *IS2 Pattern Processing Principles*, pages 837–840, 1991.

[Ass99] International Phonetic Association. *Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet*. Cambridge University Press, 1999.

[Ata74] B. S. Atal. Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. *The Journal of the Acoustical Society of America*, 55(6):1304–1312, 6 1974.

[Bae02] Christophe Van Bael. Using the keyword lexicon for speech recognition. In *Masters thesis*, pages 38–43. Department of Theoretical and Applied Linguistics, University of Edinburgh, UK, 2002.

[BD99] J A Brefczynski and E A DeYoe. A physiological correlate of the 'spotlight' of visual attention. *Nature Neuroscience*, 2:370–374, 1999.

[BGWL01] C. Barras, Z. Geoffrois, Z. Wu, and M. Liberman. Transcriber: development and use of a tool for assisting speech corpora production. In *Speech Communication*, volume 33, pages 5–22, 1 2001.

[Bil02] Jeff Bilmes. What hmms can do. Technical Report UWEETR-2002-0003, Dept of EE, University of Washington, 01 2002.

[BJP+00] C. Benoit, J.C.Martin, C. Pelachaud, L. Schomaker, and B. Suhm. Audio-visual and multimodal speech systems. *Handbook of Standards and Resources for Spoken Language Systems 2nd Ed.*, 2000.

[BK03] Christophe Van Bael and Simon King. The keyword lexicon - an accent-independent lexicon for automatic speech recognition. In *Proc. ICPhS*, pages 1165–1168, 2003.

[Bli93] J. F. Blinn. What's that deal with the dct? *IEEE Computer graphics and applications*, 4(13):78–83, 1993.

[BM01] Chris Baber and Brian Mellor. Using critical path analysis to model multimodal human-computer interaction. *Int. J. Human-Computer Studies*, 54:613–636, 2001.

[Bol84] R. A. Bolt. *The Human Interface*. Lifetime learning Publications, 1984.

[Bow74] T. G. R. Bower. The evolution of sensory systems. *Perception: Essays in honour of James J Gibson*, 5:141–154, 1974.

[Bur00] L Burnard. Users reference guide for the british national corpus. In *Technical Report*. Oxford University Computing Services, 2 2000.

[BW86] H Bourlard and C J Wellekens. Connected speech recognition by phonemic semi-markov chains for state occupancy modelling. In *Proc. European Signal Processing Conference (EUSIPCO-86)*, The Hague, 1986. Philips Research Laboratory.

[CAR04] Jeff B. Pelz Constatin A. RothKopf. Head movement estimation for wearable eye tracker. In *Eye Tracking Research & Applications Symposium 2004*. ACM SIGGRAPH, New York: ACM Press, 2004.

[Coo02] Neil Cooke. Statistical models of eye movement. In *Masters thesis*. Department of Electronic, Electrical and Computer Engineering, The University of Birmingham, UK, 2002.

[CR97] Phillip R. Clarkson and A. J. Robinson. Language model adaptation using mixtures and an exponentially decaying cache. In *ICASSP*, 1997.

[CR98] Phillip R. Clarkson and A. J. Robinson. The applicability of adaptive language modelling for the broadcast news task. In *Proceedings 5th International Conference on Spoken Language Processing, Sydney, Australia*, 1998.

[CR99] Phillip R. Clarkson and A. J. Robinson. Towards improved language model evaluation measures. In *Proceedings of EUROSPEECH 99*, 1999.

[CR05] Neil Cooke and Martin Russell. Using the focus of visual attention to improve automatic speech recognition. In *INTERSPEECH '2005 - 9th European Conference on Speech Communication and Technology, Lisbon, Portugal*, 9 2005.

[CRM04] Neil Cooke, Martin Russell, and Antje Meyer. Evaluation of hidden markov models robustness in uncovering focus of visual attention. In *ETRA 2004. Eye Tracking Research and Applications Symposium, San Antonio, Texas*, page 56. ACM SIGGRAPH, 4 2004.

