# Brains in Dialogue

Investigating accommodation in live conversational speech for both speech and EEG data.

**Vijay James Solanki**
**MA(Hons)**

June, 2017

Submitted in fulfilment of the requirements for the Degree of Doctor of Philosophy


School of Critical Studies
College of Arts
University of Glasgow

**Abstract**

One of the phenomena to emerge from the study of human spoken interaction is *accommodation* or the tendency of an individual's speech patterning to shift relative to their interlocutor. Whilst the experimental approach to the detection of accommodation has a solid background in the literature, it tends to treat the process of accommodation as a black box. The general approach for the detection of accommodation in speech has been to record the speech of a given speaker prior to interaction and then again after an interaction. These two measures are then compared to the speech of the interlocutor to test for similarity. If the speech sample following interaction is more similar then we can say that accommodation has taken place. Part of the goal of this thesis is to evaluate whether it is possible to look into the black box of speech accommodation and measure it 'in situ'.

Given that speech accommodation appears to take place as a result of interaction, it would be reasonable to assume that a similar effect might be observable in other areas contributing to a communicative interaction. The notion of an interacting dyad developing an increased degree of alignment over the course of an interaction has been proposed by psychologists. Theories have posited that alignment occurs at multiple levels of engagement, from broad levels of syntactic alignment down to phonetic levels of alignment. The use of speech accommodation as an anchor with which to track the evolution of change in the brain signal may prove to be one approach to investigating the claims made by these theories. The second part of this thesis aims to evaluate whether the phenomenon of accommodation is also observable in the form of electrical signals generated by the brain, measured using Electroencephalography (EEG). However, evaluating the change in the EEG signal over a continuous stretch of time is a hurdle that will need to be tackled. Traditionally, EEG methodologies involve averaging the signal over many repetitions of the same task. This is not a viable option when investigating communicative interaction.

Clearly the evaluation of accommodation in both speech and brain activity, especially for continuously unfolding phenomena such as accommodation, is a non-trivial task. In order to tackle this, an approach from speech recognition and computer science has been employed. The implementation of Hidden Markov Models (HMM) has been used to develop speech recognition systems and has also been used to detect fraudulent attempts to imitate the voice of others. Given that HMMs have successfully been employed to detect the imitation of another person's speech they are a good candidate for being able to detect the movement towards or away from an interlocutor during the course of an interaction. In addition, the use of HMMs is non-domain specific, they can be used to evaluate any time-variant signal. This

adaptability of the approach allows for it to also be applied to EEG signals in conjunction with the speech signal.

Two experiments are presented here. The behavioural experiment aims to evaluate the ability of a HMM based approach to detect accommodation by engaging pairs of female, Glaswegian speakers in the collaborative DiapixUK task. The results of their interactions are then evaluated from both a traditional phonetic standpoint, by assessing changes in Voice Onset Time (VOT) of stop consonants, formant values of vowels and speech rate over the course of an interaction and using the HMM based approach. The neural experiment looks to evaluate the ability of a HMM based approach to detect accommodation in both the speech signal and in brain activity. The same experiment that was performed in Experiment 1 was repeated, with the addition of EEG caps to both participants. The data was then evaluated using the HMM based approach.

This thesis presents findings that suggest a function for speech accommodation that has not been explored in the past. This is done through the use of a novel, HMM based, holistic acoustic-phonetic measurement tool which produced consistent measures across both experiments. Further to this, the measurement tool is shown to have possible extended uses for EEG data. The use of the presented HMM based, holistic-acoustic measurement tool presents a novel contribution to the field for the measurement and evaluation of accommodation.

# Contents

# List of Tables

# List of Figures

# Acknowledgements

This project was funded by the University of Glasgow's Lord Kelvin-Smith PhD scholarship programme. I would like to thank the following people for their help over the course of this project:

Firstly, I would like to thank all of my supervisors for their support, guidance and patience. I would like to thank **Prof. Jane Stuart-Smith** for regularly encouraging me to move outside of my comfort zone and for supporting me throughout the various struggles of this project. Thanks also must be made to **Dr. Rachel Smith** who has provided me with many excellent discussions that have helped me to really get to grips with the content of this thesis. **Dr. Alessandro Vinciarelli** has provided me with many hours of expert tuition and has instilled in me a new found passion for problem solving and critical evaluation, thank you. Without the foresight of **Prof. Pascal Belin** this project would not have been possible and I would like to thank him for his inital support and engagement with this work. The application of EEG to this project was only made possible through the help and guidance of **Dr. Guillaume Rousellet**, I would like to thank him for his expert guidance and openness to our non-standard EEG analysis approach. Whilst not technically a supervisor, **Dr. Ludger Evers** has been invaluable in the completion of this project and I would like to thank him for his contribution to the statistical analysis of the phonetic data.

Next, I would like to express my thanks to all of the labsters and post-docs at GULP for sharing this PhD process with me. Thank you to **Robert Lennon** for always being willing to discuss a complex problem and for not being afraid to challenge me on my views. I'd like to thank **Ewa Wanat** for her infectious personality and the brightness she brings to the lab. Thanks also go to **Dr. Farhana Alam** for showing me the ropes and for always making time for my questions. Finally, thank you to **Dr. Duncan Robertson**, **Dr. Brian José** and **Prof. Marc Alexander** for helping to get me out of the lab and for their guidance both regarding this project and life in general.

This project has been a task that I could have not completed without the help of a number of people that have provided practical and logistical support. First, I would like to thank all of the participants that took part in the experiments, without you this thesis could not have been written. I would also like to express my thanks to all of the transcribers on this project for their many hours of hard work. I would like to thank **Katia** for the time that she took to train me on the use and application of EEG systems. **Victoria Nichols** must also be thanked for her help in conducting the EEG experiments. A big thank you must also be expressed to **Adeline Callander** who was indespensible throughout the project and who would work so hard to aid me in navigating the intricacies of the university's administrative system. Finally, I

would like to thank both **Kiran Faisal** & **James Matthew** for all of their support in organising the nesessary particulars for the dissemination of my research.

I would also like to thank **Chris Forster** for his patience with me during the beginning of this project and for providing a welcome escape from the pressures of the thesis. **Garikoitz Lerma Usabiaga** deserves a mention to highlight the role that he played in helping me to integrate with the resaerch community and for providing a friendly face at many of the conferences that I attended. Thanks also go out to **Alison McNaughton** for her support in the final few weeks of the thesis write-up, those weeks would have been far more stressful without her upbeat personality and home baked banana bread. A special thanks is also made to **Marianne Alexander** who planted the seeds of my interest in linguistic research and who helped to nurture that interest throughout my time at secondary school and beyond.

Finally, I would like to say a hearfelt and sincere thank you to my family. Without their support, I don't imagine that I would have even been able to undertake this project to begin with. They have certainly helped me through a lot in the time that it has taken to complete this project and I cannot thank them enough for their support. In particular, I would like to say the biggest thank you of all to my loving and eternally patient partner, **Becca**. You have been the one constant across the most emotionally and intellectually turbulant period of my life. Thank you for your warmth, compassion and humour.

# Declaration

I declare that, except where explicit reference is made to the contribution of others, that this dissertation is the result of my own work and has not been submitted for any other degree at the University of Glasgow or any other institution.

Signature _____

Printed Name _____

# Dedication

For my grandparents.

# Abbreviations

List of definitions for the abbreviations used in this thesis:

| | |
|---|---|
| BCI | Brain computer interface |
| BFI | Big five personality inventory |
| DCT | Discrete cosine transformation |
| EEG | Electroencephalography |
| ERP | Event-related potential |
| F0 | Fundamental frequency |
| F1 | First formant frequency of a vowel |
| F2 | Second formant frequency of a vowel |
| fMRI | Functional magnetic resonance imaging |
| fNIRS | Functional near-infrared spectroscopy |
| GAM | Generalised additive model |
| HMM | Hidden markov model |
| HTK | Hidden markov model tool kit |
| IA | McCroskey interpersonal attraction questionnaire |
| ICA | Independent component analysis |
| IQR | Inter-quartile range |
| LSL | Lab streaming layer |
| MEG | Magnetoencephalography |
| MEM | Linear mixed-effects model |
| MFCC | Mel frequency cepstral coefficient |
| PSD | Power spectral density |
| VOT | Voice onset time |

# Chapter 1

# Introduction

This thesis is concerned with the role that the adaptation of speech sounds, often referred to as *accommodation*, plays in communicative understanding during dialogue. It presents novel methodologies and analysis approaches for the assessment of adaptation between speakers for both speech and neural signals. Findings validate the proposed approaches and suggest a link between speech adaptation and brain activity. What this thesis does not provide is a fully formed method of analysis. The contents of this thesis should be taken as preliminary evidence for possible alternative approaches to the evaluation of accommodation and the evaluation of links between speech based communicative phenomena and brain activity. This chapter offers brief overviews of the background surrounding the topic, the challenges that will need to be faced in order to meet the goals of the thesis and the approaches used to tackle them. It then goes on to give the main research question and details the contents of the rest of the thesis.

Since this thesis deals with both speech and brain activity, the fields of phonetics and psychology provide a great deal of the necessary background knowledge for an investigation of this type. Much of the work done on speech adaptation, or accommodation, can be found in the phonetic literature. There have been a number of studies that demonstrate the tendency of speakers to unconsciously shift their speech relative to the people that they engage with (Shockley, Sabadini, & Fowler, 2004; Pardo, 2006; Babel, 2009a; Bulatov, 2009; Bailly, Lelong, et al., 2010; Casasanto, Jasmin, & Casasanto, 2010; Pardo, Jay, & Krauss, 2010; Alshangiti & Evans, 2011; Lewandowski, 2012; Beňuš, 2014). A long held explanation for this behaviour is that it is linked to expression of social information (Giles, Mulac, Bradac, & Johnson, 1987; Giles, Coupland, & Coupland, 1991b; Gallois & Giles, 2015). That is to say that speech adaptation expresses and occurs in response to social information. This can include such things as affiliation with a particular social group (Babel, 2010), perceived closeness (Pardo, Gibbons, Suppes, & Krauss, 2012) or degree of liking (Yu et al., 2013), amongst a number of other types of social information.

Within the literature concerning the detection of speech adaptation, a term that is often used to describe this phenomenon is *accommodation*. This term encompasses

both shifting towards a speech partner and shifting away from a speech partner. However, there can be a tendency for the term *accommodation* or the usage of conjugations of the term (eg. "to accommodate") to be used to refer explicitly to a shifting towards another speaker (eg. Shockley et al., 2004, (pp.422)). Within the literature there are specific terms for each form of speech shifting. When a speaker becomes more similar to their speech partner, it is called *convergence*. The inverse of this, when a speaker becomes more dissimilar from their speech partner, is known as *divergence*. In the work presented in this thesis, the term *accommodation* is used to refer to any adaptation towards or away from a speech partner and the specific terms *convergence* and *divergence* will be used to indicate directionality.

In order to build a theory to underpin the observation of accommodative behaviours, those working in the speech sciences drew on concepts and theories proposed by social psychologists (Giles, Taylor, & Bourhis, 1973; Simons, Berkowitz, & Moyer, 1970; Byrne, 1971). The pairing of the speech sciences and psychology is somewhat parsimonious considering that language use serves a predominantly social function. Since these initial considerations regarding the links between speech accommodation and social information, there have been some considerable contributions from the field of psychology. In general, these contributions tend not to focus specifically on accommodation but rather provide considerations on the nature of the mechanisms that underlie effective communicative engagements. These considerations include processes that focus on alignment between speakers' mental representations of the topic at hand (Pickering & Garrod, 2004; Pickering & Garrod, 2013), the matching of vocal tract actions (Fowler, 1986; Fowler, 2014) and the manner in which memory traces from the perception of speech influences speech production (Goldinger, 1998). Broadly speaking, the theories and considerations presented by the psychological literature concerned with perception and memory systems involved in communicative acts provide support for the presence of accommodation in speech production (Pardo, Urmanche, Wilman, & Wiener, 2016a). Crucially though, a link between brain activity and the accommodation observed in speech has not been demonstrated. Being able to demonstrate a link such as this would go a long way to providing validation for these theories by way of a measurable neural correlate for their proposed mental mechanisms.

In this thesis, the relationship between the shifting that is observed in the speech signal and the theorised alignment of mental representations is assessed. The goal is to determine if accommodation goes beyond the expression of social information and actually contributes to the effectiveness of communicative understanding by facilitating greater speaker alignment as represented in measurable brain activity. Considering that the adjustments speakers make to their speech are, for the most part, produced unconsciously, it can be argued that they provide an implicit insight into the deeper processes of communicative understanding. In order to experimentally test this proposal, there are two main tasks that need to be performed.

The first of these tasks concerns reassessing the way in which speech adaptation is measured. Phonetic approaches tend to view the speech signal as chunked into its linguistically meaningful constituent parts. This transforms the speech signal into a more discrete signal source. What is meant by this is that whilst the local features within a phonetic unit retain their spectral properties within the time that they unfold, they are not easily interpreted in relation to the broader, ongoing trends in the speech signal. Since this thesis is concerned with ongoing brain activity as well as speech during accommodation and considering that aligning mental representations is a continuously unfolding phenomenon, the classification of speech adaptation will also need to be continuous. Currently, the detection of speech adaptation does not tend to use continuous measures for the classification of accommodation. The work presented here draws upon approaches utilising Hidden Markov Models (HMMs) that have been successfully applied to the problem of automatic speech recognition, which is based on analysis of the speech stream, to address this.

The other main task that needs to be performed is to develop an approach that can evaluate the link between speech adaptation and alignment of mental representations. Currently there are no established methods with which to accomplish this. The method used to address the previous issue surrounding speech adaptation is domain independent, meaning that it can be applied to any continuous signal, it is not limited to just the speech domain. The fact that the approach is domain independent can be capitalised upon here to make an assessment of alignment in the neural signals of the speakers. This approach is used to perform the task of evaluating the link between speech adaptation and the alignment of mental representations.

This thesis presents work that aims to tie together research on speech adaptation and alignment of neural representations through the use of HMM based tools developed for automatic speech recognition. The remainder of this chapter expands on the specific challenges facing both of the above tasks and how they are addressed in this work.

## 1.1 Measuring speech accommodation

One of the core goals of this thesis is the creation of a measure that can detect speech adaptation given the continuous speech signal of a speaker engaged in an interaction with another person. Further to this, the method developed should also be capable of classifying speaker interactions based on the level of adaptive behaviour (ie. being able to classify the interaction as being generally convergent, divergent etc.). This section presents a brief overview of the challenges facing the creation of such a measure. The topics presented here are further elaborated upon in chapter 2.

The key issues facing the creation of a continuous measure of accommodation stem from the way in which accommodation has been studied in the literature. Studies looking at accommodation have provided an excellent grounding for the inves-

tigation of this phenomenon but in doing so have developed methods of evaluation that may not be wholly compatible with a continuous evaluation of accommodation from an analysis of the speech stream.

The first issue to be tackled concerns the nature of the approaches used to evaluate accommodation. In general, the measures that are used for work on accommodation can be classified as either *perceptual* or *acoustic-phonetic*. The literature tends to provide evaluations of accommodation either from a perceptual standpoint, where judgements of accommodation are explicitly made by human listeners (eg. Namy, Nygaard, & Sauerteig, 2002; Pardo, 2006; Alshangiti & Evans, 2011), or from an acoustic-phonetic standpoint, where measures are drawn from the speech signal (eg. Shockley et al., 2004; Babel, 2010; Bailly & Martin, 2014). Both of these approaches have provided valuable insights and have helped to establish the cornerstones of research into accommodation (see section 2.2) but they do so from different ends of the analysis spectrum. Perceptual approaches consider the whole accommodative process from a holistic point of view, accounting for all elements of the speech signal that play a role in accommodation, thanks to the ability of the human perceptual system to do this. However, they can only offer insights into the specific features of the speech signal that influence the phenomenon through inference. Acoustic-phonetic approaches on the other hand, are not perceptual and are able to characterise individual aspects of the speech signal that are involved in accommodation but they do not consider the potential interactions between phonetic measures that may also play a part in accommodation. For example, it may be the case that accommodation through a lengthening of a vowel may impact on a speaker's speech rate. The ultimate goal of this thesis is to evaluate the link between accommodation and brain activity, so the phenomenon of accommodation as a whole needs to be considered. In order to provide a measurable speech signal with which to correlate the signal from brain activity, the measure of accommodation will need to be continuous and acoustic-phonetic. This is the first issue facing the creation of an appropriate measure of accommodation for this thesis, how to reconcile acoustic-phonetic measures with a holistic viewpoint of accommodation.