[CS89] Yen-Lu Chow and Richard Schwartz. The n-best algorithm: an efficient procedure for finding top n sentence hypotheses. In *HLT '89: Proceedings of the workshop on Speech and Natural Language*, pages 199–202, Morristown, NJ, USA, 1989. Association for Computational Linguistics.

[Daw92] A. Dawid. Applications of a general propagation algorithm for probabilistic expert systems. *Statistical Computing*, 2:25–36, 1992.

[Dju02] Petar M. Djuric. An mcmc sampling approach to estimation of nonstationary hidden markov models. *IEEE Transactions on Signal Processing*, page 1113, 5 2002.

[DM80] S. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 28(4):357–366, 8 1980.

[Duc03] Andrew T Duchowski. *Eye Tracking Methodology. Theory and Practice*. Springer-Verlag, London, 2003.

[ea02] Steve Young et al. *The HTK Book (for Version 3.2.1)*. Cambridge University Engineering Department, 2002.

[Fis04] John W. Fisher. Speaker association with signal-level audiovisual fusion. *IEEE Transactions on Multimdeedia*, 6(3):406–412, 6 2004.

[Fur86] Sadaoki Furui. Speaker-independent isolated word recognition using dynamic features of speech spectrum. *IEEE Transactions onAcoustics, Speech, and Signal Processing*, 34(1):52–59, 2 1986.

[GB00] Zenzi M Griffin and K Bock. What the eyes say about speaking. *Psychological Science*, 11:274–279, 2000.

[GBY+00] Ralph Gross, Michael Bett, H. Yu, Xu Zhu, Y. Pan, Jie Yang, and Alex Waibel. Towards a multimodal meeting record. In *Proceedings of the 2000 IEEE International Conference on Multimedia and Expo (ICME 2000)*, volume 3, pages 1593 – 1596, July 2000.

[GEN95] A. Glenstrup and T. Engell-Nielsen. Eye controlled media: Present and future state. `http://www.diku.dk/~panic/eyegaze/article.html`, 1995.

[GF93] J. Garofolo and J. Fiscus. Speech header resources (sphere) version 2.2. In *Documentation file sphere.doc*, 1993.

[GHM92] J. Godfrey, E. Holliman, and J. McDaniel. Switchboard: Telephone speech corpus for research and development. In *Proc. Int'l Conf. on Acoustic, Speech and Signal Processing (ICASSP)*, volume 1, pages 517–520, 3 1992.

[GJ95] Zoubin Ghahramani and Michael I Jordan. Factorial hidden markov models. Technical Report 9502, Massachusetts Institue of Technology, 05 1995.

[GJ97] Zoubin Ghahramani and Michael I Jordan. Factorial hidden markov models. In Padhraic Smyth, editor, *Machine Learning*, volume 29, pages 245–247. Kluwer Academic Publishers, Boston., 1997.

[GLF+93] J.S. Garofolo, L.F. Lamel, W.M. Fisher, J.G. Fiscus, D.S. Pallett, and N.L. Dahlgren. Darpa timit acoustic-phonetic continuous speech corpus. In *CDROM*. Dept. of Commerce, NIST, 2 1993.

[Gmb99] SensoMotoric Instruments GmbH. *EyeLink System Documentation*. SR Research Ltd, Teltow, Germany, 1999.

[GNU] The free software foundation. `http://www.fsf.org`.

[Goo53] I. J. Good. The population frequencies of species and the estimation of population parameters. *Biometrika*, 40(16):237–264, 1953.

[GS04] Zenzi M Griffin and D S Spieler. Observing the what and when of language production by monitoring speakers eye movements. *Brain and Language - special issue*, 2004.

[GW96] M. J. F. Gales and P. C. Woodland. Mean and variance adaptation within the mllr framework. In *Computer Speech & Language*, volume 10, pages 249–264, 1996.

[GWKB04] E. Gouws, K. Wolvaardt, N. Kleynhans, and E. Barnard. Appropriate baseline values for hmm-based speech recognition. *Proceedings of the 15th Annual Symposium of the Pattern Recognition Association of South Africa*, page 169, 11 2004.

[Hel61] Joseph Heller. *Catch-22*. New York: Simon and Schuster, 1961.

[Hen03] John M Henderson. Human gaze control during real-world scene perception. *Trends in Cogntive Sciences*, 17(11):498–505, 11 2003.

[HH01] John Holmes and Wendy Holmes. *Speech Synthesis and Recognition 2nd Edition*. Taylor & Francis, 2001.

[HJ98] L R Harris and M Jenkin. *Vision and Action*. Cambridge University Press, New York, 1998.

[HKPH03] E. Horvitz, C. M. Kadie, T. Paek, and D. Hovel. Models of attention in computing and communications: From principles to applications. *Communications of the ACM*, 46(3):52–59, 3 2003.

[HS04] K. Himley and J. A. Schmidt. *Opthamolgisches bibliothek*. Froman Germany: Jena, 1804.

[Hun87] Melvyn J Hunt. Delayed decisions in speech recognition - the case of formants. *Pattern Recognition Letters*, 6:121–137, 07 1987.

[HW97] J.J. Humphries and P.C. Woodland. Using accent-specific pronunciation modelling for improved large vocabularly continuous speech recognition. In *Proc. Eurospeech*, 1997.

[Jac03] Robert Jacob. Eye tracking in human-computer interaction and usability research: Ready to deliver the promises (section commentary). `citeseer.ist.psu.edu/538753.html`, 2003.

[Jai00] Anil K Jain. Statistical pattern recognition: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(1):4–36, January 2000.

[Jam90] W James. *The Principles of Psychology Vol. I*. Henry Holt, New York, 1890.

[JPG] Jpgvideo video utility. `http://www.ndrw.co.uk/free/jpgvideo/index.html`.

[Jun93] Jean-Claude Junqua. The lombard reflex and its role on human listeners and automatic speech recognizers. *The Journal of the Acoustical Society of America*, 93(1):510–524, 1993.

[KAM84] E. Kowler, A.J.Martins, and M.Pavel. The effect of expectations on slow oculomotor control-iv. anticipatory smooth eye movements depend on prior target motions. In *Vision Research*, volume 24, pages 197–210, 1984.

[Kat87] S. M. Katz. Estimation of proabilities from sparse data for the language model component of a speech recogniser. *IEEE Transactions on Acoustic, Speech and Signal Processing*, 35(3):400–401, 1987.

[Kau03] Manpreet Kaur. Where is 'it'? event synchronization in gaze-speech input systems. In *ICMI 2003. Proceedings of the 5th International Conference on Multimodal Interfaces, Vancouver, British Columbia, Canada*. ACM, 2003.

[KHDM98] Josef Kittler, Mohamad Hatef, Robert P.W. Duin, and Jiri Matas. On Combining Classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(3):226–239, March 1998.

[Kun02] Lumila I. Kuncheva. Switching between Selection and Fusion in Combining Classifiers: An experiment. *IEEE Transactions on Systems, Man, and Cybernetics - Part B:Cybernetics*, 32(2):146–156, April 2002.

[LAK+04] Benjamin Law, M. Stella Atkins, A.E. Kirkpatrick, Alan J. Lomax, and Christine L. MacKenzie. Eye gaze patterns differentiate novice and experts in a laparoscopic surgery training environment. *Proceedings of the Eye Tracking Research and Applications Symposium 2004*, page 41, 04 2004.

[Lat88] C R Latimar. Cumulative fixation time and cluster analysis. *Behaviour Research Methods, Instruments and Computers*, 20:437–470, 1988.