The second issue facing the way in which accommodation is measured in this thesis concerns the elicitation of accommodative behaviour. Accommodation can either be elicited through exposure to pre-recorded speech or it can be elicited through live interactions between speakers. Again, this distinction allows for the literature on this topic to be broadly separated into two categories: *non-interactional*, using pre-recorded speech (eg. Namy et al., 2002; Nielsen, 2011; Yu et al., 2013) and *interactional*, using live speech (eg. Collins, 1998; Purnell, 2009; Kim, Horton, & Bradlow, 2011). The use of these two elicitation approaches allows researchers to investigate different aspects of accommodation. Using a non-interactional approach allows for clear isolation of selected variables, these can be phonetic, linguistic or social. However, it is one-sided, it only considers the accommodative behaviour of

one person. It does not offer an opportunity to investigate the ongoing role that the other member of an interactional dyad plays in accommodation. Nor does it allow for the investigation of accommodation as a whole since the stimuli that are used to elicit accommodation are controlled in order to elicit specific, singular, speech features. Interactional approaches on the other hand allow for accommodation to be studied in situ. Levels of accommodation can be evaluated for both interactants and on the triggers underpinning accommodation. The consequence of this is that it is more difficult to interpret which factors are impacting on accommodation. Given that this thesis is interested in investigating how accommodation and brain activities of the speakers might be linked, it is necessary to take an interactional approach. However, in order to link accommodation and brain activity, markers for accommodation will need to be identified in the speech signal. Recording accommodation in an interactional setting whilst being able to identify markers for accommodative behaviour in the speech signal is the second issue facing the creation of an appropriate measure of accommodation for this thesis.

Finally, given the above two issues for this thesis the necessary approach required to address them will have to make use of both a combined interactional and acoustic-phonetic approach. The use of an interactional source to elicit accommodation allows for accommodation to be considered as a continuously unfolding phenomenon. That is to say that at any given point during a live interaction between two speakers, one may converge or diverge dependent on the speech of their partner. Whilst the general trend over the course of an interaction may be for that speaker to converge, there may be instances of divergence along the way. However, in order to make an assessment of this kind, a like-with-like comparison must be made. The use of acoustic-phonetic measures with interactional speech does hamper this somewhat. Allowing interactions between speakers to unfold naturally means that regular elicitation of a specific phonetic measure cannot be guaranteed. This makes assessing accommodation as a continuously unfolding phenomenon somewhat difficult with traditional methods. This is the final task for the creation of an appropriate measure of accommodation, to be able to consider accommodation as a continuously unfolding phenomenon using interactional speech and acoustic-phonetic measures.

The approach proposed for tackling these issues is based on work from the machine learning and automatic speech recognition community. The use of Hidden Markov Models (HMMs) is suggested as a possible tool to tackle the issues outlined above in the following ways:

- How does the proposed HMM based method reconcile the use of acoustic-phonetic measures with holistic approaches?

  Acoustic-phonetic measures are reconciled with holistic approaches through use of vectorised representations of the acoustic feature space. Quantitative measures of the speech signal at regular time-points are taken but are transformed to better represent what the human cochlea hears at that time-point

(Zwicker, 1961).

- How does the proposed method reconcile the use of interactional accommodation sources with the need for acoustic markers of accommodation?

  Models describing the general speech characteristics of each of the speakers in a dyad are created. Sections of the speech signal of each member of the dyad, taken at regular intervals, are compared to both speaker models to determine which model better accounts for that particular speech sample.

- How does the proposed method handle accommodation as a continuously unfolding phenomena?

  Rather than taking specific extracts of speech to evaluate for accommodation, the proposed approach chunks all speech into short, overlapping windows that are all time-tagged. When making an assessment of accommodation, this time-tagging can be taken into consideration to provide a view of accommodation levels as they unfold across an interaction.

In addition to allowing for accommodation to be evaluated in a manner consistent with the overall goal of evaluating the link between accommodation and brain activity, the proposed approach is domain independent. What is meant by this is that the approach can be applied to any signal that varies with respect to another dimension. It is this domain independence that allows for the proposed approach to be taken beyond the assessment of accommodation and to be applied to data sourced from brain activity.

## 1.2 Linking accommodation and brain activity

The second core goal of this thesis is to develop a preliminary way to assess if there is a link between speech accommodation and the alignment of observable brain activity in the speakers. Simply put, this goal of the thesis aims to ask if the accommodation observed in the speech signal of the speakers is also observable in the brain activity of the speakers. There are a number of theoretical stances that would support (and in some cases predict) the presence of such a linkage (see: Pardo et al., 2016a). However, the ability to make a direct assessment of this linkage has not yet been possible. Being able to make an assessment of this kind requires a number of challenges to be overcome. This section provides a brief overview of those challenges, providing a consideration of how theoretical implications, brain activity measurement technicalities and signal linkage technicalities contribute to these challenges. As with the previous section, section 1.1, the topics that are briefly discussed here are elaborated upon in chapter 2.

The justification for undertaking a study of the kind performed in this thesis must have theoretical backing. In this case, the phenomenon of accommodation must be

linked to the joint brain activity of interacting speakers. As outlined in the opening to this chapter, there is a close relationship between the fields of speech science and psychology. This relationship and the research produced as a result of it offers a series of potential explanations for both why accommodation manifests itself and for why this could be linked to joint brain activity. The key elements to reconcile in this literature are the stances taken on the level of automaticity in accommodation, the types of factor that modulate it and how these two things might impact on the representations in memory. Providing a theoretical backing for these three things will allow for an interpretation of mechanisms that drive accommodation, the function accommodation serves in effective communication and how a link between the brain activity of interacting speakers might come about. Taken together, it can be shown that there is good theoretical backing to assume a link between accommodation and joint brain activity (Dumas, Lachat, Martinerie, Nadel, & George, 2011; Giraud & Poeppel, 2012; Pickering & Garrod, 2013).

Given that the main focus of interest in this thesis is live interactional speech and the ongoing relationship between that and brain activity, there is a necessity for a tool that can measure both speech and brain activity as they unfold. A number of different tools for measuring the joint brain activity of interacting speakers were considered for the work presented in this thesis but the tool ultimately decided upon was Electroencephalography (EEG; eg. Babiloni et al., 2007; Kim, Lee, & Lee, 2014; Toppi et al., 2016). This particular way of measuring brain activity records the electrical activity produced by the brain and has the right balance of temporal resolution, portability and signal compatibility for use in this thesis. However, the integration of interactional speech and this method of measuring brain activity is not without its challenges.

The signal that is captured by EEG has a number of potential contaminants which can blur the electrical activity that is produced by the brain. One such contaminant comes from the electrical signal generated from the muscles used in producing speech (Vos et al., 2010; Porcaro, Medaglia, & Krott, 2015). The noise created by the production of speech is one of the main reasons that studies like the one presented in this thesis have not been more widely undertaken. The more traditional methods used in the capture, processing and analysis of EEG signals have often proved to be susceptible to these sources of electrical contamination. These traditional methods include studies investigating event-related potentials (ERPs) and the signal averaging techniques that accompany them (eg. Koelsch, Gunter, Wittfoth, & Sammler, 2005; Swaab, Ledoux, Camblin, & Boudewyn, 2012). The approach taken in this thesis differs somewhat from traditional approaches and makes use of the novel, domain independent, analysis technique developed for the interpretation of accommodation to help mitigate these effects to a certain degree. However, the implementation of the new approach presented in this thesis is not the only tool for overcoming this issue. The algorithms and signal detection approaches used to

clean the EEG signal have made considerable advances in recent years (Delorme & Makeig, 2004; Delorme, Sejnowski, & Makeig, 2007; Delorme et al., 2011; Bigdely-Shamlo, Mullen, Kothe, Su, & Robbins, 2015; Porcaro et al., 2015). This is capitalised upon in this thesis to further aid in increasing the detection of brain activity whilst minimising the influence of noise generated from the speech muscles.

As mentioned above, the method for interpretation of the EEG signal differs from traditional methods. Due to this, there are a number of considerations that need to be made in order to make the EEG signal compatible with the new analysis approach. One such consideration is the selection of an appropriate vectorisation parameter for the EEG signal. The application of the proposed approach already has some precedence for application to speech signals. However, the application to EEG signals lacks the same kind of precedence. In order to address this, a way to characterise the form of the EEG signal for each window of measurement used across the course of an interaction will need to be settled upon. This is specific to the assessment approach proposed and the method used to measure brain activity, in this case EEG. Being able to sensibly integrate the EEG signal with the approach used to assess accommodation in the speech signal is key to being able to evaluate any proposed link between accommodation and joint brain activity.

The more traditional approaches to interpreting EEG signals tend to focus on tying the EEG signal to clearly defined triggers (eg. Rousselet, Husk, Bennett, & Sekuler, 2008; Van Petten & Luka, 2012; Rousselet, Ince, van Rijsbergen, & Schyns, 2014). These studies can be conceptually considered to be similar to how a non-interactive approach to measuring accommodation might link a particular trigger to a speech sample. The measurement process is repeated many times for the same trigger across multiple participants in order to provide a strong, clear signal. Due to the nature of the phenomenon being investigated in this thesis, signal averaging in the same way as performed for ERP studies is not possible. This is due to the use of interactional speech which doesn't easily allow for triggers to be clearly defined. Further to this, interactional speech is inherently transient, it occurs in relation to the specific context at the moment in time, repetition of such speech would invalidate the use of interactional speech. Again, the domain independence of the proposed analysis approach allows this challenge to be overcome. The method that is proposed for capturing accommodation in the speech signal allows for the continuous capture of the phenomenon based on regular, overlapping time windows that are time-tagged. This same approach can be applied to the EEG signal. Both of these signals are captured during the same interaction and therefore provide the two signals necessary to link accommodation and brain activity. The speech signal provides the behavioural marker to which the EEG signal can be tied and because the same analysis approach can be applied to both signals, the time windows used for one signal can be used to temporally link the other.

This section has provided a brief overview of some of the challenges that are

faced by this thesis and has offered a cursory discussion of how they will be overcome. Chapter 2 deals with these topics is far more detail and provides information specific to the mechanisms of the approach taken. The next section provides the main research question for this thesis and details the structure of the thesis.

## 1.3 Main Research question and thesis outline

This chapter has provided a general introduction to the key themes in this thesis as well as providing a general overview of the challenges that will be faced. The overall theoretical research question for this study asks:

> Is speech accommodation linked to the alignment of mental representations as accounted for through observable brain activity?

This overall research question is then broken down into more specific research questions for the challenges presented at each stage of the thesis.

Chapter 2 presents a review of relevant literature regarding accommodation, alignment of mental representations and the use of computational approaches for social signal processing. It also outlines some important theoretical considerations for the work presented here. Finally, it situates the content of this thesis in relation to that work.

Chapters 3 and 4 break down this research question into specific elements. Chapter 3 deals with some of the experimental issues that need to be overcome in order to address the above research question. The implementation of a participant selection paradigm aimed at maximising the factors that lead to greater accommodative behaviour is evaluated. Further to this, it provides two approaches for assessing accommodation over the course of a continuous interaction. One approach is based on an adaptation of standard phonetic analysis techniques. The other provides a HMM-based approach to the problem. These two methods for measuring continuous accommodation are evaluated and the resulting findings feed into the work of the next chapter.

Chapter 4 delves into the possible relationships between speech accommodation and the neural patterning in participants. It offers a replication of the HMM-based analysis approach for the detection of speech accommodation in a continuous interaction that was presented in chapter 3. In addition to this, it provides an extension of the HMM-based approach to the evaluation of EEG signal patterning between interacting participants.

Chapter 5 draws together the work presented, makes some observations regarding its findings and offers considerations about possible theoretical implications. It also provides some conclusions that can be drawn from the work presented in this thesis and suggests areas for further investigation.

# Chapter 2

# Literature Review

As outlined in chapter 1, the main research question that this thesis asks is if accommodation has any correlates with the alignment of brain activity between speakers. This chapter aims to provide the background information required to evaluate this question by answering a number of '*What*, *Why* and *How*' questions related to the topic. Specifically, it will be answering:

- *What* is accommodation?
- *How* is accommodation measured?
- *Why* should accommodation be linked to joint brain activity?
- *How* can accommodation and brain activity be measured in tandem? onstrated an in

During the course of answering these questions, the existing literature will be explored in depth and the context for this study will be presented.

The chapter will broadly follow the structure presented in addressing the '*What*, *Why* and *How*' questions. First, an in depth look at *what* accommodation is will be presented. This is considered by recounting the key theories that underpinned the development of the study of accommodation. The next section considers *how* accommodation is currently measured, it presents the current methods used to measure and analyse accommodation and suggests a potential addition to these methods. Following this, accommodation is situated within a wider theoretical framework, demonstrating the reasoning behind *why* a link between accommodation and alignment of brain activity might be supposed. It will present the relevant literature to justify the exploration and investigation of this link. The final part of this chapter will present a consideration of *how* to measure the phenomenon in light of the considerations raised in the literature. The issues surrounding measurement of data for the type of study presented in this thesis will be detailed and a potential solution, utilising methods from computational science, will be presented.

## 2.1 What is accommodation?

Chapter 1 offered a cursory outline of what accommodation has come to be known as, here this will be greatly elaborated upon. Chapter 1 also presented a description of how the term *accommodation* would be defined and used in this thesis. Before proceeding with the main drive of this section, it is worth briefly expanding upon why *accommodation* is defined in the way that it is for this thesis. This consideration offers an insight into the number of fields and the breadth of research interested in this phenomenon and supplies a solid platform for further elaboration.

Whilst accommodation, in terms of it's observable characteristics in speech, can simply be considered to refer to the tendency of a speaker to adjust their production of speech sounds in relation to the person (or persons) they are speaking to (Giles, Coupland, & Coupland, 1991a), its social and psychological implications are much further reaching. Speakers can express accommodation by either adjusting their speech towards the speech of the other speaker, this is known as *convergence*, or by adjusting away from the speech of the other speaker, *divergence*. However, accommodation has also come to be known by a number of other names such as mimicry (eg. Kulesza, Dolinski, Wicher, & Huisman, 2015), interpersonal adaptation (eg. Burgoon, Stern, & Dillman, 1995), shadowing (eg. Pardo, Jordan, Mallari, Scanlon, & Lewandowski, 2013b), speech alignment (eg. Miller, Sanchez, & Rosenblum, 2013), entrainment (Beňuš, 2014) and imitation (eg. Delvaux & Soquet, 2007). Each of these terms has their own specific focus in the assessment of this phenomenon but still belong under the umbrella of accommodation. For instance, when the term *shadowing* is used in the work of Pardo et al. (2013b) they refer to speakers repeating words after having heard them presented over headphones. This can still be considered to be evaluation accommodation since it is assessing how a speaker adjusts their speech in relation to stimuli from another person. However, it is a very precise and targeted assessment of accommodation in a specific setting, with tightly controlled variables. This is an excellent strategy for the purposes of the study presented by Pardo et al. (2013b) and given the restrictions placed on the speaker, it is clear why this is reffered to as *shadowing* rather than accommodation. Whilst this work does place some tight restrictions on the environment in which accommodation is allowed to occur, it nevertheless provides invaluable information concerning its nature.

The work of Pardo et al. (2013b) is presented as an example of the types of work that clearly offer great insight into accommodative behaviour but which may not overtly be considered as accommodation per se. In addition to presenting an example of a specific aspect of accommodation, their work also presents a consideration of the ways in which accommodation has been evaluated, making an attempt to reconcile perceptual and acoustic-phonetic measures of accommodation. They make the point that:

> perceptual measures provide a global estimate of convergence, while acoustic measures contribute to an understanding of the attributes that talkers employ when converging. (pp.185)

This is an important consideration for the work in this thesis since the methods presented aim to help bridge the gap between these two approaches. This consideration also demonstrates the integration problem facing studies of accommodation. Since accommodation encompasses a number of acoustic-phonetic measures it is possible for them to vary in different ways. For example, one acoustic-phonetic measure may tend to converge in a given context whilst another might diverge. This presents issues for the interpretation of the findings since conflicting results are found. In the same situation, this may be considered as a meaningful level of accommodation in a perceptual measure because the combination of convergence and divergence for those particular acoustic-phonetic measures may hold a specific social meaning. However, from the point of view of acoustic-phonetic analyses, the effects may well be considered to cancel each other out in terms of overall accommodative behaviour. There are merits to both approaches and they provide information about different aspects of accommodation.

Finally, there is a tendency in the literature to focus on the directionality of accommodation. Again, this is evidenced by Pardo et al. (2013b) where they focus on the convergence of the speakers. Studies tend to look for movement of the speakers in one direction or the other, they will look for either convergence or divergence. Both of these movements are important in accommodation as they are linked to the demonstration of important social information (see subsection 2.1.2). Considerations focusing on directionality of accommodation are important and offer valuable insights into specific aspects of accommodation but if accommodation is to be assessed as a whole, it is important that they are both represented.

The reason that this thesis defines accommodation in its most broad sense of speaker adaptation in relation to their speech partner, in any direction and across multiple speech characteristics, is so as to encompass all of the above. If the work presented in this thesis is to consider the potential correlates between accommodation and brain activity, a holistic view of accommodation must be taken in order to better represent the whole of the process that might be taking place in the brain rather than trying to link brain activity to a small subset of what accommodation actually encompasses.

The remainder of this section presents how the theory of accommodation has been built and how accommodation relates to the social factors that underpin those theories. Particular focus is given to factors that have a particular influence on the work of this thesis. Subsection 2.1.1 details how the study of accommodation was born out of sociolinguistics but quickly drew on other fields to help understand what function it played in communication. It outlines how theories relied on social factors in order to explain the accommodative behaviour observed in the speech

signal. Subsection 2.1.2 focuses on how the four specific social factors of speaker sex, dialect, familiarity and dominance impact on the assessment of accommodation. Finally, subsection 2.1.3 pulls together the work presented and summarises the key points pertaining to this section. It then frames this section in relation to the themes presented in section 2.2.

### 2.1.1 A history of accommodation theory

This section aims to provide a history of the direction of travel that the theory surrounding accommodation has taken over the years. Attempts are made to keep the trajectory as linear as possible in the interests of providing a clear understanding of the work that lead to the consideration of accommodation as a phenomenon.

Labov (1966) demonstrated that changes in pronunciation patterns of his interviewees were dependent on social factors such as class and the linguistic prestige associated with a given pronunciation. Labov considered social context to be the main driving force behind the variations in people's pronunciation patterns. What he proposed was that usage of more prestige variants of pronunciation would be more likely to occur within formal contexts whereas the usage of more regional or colloquial pronunciations would occur more often when people are in less formal contexts. He went on to propose that the mechanism for this shift is based on the amount of attention paid to speech by the speaker because he observed that the formality of pronunciation tended to increase when speakers were asked to read aloud. This work was later considered by Howard Giles who would go on to become a major contributor to the work on accommodation.

Giles (1973) made the observation that some of the conditions that Labov interviewed his participants in may have had additional influences on the changes in the interviewee's accent. He suggested that it is not only social context that influences this change but also interpersonal aspects of the situation at hand. He was proposing that both social context and situational constraints have a part to play in accent variation. He referred to the interpretation of this accent variation in this way as *accent mobility*. Giles later goes on to flesh out this proposal with findings from his own data and puts forward a number of considerations on how the mechanisms of accent mobility might work.

Giles (1973) argues that the mechanisms underlying accent mobility are reliant on factors such as the linguistic level at which accent change is assessed by a native speaker (pp.101-102), the impact of cognitive and perceptual styles of interlocutors (pp.102) and the implications of visual and empathic feedback between interlocutors when assessing accent change (pp.103). Where Labov was attributing changes in accent to high level concepts such as prestige and attention, Giles aimed to focus in on these concepts in order to engage with their constituent parts. He suggested that accent mobility involved processing at the linguistic, cognitive and social levels.

Since accommodation necessitates an interaction with speech from another person, it is argued that language and social interaction are entwined in a two-way relationship. Social interaction facilitates language use and in turn, language use feeds back to influence social interaction. This work was seminal in establishing the basis for a theory of accommodation and demonstrates the beginnings of accommodation as a theoretical phenomenon. Here is the first suggestion that the use and adaptation of speech forms is dependent upon and responsive to social factors as interpreted from the speech signal.

Accommodation theory, in this form, was solely concerned with speech accommodation strategies and became known as Speech Accommodation Theory (SAT). The original ideas put forth by Giles (1973) were driven by a want to:

> […] redirect theoretical attention to more focused contextual dimensions […] and to argue the primacy of receiver characteristics over other considerations (Giles et al., 1991b, pp. 5-6)

This was essentially suggesting that his original proposals focused on how the situational constraints of an interaction influence accent mobility when considered in relation to the characteristics of the person being spoken to. As a result of this, research focused on the situational contexts in which SAT could operate (See: Powesland & Giles, 1975; Giles, Bourhis, & Taylor, 1977; Ball, Giles, & Hewstone, 1985; Coupland & Giles, 1988; Gallois, Franklyn-Stokes, Giles, & Coupland, 1988).

One of the key outcomes from the consideration of accent mobility under SAT has been to shed light on the cognitive and affective processes that occur during the use of accommodative speech strategies (Giles et al., 1991b). This could not have been accomplished without the supporting work from social psychology and psycholinguistics. The psychological literature at the time was able to offer insights that would allow researchers to build on the influence of receiver characteristics.

For example, Ball, Giles, Byrne, and Berechree (1984) built on existing frameworks by asserting that accommodation does not only offer a way to understand the reasons for modifications made to speech but also the social consequences of such modifications. Ball et al. (1984) take the view from social psychology that linguistic variations are used to either better integrate ourselves into a given community or to make ourselves distinct from a community by affirming our own community's identity (and by extension, our own identity). They argue that speakers will converge when there is a socially appropriate reason to do so, such as wishing to express affinity with a higher level of social prestige.

In SAT, it is argued that the expression of social liking can be achieved by adjusting utterances in relation to what is heard from the interlocutor. This is done to make oneself seem more linguistically attractive or desirable to said interlocutor through a shift towards their speech patterning. Importantly, the implications of accommodation are not taken further, they remain solely as a linguistic mechanism to demonstrate social information.

As detailed earlier in this section, SAT is only concerned with accommodative behaviours found in the speech signal. However, communicative actions between interlocutors often include non-linguistic aspects that may carry a large amount of communicative value that might also impact on levels of accommodation. Some of these non-linguistic factors can be seen to be addressed in the work of Simard, Taylor, and Giles (1976), who proposed a revision of the initial structure of accommodation theory in order to encompass some non-verbal forms of expression.

Simard et al. (1976) found that we are less likely to perceive a given interlocutor favourably if we believe them to be less willing to accommodate towards us. In the same vein though, when a non-accommodative act was found to be attributable to external pressures then negative reactions to the interlocutor were reduced. These effects were found to be true for the same shifts in pronunciation, which suggest that there is more involved in accommodation than purely linguistic factors. They proposed an elaboration on SAT based around these new findings and using *Attribution theory* (See: Kelley, 1973; Jones & Davis, 1965; Heider, 1958) to provide a set of guidelines for the expansion of SAT.

The new proposals suggested that our perception of the intentions and behaviours of an interlocutor can influence levels of accommodation. Studies such as that provided by Simard et al. (1976) helped to illuminate the possibility that interlocutors may not be accommodating to what was actually being said but rather, to a target percept of the interlocutor. However, this brought with it issues regarding what the contributing factors that made up the target actually were. If our interpretations of others were impacting on speech accommodation then there must be some influence of non-linguistic factors. This insight prompted researchers in the field to build upon SAT so that an understanding of accommodation's role in communication as a whole might become possible. In order to better tackle this issue, the wider notion of Communication Accommodation Theory (CAT) was proposed (Giles et al., 1987).

CAT built upon the work laid down by SAT and went on to explain aspects of communication that existed outside of the core linguistic speech factors. It also made additions to the accommodative strategies available to speakers engaged in a conversation. In addition to convergence and divergence, CAT offers *maintenance* and *complementarity*. *Maintenance* describes a situation where speakers demonstrate no shift in their pronunciation patterns, here it can be assumed that no change in social distance is being produced. *Complementarity* describes the event of one speaker becoming increasingly more similar to their speech partner whilst the other becomes more dissimilar from the first speaker. This might emerge in a situation where one speaker has reason to want to reduce social distance whilst the other wants to increase it. An example of such an interaction might be during making a shopping purchase. The cashier may be trained to make the customer's shopping experience as positive as possible and as a result may use accommodation to help reinforce this by converging towards the customer in an attempt to reduce social distance. The

customer on the other hand might be in a rush and needs to complete the transaction as quickly as possible. In this instance the customer may diverge from the cashier in an attempt to further the social distance, reinforce their own social background and implicitly signal the cashier that they do not wish to engage in pro-social small talk. Each of these methods of accommodation (Convergence, Divergence, Complementarity and Maintenance) are presented as a conceptual visualisation in figure 2.1.

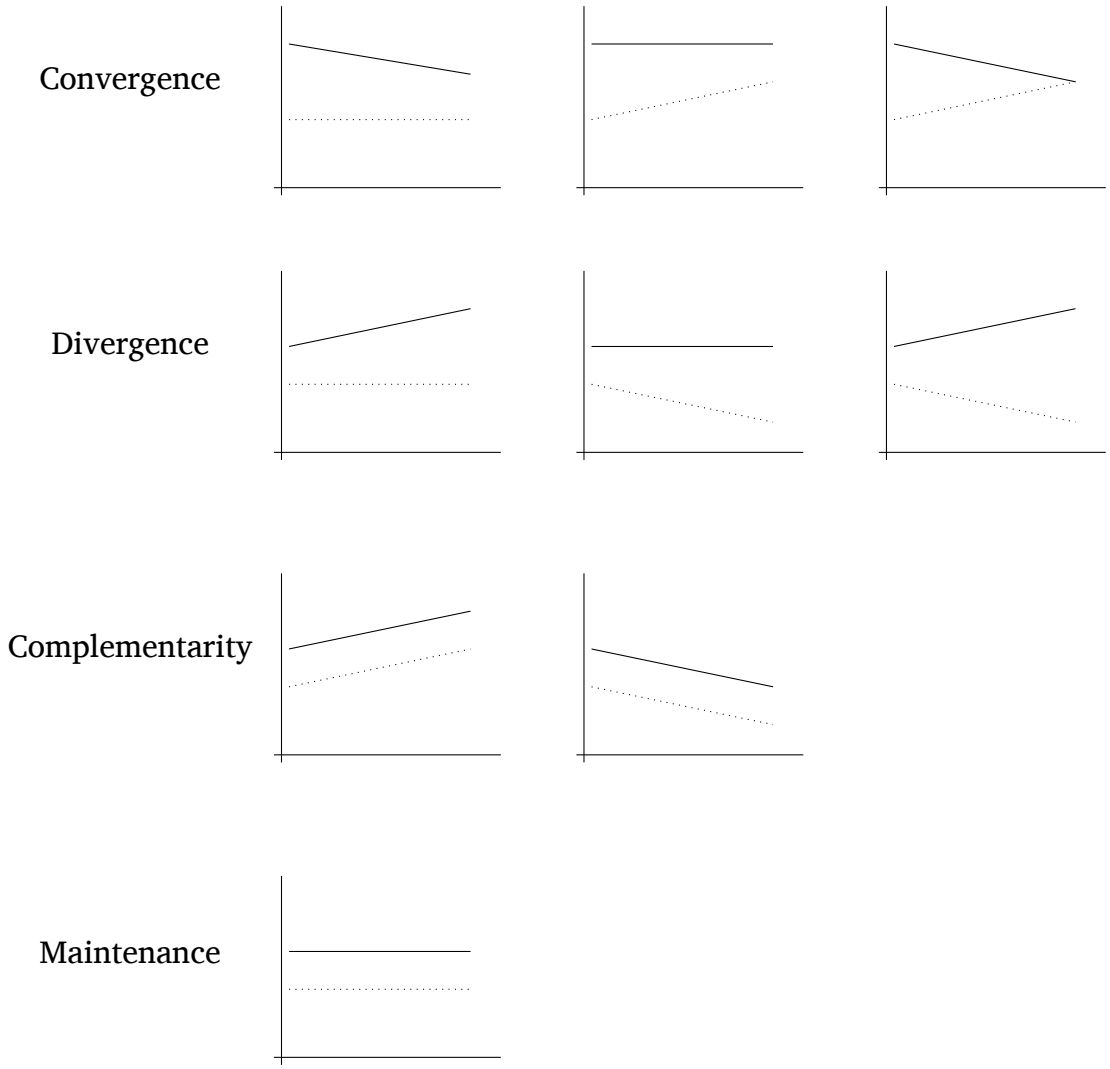



Figure 2.1: Conceptual representation of the different types of accommodation that may occur in a dyadic interaction. Each line represents the speech of one of the two speakers in an interacting dyad. The $x$-axis represents time and the $y$-axis is a conceptual dimension representing a measure of speech patterning.

Giles et al. (1991a) argue that under CAT, accommodation is utilised for a number of different communicative purposes. They suggest that because of this, it can be used to:

> Explore the complex interrelations of communication strategies and styles, the multiple social and psychological dimensions that contextualize them, and their social implications (pp. 49).

This move from considering accommodation as simply a linguistic phenomenon towards one that might offer greater insights into broader communicative strategies was an important one. It opened a new avenue of investigation for studies concerned with accommodation and provided a platform for the investigation of the relationship between accommodation and non-linguistic social factors.

### 2.1.2 Accommodation and social factors

As outlined in the previous subsection (subsection 2.1.1), CAT proposes a relationship between non-linguistic elements of communication and accommodation in the speech signal. It suggests that this is a two-way relationship, just as accommodation can convey social and psychological information, so too can accommodation be influenced by other, non-linguistic, social cues. Speakers use accommodation to demonstrate and react to indexical information about their interlocutor. This subsection of the thesis concerns the non-linguistic factors that have been shown to impact on accommodation.

#### Speaker sex

The field of sociolinguistics has long reported the differences in linguistic patterning between sexes (see: Milroy & Gordon, 2008, Section 4.4). Women have often been seen to be more likely to demonstrate more widely distributed instances of linguistic innovation and variation in relation to social and regional factors (Mees, 1987; Holmes, 1997; Chambers, 1995). However, the relationship between language use and sex in the sociolinguistic literature is not straightforward (Milroy & Gordon, 2008). This is even before gender is considered, which has been argued should be properly considered as a continuum rather than a binary variable (Eckert, 1998). It would therefore be plausible to assume that some of this innovation and variation, along with the complexity that comes with the factor of sex, will also be present in accommodation.

Namy et al. (2002) investigated whether men and women have different perceptual sensitivities to accommodation. They performed an experiment that involved three tasks: a speaker task, a shadower task and a listener task. The speaker task involved recording four speakers (two female, two male) reading out a word list. The shadower task involved sixteen participants (eight female, eight male) reading the same word list, this formed their baseline speech sample. These participants then repeated the word list after hearing each of the four speakers from the speaker task reading it, this is also known as shadowing. The listener task assessed accommodation by asking a third group of sixty-four participants (thirty-two female, thirty-two male) to listen to the original speaker's reading of a word and to assess whether the baseline or shadowed speech for the same word produced by the second group of participants sounded more like that of the original speaker.

The perceptual judgement provided by listeners was that convergence of shadowers could be found in all conditions with women being found to be more likely to accommodate than men. It was also found that women may also be more attuned to detecting accommodation in speech than men.

The authors suggest that their findings might be linked to the socialization of women to attend more to indexical features of conversational interactions such as intonation, tone of voice and speech rate. It is suggested that women may be more likely to incorporate these factors into their accommodative behaviours than men. This conclusion may have been partially driven by the work of Nygaard and Queen (2000), who found evidence that suggested that speaker identification through indexical features such as those described by Namy et al. (2002) may be a process more easily manifested in women.

Work by Bulatov (2009) found opposing trends in the level of accommodation for men and women in response to high-pass filtered and unfiltered speech. It was theorised that because the fundamental frequency of speech (F0) has been shown to facilitate the transmission of social information (see: Gregory Jr & Webster, 1996; Gregory, Green, Carrothers, Dagan, & Webster, 2001), convergence to high-pass filtered speech (removing the fundamental frequency) should be less likely. Bulatov (2009) used a similar paradigm to Namy et al. (2002), a single male served as the speaker and his speech was then presented to shadowers who either shadowed based on filtered or unfiltered speech. The shadowers consisted of nineteen participants, nine (six female, three male) in the unfiltered condition, and ten (six female, four male) in the filtered condition. There were twenty listeners asked to judge levels of convergence, ten (seven female, three male) were asked to judge the filtered condition and ten (six female, four male) were asked to judge the unfiltered condition.