[Lev86] S E Levinson. Continuously variable duration hidden markov models for automatic speech recognition. *Computer Speech and Language*, 1:29–45, 1986.

[LP94] C. J. Leggetter and P.C.Woodland. Speaker adaptation of continuous density hmms using linear regression. In *Proceedings of the ICSLP 2*, pages 451–454. ICSLP, 1994.

[LRM99] Willem J M Levelt, Ardi Roelofs, and Antje S Meyer. A theory of lexical access in speech production. In *Behavioral and Brain Sciences*, volume 22, pages 1–75. Cambridge University Press, 1999.

[LRS82] S E Levinson, L R Rabiner, and M M Sondhi. An introduction to the application of the theory of probablistic functions of a markov process to automatic speech recognition. *The BELL System Technical Journal*, 62(4), 04 1982.

[Mac03] David MacKay. *Information theory, Inference and Learning Algorithms*. Cambridge University Press, Cambridge, 2003.

[Mat74] E. Matin. Saccadic suppression: A review. In *Psychological Bulletin*, volume 81, pages 899–917, 1974.

[Mat02] The MathWorks. *MATLAB and Simulink Users Guide Version 7*. Prentic Hall, USA, 2002.

[MKJ02] Majaranta and Kari-Jouko. Twenty years of eye typing: systems and design issues. In *ETRA '02: Proceedings of the 2002 symposium on Eye tracking research & applications*, pages 15–22, New York, NY, USA, 2002. ACM Press.

[MM67] N Mackworth and A Morandi. The gaze selects informative details within pictures. *Perception and Psycholinguistics*, 1967.

[MML01] Femke Can Der Meulen, A Meyer, and W Levelt. Eye movements during the production of nouns and pronouns. *Memory and Cognition*, 29:512–521, 2001.

[MRL03] A.S. Meyer, A. Roelofs, and W.J.M. Levelt. Word length effects in picture naming: The role of a response criterion. *Journal of Memory and Language*, 47:131–147, 2003.

[MS04] Stephanie McMains and David Somers. Multiple spotlights of attentional selection in human visual cortex. *Neuron*, 42:677–686, 05 2004.

[MSL98] Antje S Meyer, Astrid M Sleiderink, and Willem J M Levelt. viewing and naming objects: eye movements during noun phrase production. In *Cognition*, volume 66, pages B25–B33. Elsevier Science B.V., 1998.

[Mur96] Robin R. Murphy. Biological and Cognitive foundations of Intelligent Sensor Fusion. *IEEE Transactions on Systems, Man, and Cybernetics - Part A:Systems and Humans*, 26(1):42–51, January 1996.

[Mur98] Kevin Murphy. A brief introduction to graphical models and bayesian networks. `www.cs.berkely.edu/~murphyk/Bayes/bayes.html`, 1998.

[MvdMB04] A.S. Meyer, F. van der Meulen, and A. Brooks. Eye movements during speech planning: Speaking about present and remembered objects. *Visual Cognition*, 11:553–576, 2004.

[NIS91] NIST. The road rally word-spotting corpora (rdrally1). `http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC93S11`, 9 1991.

[Noc02] H. J. Nock. Assessing face and speech consistency for monologue detection in video. *Multimedia 02, December 1-6, Juan-les-Pins, France*, pages 406–412, 12 2002.

[NS70] D Norton and L Stark. Scanpaths in saccadic eye movements when viewing and recognizing patterms. *Vision Research*, 11:929–942, 1970.

[NY00] H. J. Nock and S. J. Young. Loosely coupled hmms for asr. *Proc. ICSLP Beijing China*, 22(8), 2000.

[Ovi99a] Sharon Oviatt. Ten myths of multimodal interaction. *Communications of the ACM*, 42(11):74–81, 1999.

[Ovi99b] Sharon L. Oviatt. Mutual disambiguation of recognition errors in a multimodel architecture. In *Proceedings of Conference on Human Factors in Computing Systems: CHI '99*, pages 576–583. ACM Press, 1999.