Across the conditions, it was found that females were more likely to converge than men but were less likely to be able to detect convergence. To add more to this finding, Bulatov also demonstrated that females were more likely to detect convergence in the unfiltered condition than men. Both sexes performed at around chance for the high-pass filtered condition. These findings are interpreted as being linked to the emotional and social content that is conveyed in F0. It is posited that when F0 is filtered out of the speech signal, females lose their advantage to detect indexical features of speech to which they have been socially encouraged to be sensitive towards.

The role of sex was also investigated by Pardo (2006) who once again used the same general experimental format as Namy et al. (2002) and Bulatov (2009). Speakers were asked to read a list of phrases as a pre-task to obtain speech samples at baseline. They were then invited back to take part in a conversational task one to two weeks later. All pairs consisted of same sex dyads. This task elicited all of the phrases read in the pre-task. Following completion of the conversational task, the speakers were once again asked to read the same list of phrases read in the pre-task.

The data from the conversational tasks was used in the final part of the experiment. This used a group of listeners to determine if the phrases uttered in the post-task or the conversational task were more or less similar to that uttered by a task partner during the conversational task.

In a study of accommodation among both male and female pairs Pardo (2006) found that female pairs demonstrated less convergence than male pairs. She noted that this was inconsistent with other literature in the field, including the work by Namy et al. (2002). This finding is interpreted as being the result of different attentional sets for men and women. Where Namy et al. (2002) interpret their results as stemming from the perceptual sensitivity of the two sexes, Pardo (2006) suggests that both men and women may have different habitual attentional sets that ultimately influence their ability to detect perceptual differences. Pardo (2006) also makes the point that the findings might relate to differences between the use of accommodation strategies in a shadowing task and a conversational task.

In addition to studies focusing on perceived accommodation, work by Bailly, Lelong, et al. (2010) and Bailly and Martin (2014) have demonstrated the influence of sex on the spectral characteristics of accommodation. Both experiments used a form of speech elicitation known as *verbal dominoes* which was developed by the researchers. This method of speech elicitation was based on a game played in primary schools to help language learning. Participants are paired with a partner and each are presented with a word list of disyllabic words. The task is to choose the word from the word list that begins with the syllable that the previous word uttered by their partner ended with. The task in the papers presented by Bailly, Lelong, et al. (2010), Bailly and Martin (2014) used French speakers so the words used were in that language but an example of a similar chain of words in English might be something like this:

**Window** - Donate - **Eighty** - Teenager - etc…

Here **bold** typeface indicates one speaker and standard typeface indicates the other. The words uttered during the *verbal dominoes* interactions were compared to recordings of the same words from prior to this engagement. Comparisons were made based on the spectral energy characteristics (MFCC) for each word uttered by each speaker.

Bailly, Lelong, et al. (2010) found a number of strong cases of convergence and some modest cases of divergence. The strongest cases of convergence were found in same-sex pairs. Bailly and Martin (2014) found only convergence, with almost no cases of divergence. Once again, the strongest cases of convergence were found in same-sex pairs.

In sum, it would appear that in general, the literature suggests a greater tendency for females to converge than males and that greater convergence is to be found in same-sex pairs. However, evidence can also be found for the opposite of this

conclusion. This variation in the findings is likely to be tied to the social construct of gender rather than biological sex. The literature on how gender influences levels of accommodation demonstrates that there is an effect but that it is complicated (Pardo, Urmanche, Wilman, & Wiener, 2016b).

### Dialect

A key contributing factor to accommodative behaviour and to linguistic variation in general is variance in regional accent and dialect (eg. Giles, 1973; Foulkes & Docherty, 1999; Evans & Iverson, 2007). Research into dialect and accent variation is a burgeoning field and permeates a number of linguistic sub-disciplines. It has formed the basis of some of the earliest works in accommodation (eg. Giles, 1973; Ball et al., 1985; Coupland & Giles, 1988; Giles et al., 1991b). More recently however, there have been a number of studies specifically investigating the role of accent and regional dialect in accommodation.

Delvaux and Soquet (2007) performed an experiment where naturally occurring dialectal variations in Belgian French were used as stimuli to elicit accommodation. The participants assigned to the listening phase of experimentation were speakers of a different dialect of Belgian French to that used in the stimuli. They were given no instruction to accommodate or even to pay attention to the auditory stimuli which were presented to them through speakers as ambient background sounds. They were instead focused on a naming task. Convergence (or imitation, as the authors labelled the phenomenon) was found in the speech produced by the naming task although the stimuli were ambient and passive. The effect was held up to ten minutes after completion of the task.

Their results are interpreted within a number of theoretical frameworks but the important conclusions they draw are that the speakers are unaware of their general tendency to converge to the ambient speech that they heard. The authors suggest this finding implies that the process is automatic. They also add that as they demonstrated that the effect persists beyond the exposure, it must leave a memory trace. Since, there is an involvement of memory in the process, it cannot simply be a response behaviour but a more sophisticated process. Because it is a more sophisticated process the authors suggest a more appropriate term for this phenomenon might be *mimesis* after work by Donald (1991).

Although the main drive of their interpretation was focused on the cognitive mechanisms involved in convergence, their data does also suggest an impact of dialectal variation. Participants demonstrated a clear effect of movement away from their own regional dialect forms of vowels towards that of the dialect presented in the ambient stimuli. This suggests that the effect of dialect is influential even when speakers have no socially motivated requirement to accommodate (participants were only asked to complete a naming task). However, the focus of this study was not regarding the impact of dialectal variance on accommodation. Whilst infer-

ences about the role of dialect distance can be made, it cannot be claimed explicitly from this work. More direct evidence would be needed to reach the above conclusion.

In order to directly assess this effect of dialect variance on accommodation, a contrast would need to be drawn between the speech elicited from speaker pairs with similar dialectal backgrounds and for pairs who were from more differing dialectal backgrounds. Work by Kim et al. (2011) did exactly this by evaluating accommodation between pairs of native and non-native English speakers. The pairs were structured such that they represented three categories:

1. Same dialect - where speakers in a pair had the same dialect and the same first language (English or Korean).

2. Different dialect - where speakers in the same pair had different dialects but the same first language (English or Korean).

3. Different language - where speakers in the same pair had different first languages.

The participants were tasked with completion of a collaborative spot-the-difference task called the Diapix task (Van Engen et al., 2010). Their results showed a significantly higher likelihood of convergence for speakers with a similar dialectal background compared to those pairs with a more separated dialectal background or those pairs with a more separated language distance. In addition to this, they also found some evidence that speakers of different dialects of a language may diverge away from each other more so than in other conditions. They argue that the need for intelligibility is facilitated by closer dialect distance. It is suggested that the mental representations for a vast number of lexical items and phonological categories will have a higher likelihood of being shared between two speakers of the same dialect than two speakers of differing dialects.

This view is mirrored in work by Nilsson (2015) who analysed a corpus of sociolinguistic interviews between speakers from different dialect backgrounds in West-Sweden. Her investigations assessed accommodation in a number of dialectal variants between coastal and inland dialects. Findings demonstrated a clear use of convergence both towards more regional (coastal) forms and towards more standardised (inland) forms. In a traditional interpretation, the convergence towards the standardised form would be expected as it forms the more prestige variant. However, here we see convergence towards both forms. In addition, Nilsson (2015) found evidence of maintenance, where speakers did not converge at all.

The findings are interpreted within a CAT framework and suggest that these strategies are employed to facilitate the maintenance of common ground during disagreements (Clark, 1996) and establishing social affiliations. Results are complex and are influenced by additional factors but taking the stance that convergence

demonstrates shared common ground and increased social affiliation is in keeping with the suggestion of Kim et al. (2011) that similar dialect use may help to facilitate shared information.

Where Nilsson (2015) primarily argues for a socially motivated reasoning behind the accommodative strategies she observed, Babel (2010) provides evidence that suggests a simultaneous social and automated role for accommodation. Speakers of New Zealand English were recruited to take part in the experiment. Forty-two participants were recruited in total (thirty-four female and eight male) and initially read from two word lists to produce baseline utterances. This was followed by a shadowing task where they heard words read by an Australian English (AuE) speaker and were simply asked to say aloud the word that they heard. Finally, they once again read from the word lists. Prior to engaging in the shadowing task, participants were assigned to either a positive or negative group. Here they were given some information about the AuE speaker that described them as having either a positive or negative opinion of New Zealand. In addition, participants completed an implicit association task to determine their implicit bias towards Australia. Accommodation was assessed by comparing the vowels of the words elicited from the shadowing task and second word list task to elicitations from the first word list task.

Findings demonstrated that all participants tended to converge towards the AuE speaker (although not for all vowels), irrespective of whether assigned to the positive or negative group. Since convergence was observed in both groups, it can be reasoned that this behaviour is automatic rather than social since those in the negative group did not show evidence of divergence from the target speaker. However, it was also found that those participants rated as more pro-Australian would show a greater degree of convergence. This finding adds complexity to the previous result and is interpreted as the indicator of social distance. Babel (2010) suggests that the default behaviour for speakers may be convergence and that it is the degree of convergence that determines the expression of social distance. She argues for an interpretation of accommodation as both automatic and social, suggesting that 'speakers of language cannot help accommodating, but group-identity attitudes modulate this automatic process' (pp. 453).

Alshangiti and Evans (2011) add insight to this claim made by Babel (2010). They performed an experiment looking at accent accommodation in North-Eastern English speakers (NE) and Standard Southern British English speakers (SSBE) where the speakers were engaged in a conversational task. Both before and after engaging in the task, speakers completed a reading of the phonetically balanced passage Arthur The Rat (Abercrombie, 1964). All pairs consisted of one NE and one SSBE speaker, all participants were female, all NE speakers had been born in north-east England but had moved to London whereas all SSBE speakers were born and raised in London. The authors assessed accommodation both during the conversational task and whether it persisted beyond the interaction. To assess accommodation during

the interaction, snippets of speech were taken from early in the conversational task and late in the conversational task and phonetically trained listeners were asked to rate whether the snippets sounded more southern or more north-eastern. These snippets either consisted of an accent revealing (AR) or accent neutral (AN) phrase. The defining factor for whether a snippet was AR or AN was if it contained one of a number of key phonetic variables (vowels) that would identify an accent as either NE or SSBE. To assess persistence of accommodation, they performed the same comparison, using phonetically trained listeners, with sections from the pre-task and post-task readings of Arthur The Rat.

The authors observed that accommodation only took place in the NE speakers, who would accommodate towards the more prestige accent of the SSBE speakers. However, this was only found in the speech taken from during the interaction this effect did not persist into the post-task. The authors note that although the direction of accommodation was in keeping with the literature on accent mobility (Giles, 1973), the fact that it did not persist into the post-task is contrary to findings that suggest this effect even in similar accent pairs (Pardo, 2006). Further to this, the findings suggested a complex relationship between AN and AR speech segments. They found that when speakers were judged to have converged in AN snippets, they were found to have diverged in AR snippets. The authors interpret this as the NE speakers developing a 'hybrid accent, in which they used SSBE-like variants to show belonging to their new community, but retained some NE variants to show allegiance to their home community' (pp. 227). Ultimately, the authors conclude that accent accommodation is driven by short-term interaction effects that only impact long-term accent change given multiple interactions.

This work resonates with that of Babel (2010) in that accommodation is used in a targeted and context specific manner by the speakers. Whilst the NE speakers do converge to the SSBE speakers, it is nuanced and driven by the social factors of the engagement. Not all NE speakers demonstrated convergence and whilst the authors attribute this to a lack of accommodative space for the NE speakers since they had already been living in London for some time and they were friends with their conversational partners, it may also be due to the interaction between the automaticity of the system and the social factors bearing down on the speaker.

Taken together, the work reviewed on the role of dialect in accommodation presents a mixed picture. On the one hand, there is a good deal of evidence suggesting that dialects do play a role in accommodation but on the other, it is not clear how it interacts with other social factors and the cognitive systems governing communication.

#### Familiarity

Social factors by their nature require some form of interaction with another interactant. Of course, engaging with others will inevitably lead to potential for relation-

ships being built and as a consequence of this the number of engagements with that person is likely to increase, leading to more speech engagements with that person. The extent to which an increased engagement with another person has an impact on the overall ability to accommodate is a point of contention in the literature and one that will be elaborated upon here.

Research detailing increased amounts of convergence between pairs of acquainted speakers can be found in the work of Bailly, Lelong, et al. (2010) and Bailly and Martin (2014). An outline of their methodologies can be found in subsection 2.1.2. Bailly, Lelong, et al. (2010) looked at two different aspects of the speech signal, vocalic targets (vowels) and prosody. In addition to this they also performed an analysis using a form of automatic speech recognition which aimed to characterise the speech signal more holistically. The work looked at the difference between rates of both convergence and divergence in pairs of acquainted ('friends') and unacquainted ('unknowns') speakers.

Findings saw strong occurrences of convergence with some modest cases of divergence in both the vocalic target and speech recognition measures. Prosody remained largely unaffected and demonstrated only small amounts of significant accommodative behaviour. The authors note that the lack of accommodative behaviour may have been a consequence of the experimental task used. For the speech recognition results, the greatest amount of convergent behaviour was seen in the speech of the acquainted speaker pairs. The authors do not interpret these findings with a theoretical framework such as CAT since the paper focuses on practical applications of the findings. However, this demonstrates that there may be a relationship between familiarity of the speakers and convergent accommodative behaviour. By extension, according to CAT, this would mean that the acquainted pairs are actively attempting to reduce their social distance.

Further experimentation by Bailly and Martin (2014) provides additional support to their findings. Here they utilised the same experimental approach but with unacquainted pairs and familial pairs. The measurement of accommodation was focused solely on global speaker characteristics, using computational approaches to determine the direction and level of accommodation.

Using the computational approaches they employed, they were able to find more consistent results in comparison to more detailed phonetic analyses. They also demonstrated the strongest amounts of convergence between pairs that had well-established social relationships. These results further support the view that closer familiarity is linked to a greater degree of convergent accommodative behaviour. However, as cited in this study, Pardo et al. (2012) demonstrated an inconsistent effect of familiarity on accommodation.

In their work on accommodation in college (university) roommates, Pardo et al. (2012) looked at convergence rates in co-habiting pairs at four intervals over the course of three and a half months. College roommates were selected since they

are unacquainted when they begin at their college and will have daily interaction over an extended stretch of time. It was theorised that given the stance of CAT, the college roommates ought to demonstrate convergence as they have a vested social interest to reduce their social distance. To assess levels of convergence, they performed a perceptual test and a test for vowel convergence. They also examined whether the degree that the roommates felt close to one another was related to the level of convergence.

To perform the perceptual test, the authors took a baseline measure of each of the roommate's speech for four key phrases from the beginning of the academic year ($T1$). They then did the same for the remaining three time points in the year ($T2$, $T3$ and $T4$) to collect the comparison phrases. The phrases from one roommate at $T2$, $T3$ or $T4$ were then presented to listeners. These phrases were flanked by the $T1$ utterance of the other roommate and another utterance from either $T2$, $T3$ or $T4$. Listeners were required to report which of the two flanking utterances sounded more like the central one. For example, one set up might be:

$$T2_A - T2_B - T1_A$$

Where $A$ and $B$ stand for the phrases of each speaker and $TX$ stands for the time points. In this example, if the two roommates have converged by $T2$ then the listeners should report $T2_A$ to sound more similar to $T2_B$ than $T1_A$.

For the vowel measures, each roommate was asked to produce the full set of American English vowels, embedded in a carrier sentence. These were elicited at each time point and the first two formants from each time point were used for analysis. The measure for closeness was based on a questionnaire taken by the participants at $T3$.

Of these measures, consistent and robust convergence was found in the perceptual results but noted that the amount of time spent together had no effect on the level of convergence. This finding is contrary to that reported in Bailly, Lelong, et al. (2010) and Bailly and Martin (2014). However, they did report a significant interaction between the speaker pairs and time point. In addition, they also saw that the levels of convergence varied by the type of phrase assessed, with no clear pattern across phrases or pairs. This suggests that the nature and type of convergence observed is dependent on the individual pairs themselves, with overall convergence being the general trend but each pair demonstrating different methods to reach convergence. In the tests they performed on the vowel measures, the authors found complex patterns of change with each roommate pair differing in the patterning of their vowel distance from each other over time. Again an interaction between pair and time interval was found although there was no relationship between vowel measures and the perceptual results. The relationship between vowel measures and the relationship quality of the roommates revealed a correlation between rated closeness and convergence although this relationship was not found for the perceptual results.