[Ovi02] Sharon L. Oviatt. Breaking the robustness barrier: Recent progress on the design of robust multimodal systems. In *Advances in Computers*, volume 56, pages 305–333. Elsevier, 2002.

[Pad02] Murkund Padmanabhan. Large-vocabulary speech recognition algorithms. *Computer*, 35(4):42–50, 2002.

[PB92] D. Paul and J. Baker. The design for the wall street journal-based csr corpus. In *Proc. DARPA Speech and Nautral Language Workshop*, 2 1992.

[Pea88] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, San Francisco, California, 1988.

[Pea01] Judea Pearl. *Causality. Models, Reasoning, and inference*. Cambridge University Press, Cambridge, UK, 2001.

[Pos80] Michael I Posner. Orienting of attention. *Quarterly Journal of Experimental Psychology*, 32:3–25, 1980.

[PP90] Michael I Posner and Steven E Peterson. The attention system of the human brain. *Annual Review of Neuroscience*, 13:25–42, 1990.

[PS00] C. M. Provitera and L. W. Stark. Algorithms for defining visual regions-of-interest: comparison with eye fixations. *IEEE Transactions on Pattern Analysis and Machine Vision*, 22(9):970–982, 2000.

[Rab89] Lawrence R Rabiner. A tutorial on hidden markov models and selected application in speech recognition. In *Proceedings of the IEEE*, volume 77, pages 257–286. IEEE, IEEE, 02 1989.

[Ray98] Keith Rayner. Eye movements in reading and information processing - 20 years of research. *Psychological Bulletin*, 124(3):372–422, 1998.

[RFP+95] T. Robinson, J. Fransen, D. Pye, J. Foote, and S. Renals. Wsjcam0: A british english speech corpus for large vocabulary continuous speech recognition. In *Technical Report 0-7803-2431-5/95*. Cambridge University Engineering Department, 1995.

[Ric86] Alan Richter. Modelling of continuous speech observations. In *Advances in Speech Processing Conference*. IBM Europe Institute, 1986.

[RM85] Martin J Russell and R K Moore. Explicit modelling of state occupancy in hidden markov models for automatic speech recognition. In *Proceedings International Conference on Acoustics Speech and Signal Processing (ICASSP 85)*, pages 1.2.1–1.2.4. Tampa, 1985.

[RM05] Deb Roy and Niloy Mukherjee. Towards situated speech understanding: visual context priming of language models. *Computer, Speech and Language*, 19:227–248, 2005.

[RZHB97] Rajesh P. N. Rao, Gregory J. Zelinsky, Mary M. Hayhoe, and Dana H. Ballard. Eye movements in visual cognition: A computational study. Technical Report NRL97.1, University of Rochester, 1997.

[Sal99] Dario D Salvucci. *Mapping eye movements to cognitive processes(Tech. Rep. No. CMU-CS-99-131)*. Doctoral dissertation, Carnegie Mellon University, 1999.

[Sal00] Dario D Salvucci. Intelligent gaze-added interfaces. In *Human Factors in Computing Systems:CHI 2000 Conference Proceedings.* New York: ACM Press, 2000.

[Sat01] M. Satyanarayanan. Pervasive computing: Vision and challenges. *IEEE Personal Communications*, pages 10–17, August 2001.

[SB96] Peter L Silsbee and Alan C Bovik. Computer lipreading for improved accuracy in automatic speech recognition. *IEEE Transactions on Speech and Audio Processing*, 4(5):337–351, 09 1996.

[SD04] Anthony Santella and Doug DeCarlo. Robust clustering of eye movement recordings for quantification of visual interest. *ETRA 2004 Eye tracking research & Applications Symposium 2004*, pages 27–34, 04 2004.

[SH97] Ramesh R Sarukkai and Craig Hunter. Integration of eye fixation information with speech recongition systems. *EUROSPEECH*, pages 1639–1643, 1997.