The authors broadly interpret these findings as compatible with the proposals of CAT but add that it might be the case that individual speakers may 'converge on a unique set of acoustic-phonetic attributes while diverging, varying randomly, or remaining neutral on others.' (pp.196). They also stress that findings based on one acoustic-phonetic feature must be interpreted against the full phonetic repertoire of the speaker because patterning found for one acoustic-phonetic feature may not be used consistently by every speaker. Overall, the work demonstrates a more complex picture than that presented by Bailly, Lelong, et al. (2010) and Bailly and Martin (2014) and highlights an important issue regarding the nature of the relationship between perceptual measures and more holistic or perceptual measures (expanded upon in section 2.2).

Once again, results from exisiting studies paint a complex picture of the interaction between this factor and accommodation. Whilst it seems that greater convergence generally appears to be present in pairs of acquainted individuals, the evidence is certainly not conclusive. It may be the case that the type and length of relationship that a pair of speakers have plays a role in the accommodative strategies that they use. There may be differences in the accommodation expressed by friends and by family members for instance. It may even be the case that the type of familial relationship could impact on accommodation, where some more closely knit families demonstrate different patterns to those who are less close. On the other hand, unaquainted participants enter the experimental setting with the same amount of knowledge about each other and will therefore have less shared history impacting on their accommodative behaviour. Familiarity is a difficult factor to control for in an experimental setting and whilst it is most certainly worth investigating further, the use of unacquainted participants makes results between studies more comparable.

**Dominance**

Dominance describes the degree to which one speaker in a pair will tend to lead (or dominate) an interaction. It is something that can either be inherent in the social make-up of the pair or related to the particular task that the pair are engaged in (or a combination of the two). If one speaker is too dominant in an experiment, it can lead to an over representation of that speaker's speech. This is a problem when one wishes to investigate a phenomenon such as accommodation where a good sample of speech from both speakers is required. The natural dominance of a speaker within an interaction is difficult to ascertain before they actually engage in the task however, the influence of the task on the dominance is something that experimenters can control. This subsection is concerned with the role that dominance plays in accommodation and how it has been handled in the literature.

In studies that use pre-recorded read speech to assess accommodation (eg. Namy et al., 2002), dominance is not an issue because the people doing the accommodat-

ing do not engage in a conversation with the person that they are accommodating to. Dominance only presents itself as an issue in studies using spontaneous speech. Baker, Gallois, Driedger, and Santesso (2011, pp.762-763) offer a good overview of the methods that tend to be used to elicit spontaneous speech, noting that the most widely used approach is the Map Task (Anderson et al., 1991). This involves two participants, both of whom are presented with a map. Both of the maps have a series of landmarks on them, indicated by both a picture and some text (eg. a picture of three boats with the text 'moored boats' underneath). The two maps have a number of key differences between these landmarks. One of the maps has a path through these landmarks marked on it, the person that has this map is designated the instruction 'giver'. The other map does not have a path marked on it, the person with this map is designated the instruction 'receiver'. The participants cannot see each others maps. The task is for the receiver to draw a line on their map according to the instructions provided by the giver so that the two paths on the maps match up as best as possible. The level of success is determined by the amount of deviation between the two paths. This task has been widely used (eg. Kemper, Othick, Warren, Gubarchuk, & Gerhing, 1996; Koiso, Horiuchi, Tutiya, Ichikawa, & Den, 1998; Pardo et al., 2010; Heldner & Edlund, 2010; Aguilar et al., 2015) but the problem of speaker dominance is present in the data. Anderson et al. (1991) themselves note that in the Edinburgh Map Task corpus around $55,000$ word tokens were produced by the task receivers compared to $80,000$ tokens by the task givers.

This effect was investigated by Pardo et al. (2013a) where participants were asked to take part in a Map Task but switched roles between different task trials (ie. both participants played the role of giver and receiver an equal amount of times). The authors hypothesised that the role-switching would help to mitigate the effect of speaker dominance since each speaker should speak more when in the giver role. However, what was found was that the speaker assigned to the role of giver in the first instance maintained dominance throughout the experiment. Even when the roles were switched and the original giver was now in the role of the receiver, the original giver still continued to speak more than the current giver. In addition to this, it emerged that the accommodative tendencies linked to the original role attributed to the conversational partners persisted across trials. This study explicitly investigated the influence of conversational role on speaker dominance and its effect was found to be robust and pervasive. The findings of Pardo et al. (2013a) are supported by earlier findings by Pardo (2006) where the role of speakers demonstrated significant effects on the amount of convergence. However, in this earlier study the authors noted an interaction between speaker role and sex, suggesting that the role attributed to the participants and the effect that it has on convergence is not straightforward.

### 2.1.3 Summary

This section has explored what is meant by the term accommodation when considered in relation to the speech signal. It has offered a history of the theory underpinning the study of accommodation and has presented a consideration of the non-linguistic factors that impact on accommodation. The key points from this section are briefly summarised here.

Subsection 2.1.1 offered a history of accommodation theory and presented SAT and CAT. SAT was the first theory of accommodation and was originally focused solely on how situational constraints influenced accent mobility. It grew and was developed to include a mechanism by which social liking could be expressed by incorporating work from social psychology. CAT went beyond the remit of SAT by aiming to explore the communicative aims that underlie accommodative behaviour. It proposed that when accommodation takes place, it is in relation to a number of factors that can be discerned from the speech signal.

Following on from this, subsection 2.1.2 explains the links between non-linguistic factors and accommodative behaviour. It offers considerations on the impact that sex, dialect, familiarity and conversational role have on accommodation. Each of these factors will be important for the experiments presented in this thesis.

Overall, the existing literature has demonstrated that social factors do indeed have an impact on accommodation. However, the findings are complex and do not seem to follow a simple pattern. There are effects that can be seen to occur at more holistic levels of measurement that may then not be measurable or involve a complex interplay between more fine-grained acoustic-phonetic measurements. This is related to the wide array of methods and techniques used in the literature which makes direct comparison between studies somewhat difficult. For instance, comparisons between the findings of a perceptual study using pre-recorded samples of speech and findings from an acoustic-phonetic study that used live interactional speech can both be said to be studying accommodation but are looking at it from different perspectives. The perspective of accommodation that researchers evaluate is often linked to their theoretical stance on accommodation. Some may consider accommodation to be primarily automatic (cf. Pickering & Garrod, 2004) and may therefore take advantage of the control offered by laboratory settings and pre-recorded speech. Others may not share that view and might argue that accommodation is a more involved process (eg. Pardo et al., 2010), opting for the greater ecological validity offered by live interactional speech.

Clearly the way in which accommodation is measured has an important role to play in its evaluation. Understanding the variety of methods and techniques used in research on accommodation is essential. The following section aims to develop this understanding by evaluating the methods used to assess accommodation and highlighting what aspects of accommodation each type of approach is best suited for.

## 2.2 How is accommodation measured?

Where section 2.1 explored the theoretical links between accommodation and social factors, this section is concerned more explicitly with the process of measuring accommodation. It provides a framework within which studies concerning accommodation can be situated (see Table 2.1 for on overview) and offers interpretations of the merits and drawbacks for each approach.

To form this framework, the field has been divided by four key criteria that intersect with each other. Two of these criteria relate to the nature of the method used to assess accommodation. There are broadly two primary ways to do this, one involves asking listeners to make a judgement regarding whether accommodation has taken place, this is labelled as *perceptual*. The other main method involves looking at acoustic-phonetic features of the speech signal to draw a measure of accommodation, this is labelled as *acoustic-phonetic*. The other two criteria relate to the stimuli used to elicit accommodation and concern whether this stimuli was drawn from a pre-recorded source or a live engagement, they are labelled as *non-interaction* and *interaction* respectively. Bringing these four criteria together allows for the formation of a two-by-two matrix that categorises most of the field and provides four categories of investigation: *Perceptual Interaction*, *Perceptual Non-Interaction*, *Acoustic-Phonetic Non-Interaction* and *Acoustic-Phonetic Interaction*. These distinctions are conceptual ones that allow for current studies to be situated in relation to one another and to provide a landscape in which to situate this thesis. However, before considering the work in each of these four categories, it is worth outlining exactly what is meant by each of the four constituent criteria (ie. perceptual, acoustic-phonetic, non-interaction and interaction) and where each of their strengths and weaknesses lie.

### Perceptual

Studies using this approach take advantage of the human perceptual system's ability to detect variations in the speech production of others. The key reasoning behind running experiments of this nature is that it may not be possible to discern accommodative mechanisms from a single acoustic-phonetic measure or even a number of acoustic-phonetic measures (Babel & Bulatov, 2012).

The use of perceptual measures generally assumes that the full phonetic repertoire of an individual needs to be accounted for. This is something that is effortless for human evaluators but proves difficult when using acoustic-phonetic measures because an evaluation of a speaker's full phonetic repertoire would be infeasible to produce and then to evaluate statistically due to the large number of non-independent, co-varying measures. In addition there may be interactions between the acoustic-phonetic measures which relate to social factors that the speaker is expressing. Dis-

crete acoustic-phonetic measures would not be able to account for these complex interactions, human evaluators on the other hand have an ability to do this.

A much used and common assessment paradigm which used perceptual measures is the *AXB* paradigm developed by Goldinger (1998). This type of test has already been briefly outlined in the reporting of works in subsection 2.1.2 but was not referred to explicitly as an *AXB* paradigm. The general format of an *AXB* experiment involves making a series of pre-recordings of target stimuli from a speaker or speakers, this forms the *X* of the *AXB*. Another set of recordings is then made from another group of participants using the same stimuli that are used for the pre-test, forming the *A* of the *AXB*. They are then exposed to the target recordings (the *X*) and are asked to repeat the content of the target recording as quickly and accurately as possible after hearing it. This forms the *B* of the *AXB*. These pre-recorded target words are then presented to yet another group of participants in the order $A-X-B$ and they are asked to indicate which utterance is more similar to the target stimuli, the *X*. The *A* and *B* recordings have the same target word or phrase as the *X*. The *A* and *B* stimuli are counterbalanced so that the pre and post exposure recordings are presented a balanced number of times on either side of the *X*. Different elements of the *AXB* can be adapted to suit the needs of particular research questions. For instance, Pardo et al. (2010) used early and late excerpts from a conversational interaction to form the *A* and *B* elements in the test and Alshangiti and Evans (2011) used phrases rather than single words. Use of the *AXB* paradigm allows for the interpretation of how much speakers producing the *A* and *B* elements of the test adjust their speech in response to hearing the speaker producing the *X* element. This is done by allowing listeners to make a perceptual judgement of the similarity of pre and post-exposure recordings to the target recording.

Use of this form of perceptual assessment of accommodation has been influential in studying this phenomenon. Understanding the way this method of testing works also helps to demonstrate what is meant in this thesis by a perceptual measure. It is any measure that uses the human perceptual system to assess accommodation. Whilst it is possible for perceptual methods to be constructed such that they target specific acoustic features by artificially manipulating stimuli, this has not been the case in the literature that is reviewed. In this thesis, perceptual measures are those that do not use acoustic measurements of the speech signal and only make global estimates of accommodation rather than assessing the attributes that speakers use to accommodate. It allows for researchers to avoid making assumptions about the phonetic elements that underpin accommodation and instead focus on evaluating it as a whole. However, this benefit comes at the cost of losing the ability to make an assessment of the speech attributes that contribute to accommodation. It is an excellent tool for determining if accommodation has taken place but it does not provide answers as to how it took place in terms of the phonetic features employed.

### Acoustic-Phonetic

Using acoustic-phonetic measures to evaluate accommodation allows researchers to make assessments of the specific speech attributes that speakers are using to accommodate. They are defined in this thesis as any measure that makes an assessment of accommodation through a measurement of the speech signal. These methods are more accurate at pinpointing acoustic speech features liked to accommodation but in order to do so they must chunk the speech signal in order to evaluate it. For instance, a study might look at whether speakers accommodate their speech rates. However, in order to do this, other elements of the speech signal are excluded from analysis. This means that all accommodative information contained in any other attributes of the speech signal is, to some extent, lost. In contrast to the perceptual measures, acoustic-phonetic measures are unable to account for accommodation as a whole.

Of course, one could aim to evaluate a number of different acoustic-phonetic attributes of the speech signal with the aim of accounting for as much of the accommodative effect as possible. In practice though, doing this would prove very difficult for a number of reasons. Firstly, the process of extracting all the necessary phonetic attributes would be highly time consuming and labour intensive. There are a vast array of phonetic measures that can be extracted from the speech signal and even with the use of automated (eg. Rosenfelder, Fruehwald, Evanini, & Jiahong, 2011) and semi-automated (eg. Sonderegger & Keshet, 2012) measurement techniques it would still be an incredibly sizeable task. Secondly, as stated in the phonetic literature (Öhman, 1966; West, 2000; Lien, Gattuccio, & Stepp, 2014), the surrounding phonetic context from which a measure is taken has an influence on the realisation of certain phonetic measures (eg. West, 1999a, 1999b). Accounting for all possible effects of surrounding phonetic context in order to isolate the true degree of accommodation in the phonetic measure of interest would again be quite a challenge. Finally, accommodation occurs across a range of phonetic attributes of the speech signal. Perceptual measures can account for the interplay between all of these since this is an feature of the human perceptual system. Accounting for all of these interactions when looking at multiple isolated phonetic measures would add further complexity to an already difficult problem. So, whilst making an assessment of accommodation in a speaker's full phonetic repertoire using isolated phonetic measures may not be impossible, it does present a potential hurdle. Acoustic-phonetic measures of accommodation are best suited to the evaluation of accommodation in studies targeting specific speech attributes.

### Non-Interaction

Where *perception* and *acoustic-phonetic* refer to methods of measurement for accommodation, *non-interaction* is in reference to the type of engagement used to elicit ac-

commodation. Specifically, non-interaction refers to research that has used a form of pre-recorded stimuli in order to elicit accommodation from a speaker. Experiments using these types of stimuli will generally refer to this particular type of accommodation as *imitation* (Babel, 2009a). The reason for this is that when speakers interact with a pre-recorded stimulus, the social imperative to accommodate has been removed. A benefit of this is that studies using non-interactive stimuli are able to make evaluations of the automaticity of accommodation. With the joint social imperatives that are present in a live interaction removed, it can be argued that any remaining accommodative effects are the result of an automatic accommodative mechanism. The speakers doing the accommodating do not have access to any additional social information about the speaker that they are accommodating to and must therefore make any accommodative choices based on the available information in the speech signal. In addition, the use of immediate repetition upon hearing the target utterance further helps to focus findings on the automaticity of accommodation given the short period in which to make an assessment of an appropriate level of accommodation and provide a response. One further benefit of this elicitation method is that researchers have the ability to control the stimuli. Selections can be made with regards to the exact stimuli presented to the speaker doing the accommodating. By doing this, the experimenter can make an assessment of specific speech attributes of interest.

The main drawback with using non-interactional stimuli is that the adaptive nature of accommodation is lost. Speakers are unable to respond to each other in a dynamic and socially meaningful way. A conversational interaction necessarily involves at least two participants and they will each demonstrate their own patterns of accommodation in response to their partner. The use of non-interactional stimuli loses this inter-dependent element of accommodation, it only allows for accommodation to be assessed from the perspective of a single person. Further to this, it does not allow for an assessment of the dynamics of a natural interaction. Over the course of an interaction speakers express a general accommodative trend but, as with all generalisations, this will consist of a number of differing movements. In live speech, the course of a conversation will involve a number of accommodative movements towards and away from an interlocutor. This is an important feature of accommodation because the desired social distance between speakers will often change dependent on the content of the interaction. For example, two speakers may begin a conversation in agreement but then come to disagree at a later point. At the beginning of this conversation, one might expect a small social distance as evidenced by a high degree of convergence, this may then change as the speakers begin to disagree. Using non-interactional stimuli would not be able to capture this change since the data would have been drawn from read speech. This highlights another issue with non-interactional stimuli which is that they are often produced from having a speaker read a word-list. There have been a number of studies document-

ing the differences between read and spontaneous speech (Howell & Kadi-Hanifi, 1991; Hirose & Kawanami, 2002; Walker, 1988; Laan, 1997) and asking speakers to accommodate to read speech may not provide results that can be generalised to a broader notion of accommodation.

### Interaction

Those studies classed as using *interaction* as stimuli to elicit accommodation are defined in this thesis as using speech stimuli drawn from a live speech engagement between at least two speakers. Methodologies employing this form of elicitation allow for accommodation to be evaluated in a form that more closely resembles that found in real-world conversational settings. Where those studies using non-interaction to elicit accommodation have issues regarding the loss of the dynamic nature of accommodation, studies using interaction do not. Here, speakers can adapt to the utterances of their interlocutor in real-time and may do so in a socially meaningful way. This allows for accommodation to be considered in a more complete form where both the automaticity and social impacts on its realisation are allowed to manifest. Further to this, it allows for accommodation to be evaluated from the point of view of each of the speakers engaged in the interaction. This offers a more complete picture of how speakers accommodate in relation to one another to be built. Finally, the use of stimuli from an interactional source eliminates the issues surrounding the use of read speech.

Whilst the use of interactional stimuli provides researchers with a more ecologically valid way of evaluating accommodation it isn't without its own pitfalls. One key issue with the use of interactional stimuli is that by allowing speakers to interact with one another a certain amount of experimental control is lost. Whereas those studies using non-interactional stimuli had the ability to control the stimuli to which speakers were accommodating, studies using interaction to elicit accommodation do not. Control over the exact context in which a speech variable is produced is lost and so holding the stimulus to which a speaker accommodates to static is not an option. This is a feature that is inherent to interaction since conversation unfolds dynamically and accommodation is produced 'on-the-fly'.

The remainder of this section considers the measurement and elicitation techniques discussed above in relation to each other. This allows for research in the field to be considered under the framework as summarised in Table 2.1. Considerations begin with the top-left hand quadrant (Perceptual Interaction approaches) of the table and work around in a clock-wise manner, ending with a consideration of the types of methods represented by the bottom-left hand quadrant (Acoustic-Phonetic Interaction approaches).

Subsection 2.2.1 details the approach taken by studies using perceptual measures of accommodation based on interactional data. It provides information regarding

| | | Stimuli | |
|---|---|---|---|
| | | Interactional | Non-Interactional |
| Measure | Perceptual | Perceptual interaction approaches | Perceptual non-interaction approaches |
| | Acoustic-Phonetic | Acoustic-phonetic interaction approaches | Acoustic-phonetic non-interaction approaches |

Table 2.1: Conceptual map outlining the broad approaches to assessing accommodation that are currently used.

the strengths and drawbacks of these methods and reports the findings of such studies.

Subsection 2.2.2 continues to evaluate perceptual measures of accommodation but shifts its focus to those studies that use non-interactional data as the stimulus to accommodate to. It again considers the strengths and drawbacks of these methods in relation to their findings.

Subsection 2.2.3 changes the focus of the measure to acoustic-phonetic measures but holds the accommodative stimulus as a non-interactive data source. Findings are reported and the strengths and drawbacks of these methods are discussed.

The penultimate part of this section, subsection 2.2.4 looks at the final iteration of the four categories, acoustic-phonetic measures using interactive data as the stimulus to accommodate to. Once again, strengths and drawbacks are discussed and findings reported.

Finally, subsection 2.2.5 draws together the various streams of methodologies presented in this section and recaps the key points discussed. It also situates the material presented here in relation to the following section, section 2.3, which discusses the link between accommodation and joint brain activity.

### 2.2.1 Perceptual interaction approaches

Perceptual interaction approaches to measuring accommodation are presented here as studies that have used perceptual evaluations of data drawn from an interactional setting to make an assessment of accommodation. These measures have the combined benefits and drawbacks associated with the use of perceptual evaluation and interactional stimuli, as described in the beginning of section 2.2. The main aim of this subsection is to highlight what elements of accommodation this method of elicitation and measurement can reasonably expect to be assessing.

Two studies that make an assessment of accommodation using a perceptual interaction approach are Alshangiti and Evans (2011) and Kim et al. (2011). Both studies

make use of the conversational speech elicitation task known as Diapix (Van Engen et al., 2010). It should be noted that whilst Kim et al. (2011) used data sourced from the original Diapix format produced by Van Engen et al. (2010), Alshangiti and Evans (2011) used a modified version designed specifically with a British audience in mind called DiapixUK (Baker & Hazan, 2011). The core aim of the Diapix task remains the same in both forms. The task itself is a collaborative 'spot-the-difference' task where speakers must communicate with one another in order to find all of the differences between two pictorial images of the same scene but with minor differences between them. The differences that the speakers were required to find are able to serve as target words that elicit key phonetic variables. Participants are each given a different version of the scene and cannot see the other person's scene. The use of this type of task allows for speakers to take part in an unscripted conversational interaction whilst retaining some experimental control. Additionally, as noted in Kim et al. (2011), this task does not lead to speaker dominance and allows for a more balanced collection of speech samples than other method of eliciting spontaneous speech (see subsection 2.1.2). Ultimately, as with all methods using interaction based elicitation methods, the aim is to allow accommodation to emerge as a result of the interaction rather than inducing it through controlled exposure.

The goal of Alshangiti and Evans (2011) was to investigate phonetic alignment in spontaneous speech through the use of accent ratings. In brief, the task consisted of pairs of North-Eastern British English (NE) and Standard Southern British English (SSBE) speakers who performed individual pre-interaction and post-interaction reading tasks with the main task being the DiapixUK task. The structure of this study was described in detail in subsection 2.1.2. Analysis of the recorded data was performed by phonetically trained listeners who were asked to assess the level of convergence during the DiapixUK task based on extracts of approximately two to three seconds in length from early and late on in the conversation. These extracts were split into accent neutral and accent revealing categories such that the accent revealing category contained one of the key phonetic variable words. The listeners were also asked to assess the level of convergence between the pre and post tasks. The use of DiapixUK to elicit accommodation and then human listeners to evaluate it are the criteria that define this study as perceptual interaction.

In the analysis of the conversational speech, evidence for convergence of the NE speakers towards the SSBE speakers was found. However, it was small, present in only some NE speakers and was modulated by the presence of accent revealing stimuli. Looking at accommodation as elicited from the conversational speech allows this study to investigate the changes made by speakers to the realisation of specific target elements of their accent (as dictated by the task key words) as a result of accommodation in response to a real interpersonal engagement. As a result, a greater number of factors impacting on accommodation are allowed to vary and

as discussed in section 2.1 social factors play a key role in accommodation. The authors make note of these as possible explanations for the detection of only small effects. They note that the pairs used were 'friend' pairs who had known each other previously and may have therefore already accommodated towards one another outside of the testing paradigm. Additionally, most of the NE speakers had already been living in London for a minimum of four months prior to experimentation and may have already shifted their speech somewhat towards the southern form. In addition to these explanations, the total time of exposure was at most only ten minutes. This may have impacted on levels of accommodation since it doesn't allow a great deal of time for accommodation to emerge. Further to this, the study treats time as a binary variable, looking at only early and late excerpts of the conversational data. Since speakers respond to one another in real-time, it is likely that accommodation towards or away from an interlocutor varies from moment to moment based on the desired social distance at that particular time. These two remarks about the short interaction time and the treatment of time as a binary variable can be taken together as a statement regarding the time-scale of accommodation. Although there may be a general trend of accommodation in one direction, this is not to say that all accommodative movements in that engagement run in the same direction. Over the course of the engagement there are likely to be movements towards and away from each other dependent on the content of what is being said and how this impacts on the desired social distance. Treating time as binary removes the possibility of looking at these smaller movements and allowing only ten minutes for an interaction may not provide enough time for effects in a particular direction to build up in all speakers.

Further to their findings in the conversational data, the authors did not find evidence of the persistence of convergence into the post-task. However, this comparison was made between the read speech of the speakers. The degree to which this form of speech is comparable to the accommodative behaviour seen in the conversational task is difficult to assess given the evidence for changes in speech during read speech (de Ruiter, 2015; Howell & Kadi-Hanifi, 1991; Mehta & Cutler, 1988). It may have been the case that the authors were looking to make a comment on the transferability of accommodative effects between conversational and read verbalisations but this is hard to determine from the article itself as it is not explicitly stated.

In sum, whilst this study may not provide conclusive evidence for accent accommodation between NE and SSBE speakers during a conversational interaction, it does provide insights into a number of considerations that need to be made in order to develop a measure of accommodation for spontaneous, conversational speech.

Work by Kim et al. (2011) looked at the impact of interlocutor language distance on accommodative behaviour. The conversational pairs here consisted of American English – American English pairs, Korean-Korean Pairs and American English-Korean, American English-Chinese pairs. Of the cross language pairs, all trials were

conducted in English such that the conversation represented communication between a native and non-native speaker. The speakers were all drawn from the Wildcat corpus (Bradlow, Baker, Choi, Kim, & Van Engen, 2007) and the conversational pairs varied in the level of regional dialect matching. Similarly to Alshangiti and Evans (2011), they used early/late excerpts from the Diapix interaction itself in order to assess accommodation. They then took these excerpts and used them in an *AXB* paradigm in order to test perceived levels of accommodation. The early and late excerpts became the *A* and *B* elements of the paradigm. In order to reduce the working memory load on the participants (as the excerpts were slightly longer than single word examples), the stimuli were re-ordered into an *XAB* presentation order.

When these stimuli were presented to listeners to rate which of the *A* and *B* excerpts was closer to the *X* target accommodation patterns were found in the data of most of the interacting pairs, with a strong contributing factor from the language/dialectal distance of the interlocutors. Specifically, it was found that a match in regional dialect facilitated convergence whereas those pairs who did not share a regional accent or came from different native language backgrounds showed a lack of convergence. The authors point to the need for intelligibility and the increased perception/production demands of interacting with a non-native speaker as possible explanatory factors accounting for the results of their experimentation.

This paper presents some interesting findings and provides some carefully considered interpretations of the accommodation found in terms of the implications of the use of accommodation outside of the notion of social distance mediation. The fact that the authors are able to make reference to how accommodation might play a role in intelligibility demonstrates one of the key advantages of using a perceptual interaction approach. The participants are engaged in a task with a joint goal (ie. to find all of the differences between their two pictures) and are free to use their full linguistic repertoire to achieve this goal. Intelligibility is an important factor in the completion of this task as greater intelligibility will help to reach completion faster. In order to reach this conclusion, experimenters need to understand the relationship between the interlocutors and the likely motivating factors behind their accommodative acts. This conclusion could not have been achieved from a non-interactive stimulus since the necessary extra linguistic information and context is not available. What Kim et al. (2011) offer here is an insight that looks at a more fundamental aspect of accommodation concerning its use as a tool for communicative effectiveness.

Further to this, the assessment of accommodation is being made by making use of the human perceptual system rather than individual phonetic elements of the speech signal. This allows for all of the speech attributes contributing to accommodation to be accounted for. Although as the authors note, as a consequence of this they cannot trace the 'perceptual patterns to specific acoustic-phonetic features in the diapix recordings' (pp.144) and are therefore unable to qualify the perceptual notion

of accommodation. However, as they also note, this is a challenge that as of yet has not been resolved in the literature.

This paper offers some valuable insights not only into the relationship between interlocutor language/dialect distance but also into the possible broader implications of accommodation and the challenges still facing its measurement and interpretation. The authors also highlight that the fine-grained phonetic adaptations that are clearly observed in non-interaction studies become less clear in studies using interactional spontaneous speech. This is something that is of general interest to researchers in the field since it has important implications for the automaticity of accommodation. Nevertheless, the authors suggest that their findings demonstrate some interlocutor oriented speech adjustments are present in spontaneous speech and point to Pardo (2006) as further evidence of this.

In addition to the Diapix/DiapixUK task, researchers have also made wide use of the Map Task (Anderson et al., 1991) and this is the method of elicitation used in Pardo (2006). The Map Task is described in subsection 2.1.2 but a brief recap is provided here. The task consists of two maps, both of which have a series of landmarks on them but importantly, the two maps have a number of key differences between these landmarks. One of the maps has a path marked on it the other does not. It is the job of the person with the path to describe it to the person without the path so that they can draw it on their map as best as possible. The participants cannot see each others maps. The level of success is determined by the amount of deviation between the two paths. Pardo (2006) used this task in addition to a pre and post task reading task to evaluate accommodation. Prior to the participants taking part in the Map Task, they were asked to read a word list containing the phrases used to describe the landmarks in the maps. This word list reading was performed again upon completion of the Map Task. Recordings of the landmark phrases were extracted from the Map Task data and were used in conjunction with the pre and post task word list data as stimuli for an *AXB* perceptual listening task. Pardo (2006) had three different types of *AXB* stimuli that she used to evaluate accommodation, they are summarised in figure 2.2. In the figure, the terms 'Pre-Task' and 'Post-Task' refer to the word list reading tasks undertaken by participants before and after the Map Task, respectively. The term 'Sample' refers to speech samples of target landmarks taken of one speaker and 'Task-Rep' refers to the same target landmark speech sample taken from the other speaker. She used the 'Pre-Task - Sample - Task Rep' *AXB* trial structure to evaluate if speakers were rated as sounding more similar to their partner during the Map Task. The 'Post-Task - Sample - Task Rep' *AXB* trial structure was used to evaluate if speakers were rated as more similar to their partner during the task or during the post task. Finally, the 'Pre-task - Sample - Post-task' *AXB* trial structure was used to assess if convergence effects persisted into the post task.

Pardo (2006) found evidence for increased similarity in pronunciation during

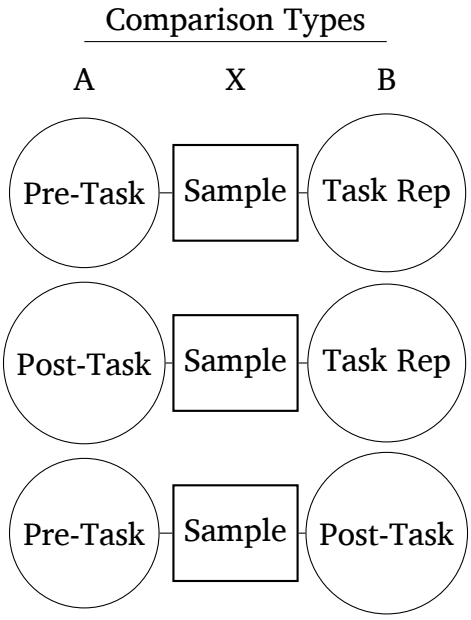



Figure 2.2: This figure presents the AXB trial structures used to assess accommodation in Pardo (2006). 'Pre-Task' refers to the word list speech recorded prior to the Map Task and 'Post-Task' refers to the word list speech recorded after the Map Task. 'Sample' refers to the speech sample taken of a landmark phrase produced by one speaker during the Map Task and 'Task Rep' refers to the landmark phrase produced by the other speaker during the Map Task. The figure is adapted from Pardo (2006, (pp.2386)).

the conversational interactions with a greater similarity detection rate for samples drawn from later in the conversational interactions. She also demonstrated that this similarity in pronunciation persisted beyond the conversational interaction into the post-task. These findings are taken as evidence for a rapid (the length of a conversation) process of convergence. However, it was also found that this process of convergence was modulated by the role that the speaker played in the Map Task (giver or receiver) and the sexes of the speakers. Both of which were found to differ from previously published research by Namy et al. (2002) and Nygaard and Queen (2000). Where previous research suggested that females were more likely to converge than males due to greater perceptual sensitivity or attention to the indexical features of the other speaker, results from Pardo (2006) suggest that the attentional factor may be more important than absolute perceptual sensitivity. Similarly, the findings concerning speaker role and the level of dominance associated with it also presented findings that differed from previous literature. These were interpreted as further evidence for the influence of functions from outside the domain of perceptual sensitivity on levels of convergence.

Where Kim et al. (2011) presented evidence for the automaticity of accommodation, Pardo (2006) makes a case for it being subject to situational constraints. She argues that the link between perception and production in spoken communication is not automatic and that since this is necessarily involved in accommodative behaviours the direction and magnitude of accommodation must also be, to a cer-

tain extent, non-automatic. Further to this, she also makes the point that whilst speakers may have been shown to converge as determined by a perceptual assessment of their speech, it is not possible to determine which phonetic features are being used to produce this phenomenon. Although the speakers demonstrate convergence, they will clearly not be tending towards similar realisations of all phonetic features. Understanding which aspects of speakers' phonetic repertoires are used to apply adaptation strategies and if they vary with respect to the situational constraints (eg. speaker role, sex) would be an important addition to the perceptual findings of interactional speech.