[SJ99] Lawrence K Saul and M.I. Jordan. Mixed memory markov models: decomposing complex stochastic processes as mixtures of simpler ones. In Padhraic Smyth, editor, *Machine Learning*, volume 37, pages 75–87, 1999.

[SPH98] Rajeev Sharma, Valdimir Pavlovic, and Thomas Huang. Toward multimodal hci. *Proceedings of the IEEE*, 86(5):853–869, May 1998.

[ST00] S.K. Schnipke and M.W. Todd. Short talk, trials and tribulations of using an eye-tracking system. In *CHI 2000*, 2000.

[SVN37] S. S. Stevens, John Volkmann, and E.B. Newman. A scale for the measurement of the psychological magnitude of pitch. In *The American Journal of Psychology*, number 8 in 12, pages 185–190, 1937.

[SWS99] Barry Stein, Mark Wallace, and Terrence Stanford. Development of multisensory integration: Transforming sensory input into motor output. *Mental Retardation and Developmental Disabilities Research Reviews*, 5(1):72–85, 1999.

[SYW97] R. Stiefelhagen, J. Yang, and A. Waibel. Tracking eyes and monitoring eye gaze. In *Proceedings of the Workshop on Perceptual User Interfaces (PUI'97), Alberta, Canada.*, pages 98–100, 1997.

[TC00] Kari Torkkola and William M. Campbell. Mutual information in learning feature transformations. In *Proc. 17th International Conf. on Machine Learning*, pages 1015–1022. Morgan Kaufmann, San Francisco, CA, 2000.

[TKB92] K. R. Thorisson, D. R. Koons, and R. A. Bolt. Multimodal natural dialogue. *Proceedings of CHI92*, pages 653–654, 1992.

[Tom96] Michael John Tomlinson. *A study into the Audio and Visual Integration of Speech for Automatic Recognition.* PhD dissertation, University of Bath, 1996.

[Vir] Virtualdub video capture/processing utility. `http://www.virtualdub.org/`.

[Wan95] Jian Wang. Integration of eye-gaze, voice and manual response in multimodal user interface. *IEEE International Conference Systems, Man and Cybernetics.*, 5:3938–3942, 10 1995.

[Wei99] Mark Weiser. The computer for the 21st century. *SIGMOBILE Mob. Comput. Commun. Rev.*, 3(3):3–11, 1999.

[WOC00] Lizhong Wu, Sharon L Oviatt, and Phillip R Cohen. Multimodal integration - a statistical view. In *IEEE Transactions on Multimedia*, volume 1 of *4*, pages 334–341. IEEE, 12 2000.

[Woo02] D S Wooding. Fixation maps: quantifying eye-movement traces. *Proceedings of the Eye Tracking Research and Applications (ETRA)*, pages 31–36, 2002.

[YBZ02] Chen Yu, Dana H Ballard, and Shenghuo Zhu. Attentional object spotting by integrating multimodal input. In *Proceedings of the Fourth IEEE International conference on Multimodal Interfaces (ICMI'02).* IEEE Computer Society, IEEE, 2002.

[You95] S. Young. Large vocabulary continuous speech recognition: A review. In *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 3–28. IEEE, 12 1995.

[YOW94] S. Young, J. Odell, and P. Woodland. Tree-based state tying for high accuracy acoustic modelling. *Proceedings ARPA Workshop on Human Language Technology*, pages 307–312, 1994.

[YV02] Q. Yan and S. Vaseghi. A comparative analysis of uk and us english accents in recognition and synthesis. In *ICASSP*, 2002.

[ZIGM04] Q. Zhang, A. Imamiya, K. Go, and X. Mao. Designing a robust speech and gaze multimodal system for diverse users. In *IEEE International conference on Information Reuse and Integration 2003*, pages 354–361. SIGGRAPH, 2004.