Additional work on the role of social norms in the realisation of accommodation was presented by Pardo et al. (2012). In this paper, the authors investigate the amount of accommodation found in college (university) room-mates, the details of which can be found in subsection 2.1.2. Although the paper provides assessments of accommodation with both perceptual and acoustic-phonetic measures (which is something that Pardo (2006) suggests to be appropriate), only the perceptual results will be discussed in this subsection. The authors recorded students at four intervals across an academic year, taking their first recording as a baseline measure of the participants' speech at the beginning of the year before they met their room-mates. These measures were taken by asking the participants to read five sets of American English vowels embedded in the carrier sentence, 'Say __ again'. Participants also provided two recordings of the two sentences, 'She had your dark suit in greasy wash water all year' and 'Don't ask me to carry an oily rag like that'. These were the stimuli used to present *AXB* paradigms to a separate set of listeners, again details of this can be found in subsection 2.1.2. The general findings of the perceptual tests for accommodation demonstrated no statistically significant increase in similarity between speakers across the time intervals studied. The detection rates for accommodation were above chance however. In addition, there was a good degree of difference in the levels of accommodation detected in each pair and for each of the presented speech phrases.

The variation found in both the detected levels of accommodation and the detected trends of accommodation over time may represent some of the complexities involved in the use of accommodation to serve social norms and in accessing and maintaining cultural communities. As the participants in this study progress through the academic year, there are a number of social and cultural impacts on their speech. Whilst the participants may be aiming to build and maintain a working relationship with their respective room-mates, they are at the same time faced with a number of other social groups and communities with which they will be looking to engage with (eg. their new course cohort, the new college societies that they have joined etc.). Each of these will have different linguistic norms and if participants are also aiming to accommodate towards these as well, it would be expected that variation in accommodation towards their room-mates would be found. In addition to the

social and cultural reasoning for finding variation in accommodation, the authors point out that the degree of convergence found was relatively modest. They go on to remark that if convergence is automatically evoked by perceptual resolution of phonetic forms (Pickering & Garrod, 2004) or by automatic activation of recent episodic traces (Goldinger, 1998) then speakers who live together should reliably converge and that this convergence should not vary across utterances.

To summarize, studies using perceptual interaction approaches to investigate accommodation provide a source of accommodation that is ecologically valid and measures it in a way that preserves the full range of speech attributes that contribute to it. It also allows for accommodation to be assessed from the perspective of both speakers in an interaction. However, it cannot provide information about the exact acoustic-phonetic features that contribute to accommodation and it also has some loss of experimental control given that the speech material being accommodated to is sourced from an interactive setting. The key findings that these works offer are that accommodation can be seen in short-term interactional settings such as conversations even though they present somewhat of a more noisy source of data than a more controlled non-interactional approach. They have also demonstrated that by using a perceptual interaction approach, insights can be gained into the broader social and task oriented goals associated with accommodative behaviours. However, there are some drawbacks of the approach and one key issue is that the methods of accommodation cannot be quantified in terms of the phonetic variations that speakers produce. Additionally, the signal from which accommodation is being assessed does not allow for the isolation and testing of a given phonetic form that might be theorised to be part of an accommodative strategy. The following section addresses the latter of these two drawbacks by considering the work that has been done using perceptual non-interaction approaches.

### 2.2.2 Perceptual non-interaction approaches

The work reported in this subsection concerns studies that have measured accommodation by making an assessment of the degree of perceived change in speakers' pronunciation in response to stimuli drawn from a non-interactive source. Similar to methods described in subsection 2.2.1, the assessment of accommodation is made by perceptual judgement of similarity to a stimuli for speech recorded before and after exposure to that stimuli. The important distinction between the methods presented in this subsection and those presented in the previous subsection is that the stimuli used is drawn from a non-interactive source (eg. word-list reading task). This subsection aims to outline the key findings of studies using this approach as well as detailing the advantages and drawbacks.

A good example of what is meant by 'perceptual non-interaction approaches' can be found in the work of Namy et al. (2002). Using the previously detailed

*AXB* paradigm (see opening to Section 2.2, specifically the part dealing with the key criteria labelled as *perceptual*), they investigated the role that sex (although the authors refer to this as gender) plays in accommodation. The findings of this study in relation to the impact of sex are reported in subsection 2.1.2 and are not the focus of this subsection. The fact that convergence was detected using this approach demonstrates an important aspect of accommodation. As the authors note the use of non-interactive stimuli allowed for them to test for accommodation when 'social interaction is severely limited, minimizing the opportunity for social motives to arise.' (pp. 424). Social motives are one of the key driving forces behind theories of accommodation although are not solely responsible for it (see: Gallois & Giles, 2015). By removing their participants from social engagement with each other and by making use of non-interactive stimuli, the authors are able to shift the focus of their study away from the social motives of the speakers. In this study they do this by having the shadowers repeat pre-recorded words rather than engaging in a live interaction. In this case, doing so allowed the authors to assess the impact of speaker sex on accommodation without the additional confound of social motives.

This line of thought can be taken somewhat further however, it could be argued that by restricting the shadowers' access to social information in the way that they did, the authors are actually able to look at the automatic nature of accommodation. That is to say that since speakers can be seen to accommodate when the main driving factor for accommodation is minimised, the drive to accommodate may function irrespective of the stimulus source and/or context. It suggests that accommodation may take place automatically. If this is the case then it could be that accommodation plays a more fundamental role in communication than previously thought. However, again as the authors point out, the results that they present may still be linked to socialisation and social motives. They point to differences in perceptual sensitivity between sexes as a contributing factor to accommodation that cannot be removed even when the opportunity to express social motives has been severely reduced. This can be evidenced through their use of perceptual rather than acoustic-phonetic measures.

As discussed in subsection 2.2.1, the use of perceptual measures rather than acoustic phonetic ones allows for the whole range of elements in the speech signal that contribute to accommodation to be taken account of. This is because experimenters are making use of the speakers' ability to produce and detect accommodation. There are likely to be individual differences in the ability of different speakers to detect and produce accommodation. The work of Namy et al. (2002) applies this reasoning to groups of speakers, looking at the ability of different sexes to both produce and perceive accommodation. They argue that their results not only demonstrate that females are more likely to accommodate but that they are also more likely to be able to detect accommodation. The latter of these two assertions is one that can only be made through the use of perceptual measures of accommodation. Assessing

accommodation using acoustic-phonetic approaches uses the speech signal to assess accommodation. Assertions about the perception of accommodation necessitate the use of a measure using the human perceptual system. Acoustic-phonetic measures can only make an assessment of levels of accommodation, they cannot assess the ability of a given person or group to detect it.

The authors demonstrated a difference between accommodation detection rates in females and males. They suggest that these differences may be related to differences in the way that males and females are socialised. It may be the case, they suggest, that females develop a greater perceptual sensitivity for accommodation due to reinforcement of the need to attend to indexical features of speech such as emotional tone of voice. If this is the case then it poses a problem not only for evaluating the perception of accommodation but also for evaluating the production of accommodation. Differing levels of perceptual sensitivity between groups such as males and females implies that the total available information about the speaker with which one is engaging may differ between people and groups. This means that the ability of a person or group to adequately accommodate in an appropriate manner would also vary since one can only accommodate relative to the information available. It could be argued that these effects will balance out given a large enough sample however, if studies such as this one find differences between sexes it could equally be argued that the perceptual sensitivity of certain groups may be a sensible consideration when planning experiments evaluating accommodation.

Further work addressing the issues of automaticity of accommodation and the perception of accommodation can be found in Shockley et al. (2004). The work presented by the authors consisted of two experiments. The first of these experiments aimed to replicate the findings of Goldinger (1998) which demonstrated that when asked to repeat a word heard over headphones, with no explicit instruction to imitate that word, repetitions were consistently judged to be more similar to the speech of the speaker producing the target word. In other words, it demonstrated that speakers accommodated to the speech used to produce the target word. The second experiment focused on expanding the findings to shed light on what aspects of the speech signal are imitated by speakers, specifically the authors looked at voice onset time (VOT). VOT is the 'time interval between stop occlusion and the onset of vocal fold oscillation' (Port & Preston, 1972, pp.126). Both of the experiments reported used the $AXB$ paradigm to make an assessment of speech similarity. The $X$ elements of the paradigm were produced from readings of a word list containing only bisyllabic words with word initial voiceless stops and therefore constitute a non-interactive stimulus source. The baseline speech sample for those participants performing the shadowing was also produced from a reading of the same word list.

The results of the first experiment provided support for the earlier findings of Goldinger (1998). It was found that speech produced after hearing the stimulus source, tended to be rated as more similar to the stimulus source than the baseline.

This finding also helped to broaden the claims that speakers tend to imitate those that they hear since the word list used to produce the *AXB* paradigm differed from that of Goldinger (1998). Although the findings of the first experiment broadly replicated the findings of Goldinger (1998) the authors did find an effect of presentation order. However, the effect of presentation order was not found in the second experiment. In the second experiment the authors artificially manipulated the length of the VOTs in the *X* element of the *AXB* paradigm so that they were twice as long as initially produced. The *A* and *B* elements of the paradigm were then produced as in the first experiment. When asked to rate which speech extract sounded more similar to the stimuli, *X*, listeners were found to consistently choose the speech extract that was produced after hearing the source stimuli. The authors also performed acoustic-phonetic analyses of these data but they are not discussed here since they constitute a different approach to the assessment of accommodation (see subsection 2.2.3 for this information).

Taken together, the findings of Shockley et al. (2004) provide further support for the assertion that accommodation takes place even in the absence, or at least restriction of, social motives. The inclusion of an experiment to assess if VOT is imitated allowed the authors to begin to evaluate what elements of the speech signal might be being imitated in accommodative acts. However, the fact that listeners rated the post-exposure speech as more similar to the stimuli with extended VOTs does not necessarily indicate that VOTs are being imitated specifically. It could be the case that this is simply the same effect that was being seen in the first experiment. In order to evaluate if VOTs are actually being imitated, an acoustic-phonetic analysis is required. Nevertheless, the findings from the purely perceptual analyses of this data allowed the authors to reason that since the majority of the social motivation to accommodate has been removed, there must be an additional reason to accommodate. The suggestion offered for this is that accommodation is linked to the vocal tract gestures made by the speaker. The reasoning provided draws on the motor theory of speech perception (Liberman & Mattingly, 1985) and suggests that as a listener hears a speech sound, their own speech motor system is partially activated as part of the process used to evaluate the incoming speech signal. This activation leaves a trace in the memory of the listener and this trace influences upcoming utterances, thus leading to a degree of accommodation. However, in this paper the authors stated that 'Speakers 'accommodate' (converge) in their accents, speaking rates, rates of pausing and vocal intensity' (Shockley et al., 2004, pp.422) this wording suggests that the authors consider accommodation as a tendency to shift pronunciation towards the interlocutor. Indeed this is reflected in the wording that they use to describe accommodative behaviours, referring to it mainly as 'imitation'. When considered in this light the above explanation of accommodation fits very nicely but directions of accommodation other than convergence are more difficult to reconcile with this explanation. Having said that, the authors do also note

that alternative explanations might include imitation that is linked more closely to the acoustic signal that is produced rather than the underlying gestural information contained in the signal. The underlying cause for this, as the authors consider it, is inherently social. The example they offer is that imitation of other speakers is important in guiding children's entry to a cultural community. If imitation can be considered to play a role in social and cultural integration, the authors argue that it might be the case that speakers are generally predisposed to imitate.

The benefits of using a perceptual non-interaction approach as highlighted by the studies reviewed in this subsection include the ability to evaluate accommodation as a composite of all the relevant acoustic features that contribute to it and the ability to restrict the potential for social influence on accommodation to a minimum. Used together, these two benefits mean that studies using this approach of assessment can begin to answer questions concerning the automaticity of accommodation and if there are additional factors that drive accommodation beyond those detailed through a social explanation. However, these approaches cannot make an assessment of the aspects of the speech signal that speakers employ to produce accommodative behaviour and they cannot make assertions regarding the role of accommodation in genuine human communicative interactions. The following subsection, subsection 2.2.3, addresses one of these drawbacks by looking at what can be learned about accommodation through the use of acoustic-phonetic measures.

### 2.2.3 Acoustic-phonetic non-interaction approaches

One of the key advantages of employing an approach that utilises perceptual measures of accommodation (as discussed in subsections 2.2.1 and 2.2.2) is that all aspects of the speech signal pertaining to accommodation can be said to be considered. However, the ability to determine which aspects of the speech signal are being manipulated by speakers to produce accommodative behaviour cannot be determined through perceptual measures alone. In order to do this, acoustic-phonetic measures are required. This subsection deals with those studies that employ acoustic-phonetic measures of accommodation whilst using stimuli drawn from non-interactive sources to elicit accommodation. The aims for this subsection are to report the main findings of these studies and to consider their advantages and drawbacks. Since this subsection deals with acoustic-phonetic measures of accommodation, it is divided in relation to the particular acoustic-phonetic feature being considered. The acoustic-phonetic features focused on here are voice onset time (VOT), vowels (as measured by the first two formant frequencies, F1 and F2), speech rate and fundamental frequency (F0). There are a some additional measures that could also be discussed (eg. prosody/syllable timing) but there is more work regarding accommodation for the acoustic-phonetic measures chosen. A discussion of additional measures would not add much value to this review of acoustic-phonetic

measures of accommodation.

**VOT**

The work of Shockley et al. (2004) was discussed in subsection 2.2.2 but only the findings of the perceptual study were considered. Here, the additional findings concerning the VOT measurements reported in the study will be discussed. To briefly recap, Shockley et al. (2004) utilized the *AXB* paradigm to both replicate the findings of Goldinger (1998) whilst also looking to evaluate to degree to which VOT is imitated. The key manipulation here was that the target speakers' utterances, *X*, were adapted such that the VOTs were extended to twice their original length. This made the utterances sound 'noticeably breathier than the original productions. Although they did not sound unnatural' (pp.425). The reasoning behind this manipulation was to see if VOT is an aspect of speech that is imitated by others. If it is, then the listeners would rate the post-exposure recording as more similar to *X* than the pre-exposure condition where the VOT was not manipulated.

The authors found significantly longer VOTs in the post-exposure condition than for the baseline condition. However, they also found the same result in their first experiment in which the VOT of speakers had not been manipulated. The authors interpret this finding as possibly being due to the different manners in which the tokens were collected. The baseline tokens were read whereas the post-exposure tokens followed spoken words. They propose that their instructions to speak 'quickly but clearly' may not have been enough to overcome the different speaking styles of read and shadowed speech. In addition, upon closer inspection of the data, it was found that the target *X* tokens in the first experiment tended to have a longer VOT that the baseline VOTs of the shadowers. In any case, given that the VOTs for the *X* token in the second experiment was extended, the difference between the baseline and post-exposure conditions in the second experiment should be larger than the difference between the baseline and post-exposure conditions in the first experiment. In order to evaluate this the authors compared the differences in VOTs for the baseline and post-exposure conditions between experiments. Results demonstrated that the difference between baseline and post-exposure VOTs from the second experiment were significantly larger than the differences found in the first experiment. The authors were particularly rigorous in their testing and also tested to see if the extended VOTs found in the second experiment were due to the influence of a few, faithfully reproduced tokens, this was not the case. Further to this, the effect of word duration was evaluated. Authors found that post-exposure words tended to be longer than baseline words but that this increase in duration did not differ significantly beyond the lengthening due to VOT increase. Taken as a whole the findings provide support to the hypothesis that VOT is a phonetic feature of accommodation.

Although this paper provides support for VOT being part of the accommodative process, the contexts in which VOT is realised are somewhat restricted and it

is difficult to extend these findings to a more generalised view of VOT. Additionally, it only considers adaptation to VOT in a single dimension, lengthening. The fact that the shadowers were exposed to extended VOTs multiple times may have compounded the amount of exposure, leading to greater convergence. The work of Nielsen (2011) provides additional insight into this issue by looking at accommodation to both artificially lengthened and artificially shortened VOTs.

Nielsen (2011) was interested in evaluating accommodation of VOT across a number of different areas. She looked to ascertain if accommodation was present at the sub-lexical level, if lexical frequency had an impact on degree of accommodation and if accommodation is phonetically selective. The approach that she took to investigate this involved two experiments, one in which VOTs were extended and one in which they were shortened. Both experiments source their target stimuli from the same phonetically trained male speaker. This speaker was asked to read aloud 80 target words, all of which began with the phoneme /p/. Half of these words were low-frequency words and half were high frequency words. The speaker was asked to read these words twice with normal aspiration and twice with extended aspiration. The most subjectively clear of these readings were used to construct the artificially lengthened/shortened stimuli. To keep the lengthening of the VOTs consistent, the burst and aspiration of the normally aspirated recordings were replaced with a section of that from the recordings with extended aspiration such that all VOTs were extended by exactly $40\,\mathrm{ms}$. This approach was taken to best preserve natural formant transitions. To produce the shortened stimuli, the normally aspirated recordings were simply trimmed by exactly $40\,\mathrm{ms}$.

The procedures for both experiments were the same except that experiment one used lengthened VOT stimuli and experiment two used shortened VOT stimuli. The procedure consisted of four blocks: warm-up reading, baseline recording, target exposure and post-exposure recording. In the warm-up reading block participants were visually presented with a list of words and were asked to read them silently without pronouncing them. This was done to help mitigate the effects of hyperarticulation of low-frequency words. The list of words presented to the participants consisted of 150 words containing 120 test words (80 matching the stimuli, 20 novel /p/ initial, 20 novel /k/ initial) and 30 filler words. The baseline recording block the same words were again visually presented to the participants but this time they were asked to read them aloud. The target exposure block had the participants listen to two repetitions of the artificially manipulated target words along with 40 filler words to make for a total of 120 words. Extended VOTs were used in experiment one and shortened VOTs in experiment two. The final block, post-exposure recording, participants were asked to perform the same task as in the baseline recording block. Importantly, at no point were participants explicitly asked to imitate the speech of the target speaker. Unlike experiments utilising the *AXB* paradigm, the participants were not asked to repeat the words directly after hearing them.

Results of experiment one demonstrated that post-exposure VOTs for the participants were consistently longer than those for baseline. This was found to be true for the target /p/ initial target words as well as the novel /p/ and /k/ initial words. This finding suggests that accommodation not only takes place in the words that are being presented to participants but is also generalised at a sub-lexical level. The author also tested for differences between the levels of imitation between the novel /p/ and novel /k/ categories and found a significant difference. This is taken to suggest that the extended VOT was coded at both the phoneme and feature levels and thus impacted /p/ and /k/ differently. Further to this, the author also found an effect of lexical frequency with low-frequency and target words demonstrating stronger imitation effects. The results of experiment two were quite different to that of the first experiment. Here, when the same procedure was applied with shortened VOTs, the author found no evidence for VOT imitation.

These results demonstrate some interesting findings regarding the mechanisms of accommodation. The fact that this work did not ask its participants to repeat the words that they heard immediately after hearing them lend support to accommodation persisting beyond the initial interaction. The results from experiment one provide evidence of word, phoneme and feature-level representations playing a role in accommodation with adaptation taking place across all of these. It shows that there is an interplay between phonemic categories since adaptation was found to extend to /k/ and that this adaptation level differed from patterns of adaptation found for /p/. This finding is further reinforced by the results of experiment two in which no adaptation was found towards shortened VOTs. In the paper it is suggested that this may be the result of potential conflicts in voiceless versus voiced phoneme categories (ie. /p/ vs. /b/) and that participants do not accommodate towards a shortened VOT in order to maintain a perceptual distinction between these categories. This suggests that adaptation may not be as automatic as suggested by Shockley et al. (2004). Whilst the author does not suggest that her findings are incompatible with the gestural theories for accommodation proposed by Shockley et al. (2004), she does make it clear that she considers the process to be selective and modulated by linguistic factors. One aspect that was consistent across both experiments presented in Nielsen (2011) was that there was a good degree of individual variability found in VOT differences between baseline and post-exposure conditions. However, only a cursory interpretation of these findings is offered. The author says only that it replicates previous findings and that further work is required to ascertain the role of this variability in accommodation. It is likely that individual variation is found in all experiments considering human behaviour. However, given that the biggest factors influencing accommodation are said to be social (Giles, 1973; Giles et al., 1973; Bourhis & Giles, 1977), the attitudes and social ratings that speakers have of each other may well play a significant role in accommodation. The work of Yu et al. (2013) set out to investigate this.

Yu et al. (2013) considered the individual variability of speakers and how this might apply to convergence. The experiment they ran consisted of three blocks: a baseline production block, a listening block and a post-task test block. The baseline production block had participants produce a list of 72 $\mathrm{p/t/k}$-initial target words in a carrier phrase. For the listening block of the experiment participants were separated into one of four groups and each group listened to a different version of a scripted first-person narrative as read by the same male talker. The narrative recounted a blind-date and was edited such that it either had 'positive' or 'negative' outcome and was from the perspective of a 'heterosexual' or 'homosexual' narrator. This provided the four groups for the participants: 'positive-heterosexual','positive-homosexual','negative-heterosexual' and 'negative-homosexual'. The narrative used contained all of the 72 $\mathrm{p/t/k}$-initial target words used in the baseline production block. In all instances, the VOTs of the male speaker were extended by $100\,\%$. For this listening block, the task for the participants was simply to listen to the narrative. Given that a pre-recorded narrative was used to elicit accommodation, this can be considered as a non-interactional stimulus source. The final post-task test block was a repetition of the task in the baseline production block. Following these three blocks participants were asked to take a number of tests and questionnaires. The participants were asked to complete the Automated Reading Span Task, to measure working memory capacity, the Big Five Inventory (BFI), to assess the personality traits of the participants and the Autism Spectrum Quotient, to assess the degree to which participants demonstrated autistic traits.

The authors were interested in addressing two questions. The first of these concerned the general effect that the narrative had on the participants' VOTs. The second question looked at determining what factors affect the amount of VOT shift in the participants. In order to do this, the authors made use of linear mixed effects modelling (MEM) implemented using the $\mathrm{lmer}$ function from the $\mathrm{lme4}$ package (Bates, Mächler, & Bolker, 2011) in R (R Core Team, 2016). This method of statistical analysis allowed the authors to ascertain the general effect of the narrative on VOT whilst controlling for additional factors such as speaking rate, place of articulation and lexical frequency. Doing so allowed them to address both questions with a single model. In order to do this they employed a two-step approach. The first step of this approach was to produce a MEM that accounted for the effects of all word and utterance-level properties *except* for exposure to the narrative. This provided the authors with residuals of this model that represented the data stripped of the word and utterance-level effects on VOT, allowing them to more accurately determine the effect of subject-level properties on VOT change due to exposure to the narrative. Interpretation of the first model confirmed previously established findings about the effect of properties such as place of articulation and speaking rate on VOT (Cho & Ladefoged, 1999; Miller, Green, & Reeves, 1986). The second model that the authors reported was more informative in terms of the questions

that they were asking and reported the effect of narrative exposure on VOT values. This model reported no overall effect of narrative exposure on VOT lengthening, that is to say that simply hearing a narrative with lengthened VOTs was not a core predictor of participants lengthening their own VOTs. However, the subject-level predictors of attitude, openness, narrative outcome and attention switching were all found to significantly impact on VOT shifting. Participants with a more positive attitude of the narrator tended to be more likely to shift their VOTs towards those of the narrator than those participants with a more negative attitude. Those with higher levels of openness, as determined by the BFI, tended to be more likely to shift their VOTs towards that of the narrator, those with lower openness scores tended not to. The outcome of the narrative, recall that the narrative either had a positive or negative outcome, also had an effect on VOT. It was found that the negative narrative showed an average increase of $2.58\,\text{ms}$ VOT compared to the positive narrative. Finally, attention switching also had a significant effect on VOT. Higher levels of attention switching were associated with longer VOTs.

The authors take these findings as evidence that shifts in VOT between the production blocks may be modulated by 'disincentives and obstacles that conflict with goals, attention, and liking' (pp. 10). To justify this conclusion, they point to the significant influence that attitudes towards the narrator and narrative outcome had on VOT imitation. They make the point that increased liking and negative narrative content, both of which showed increased VOT imitation, seem like counter-intuitive companions in accommodative behaviour. However, they point to work on *automatic vigilance* (Pratto & John, 1991), which posits that there is an unintentional attentional bias towards undesirable stimuli which may be driving the effect of the narrative outcome. So, irrespective of the attitude that the participant has towards the narrator, which still has a significant effect on VOT in itself, there is an additional unintentional effect associated with attention to negative stimuli. Ultimately, the authors suggest that the influence of the additional significant properties identified by the model relate to the specific social and cognitive make-ups of the participants themselves. The authors further suggest that these findings are not completely compatible with simple automatic gestural accounts of imitation since simple exposure to another speaker did not elicit imitation of VOTs.

This study stands in contrast to that of Nielsen (2011), where imitation as a result of exposure was found. Yu et al. (2013) suggest that this may well be due to the type of exposure that the participants were exposed to. Whereas Nielsen (2011) presented participants with words in isolation, the present study under discussion had the target words embedded in a meaningful narrative. The authors suggest that it is the use of a meaningful narrative that allows participants to make evaluative judgements about the narrator and to therefore introduce some social mechanisms for the mediation of the automatic imitative responses to stimuli found in Nielsen (2011). Additionally, the statistical techniques used in the two studies differ con-

siderably and the authors point to this as another potential source of variation. However, the point that the authors make regarding content of the stimuli used to elicit accommodative behaviours is an important one. Whilst studies that make use of isolated target words that lack social meaning provide useful evidence for the potential underlying mechanisms of accommodation, researchers must be careful not to overstate the findings without consideration of the role that social context plays in communicative acts. An ability to evaluate the driving mechanisms behind accommodation whilst participants are involved in a social engagement would go quite some way to contributing to the discussion surrounding accommodation.

#### Vowels

Further evidence for the influence of social factors on accommodation can be found in the work of Babel (2010). In her work, the effect of both implicit social attitudes towards a specific social group as well as explicit flattery and insult were considered as potential modulating factors on the accommodation of vowels in New Zealand English (NZE). For this experiment participants, all NZE speakers, were separated into two groups, positive and negative, and each group took part in a three block auditory naming task. The first block, the pre-task block, was the same for both groups. Here participants were visually presented with a series of hVd words (eg. hid, had, head) followed by a target word list, participants were tasked with reading them aloud. The second block, the shadowing block, varied between the groups. For the positive condition, the participants were presented with a text which described the target speaker as having a positive attitude towards New Zealand. For the negative condition, the participants were presented with a text which described the target speaker as having a negative attitude towards New Zealand. After reading the presented text, participants were aurally exposed to isolated target words spoken by the Australian English (AuE) target speaker. The task for the participants was to identify the word by speaking it aloud. Participants then completed the third block of the experiment, the post-task block, which consisted of the same task as the pre-task block. Following this, all participants completed an Implicit Association Task (IAT; Greenwald, McGhee, & Schwartz, 1998) designed to assess participants' implicit and subconscious attitudes towards both Australia and New Zealand.

The results from this study considered both the speech produced during the shadowing block and the speech produced during the post-task block for evidence of accommodation. The vowels produced by the participants in the shadowing block showed an overall trend of convergence towards the AuE speaker although all vowels were not convergent to the same extent. Both the outcome of the IAT and word-frequency were significant predictors of the shifts in vowel realisation. As IAT scores shifted towards a pro-Australia bias, the shift towards the speech of the AuE speaker was greater. Word-frequency demonstrated the same trend as seen in Goldinger (1998), as the word-frequency became smaller, the degree of shift

towards the AuE speaker became greater. For the post-task block productions, the results demonstrated that participants retained the more Australian-like vowels that they had converged towards in the shadowing block. However, unlike the results from the shadowing block, the only significant predictor for this observed shift was IAT score and there was no observed difference between the degree of convergence between the vowels. In both the productions from the shadowing block and the post-task block there was no significant effect of participant group (ie. AuE speaker having either a positive or negative attitude towards New Zealand).

The author interprets the results of this experiment as a evidence for accommodation being automatic but with certain nuances. It is suggested that accommodation, here specifically convergence, is automatic in so far as the participants do not know that they are doing it and that any social biases impacting on accommodation are not elicited from explicit decision making (Dijksterhuis & Bargh, 2001). However, it is pointed out that the work provides evidence that the process does not occur at all times and that biases impacting on accommodation exist prior to interaction (IAT scores were the only persistent predictor of accommodation). So, unlike the suggestions made by some works that automatic accommodation can lead to the development of social ties (eg. Trudgill, 2008), the work presented by Babel (2010) suggests that whilst speakers may naturally tend towards accommodation, it is modulated by group-identity attitudes. The use of IATs in this study has allowed for some tentative progress to be made towards determining the directionality of the relationship between social factors and accommodation. This work is built upon by Babel (2012) where further investigation of the effect of social factors on accommodation is offered and further evidence for selective imitation of vowels is presented.

Babel (2012) presents an experiment with three core goals, (1) to expand understanding of what phonetic features are imitated, (2) to assess if different levels of social information lead to different behaviours and (3) to evaluate if the degree of liking, as measured by social attractiveness, impacts on levels of accommodation. The study looked exclusively at the vowels /i æ ɑ o u/ measuring the degree of accommodation by way of an *AXB* paradigm. The procedure itself consisted of a pre-test reading task where participants (111 participants, 51 male, all self-identified white) were presented with a sequence of words on a screen and were asked to read them as accurately and clearly as possible. This was used to gain a baseline measure for each of the participants. The pre-test reading task was followed by three *AXB* shadowing tasks where participants listened to words produced by model speakers and were asked to repeat the word as clearly and naturally as possible. The model speakers were both male, one was white and the other was black, both spoke the same standard variety of California English. Within these shadowing blocks the participants were placed in one of three categories. They were either presented with an image of the speaker they were listening to or they received no additional visual stimuli. The final stage of the experiment was the post-task, which was the same

as the pre-task. After the experiment had finished, participants who received no visual prompt were asked to identify the race of the speaker and participants who did receive a visual prompt were asked to rate the attractiveness of the model speaker. The participants in the condition with no visual prompt could not reliably identify the voices of the speaker they heard as either black or white.

The author of this paper considered a number of potential impacting factors including speaker voice, available social information (visual prompt), vowel type, the effect of testing block and participant sex. The high number of factors involved in this study meant that the experimental design was highly complex (a $2 \times 2 \times 5 \times 4 \times 2$ design) this made results difficult to interpret directly but the author handles the complexity of the design well in the analyses. The general trend of the participants was convergence towards the model speaker with each of the tested vowels (/i æ ɑ o u/) individually demonstrating a significant amount of convergence. However, given the complex nature of the experiment, it was possible to explore the differences and patterns across groups. An analysis of the speech data collected from the participants with no visual prompt demonstrated that imitation of the vowels occurred more strongly in shadowing blocks than in post-task blocks. It was also found that there were vowel specific patterns of accommodation which demonstrated an interaction effect with both block and participant sex. These findings demonstrated that there was more convergence in earlier (shadowing) blocks than in later (post-task) blocks and that there were differences in the accommodation strategies of the participant sexes. Males were found to converge more to /ɑ/ than females and that females would converge more to /æ/ than males.

The results for the participants in the visual prompt condition are more complex than that of the no visual prompt condition but more convergence was found overall in this condition than in the no visual prompt condition. There were a number of complex interactions between the factors of interest but the main trends found in the no visual prompt conditions were also found here. That is to say that shadowing blocks showed more convergence than post-task blocks and that accommodation varied by vowel. The factors of speaker voice and participant sex interacted with the effects of block and vowel. This suggests that accommodation is vowel specific, changes for different voices and is different for males and females. Of course, further to this, each of these factors impacts on each other as well to provide an interrelated picture of accommodation in this context. However, it was found that overall /æ/ was converged towards more than other vowels and that /ɑ/ was converged to more than /i/ and /o/. The author notes that the female participants do not ever directly overlap with the model speaker in the production of their vowels, which is to be expected given physiological differences, but would travel a further comparative distance towards the model speaker than the males. In addition, it is also noted that for the most converged towards vowel, /æ/, the males in the no visual prompt condition had a complete overlap with the vowel production of the white speaker.

However, this is not seen as a quantitatively large shift towards the model speaker due to the relative starting point of the male speakers in the vowel space.

The data on the degree of liking, as measured by attractiveness ratings, was also evaluated. No significant trends were found for the visual prompts containing the black speaker. For the visual prompts containing the white speaker, it was found that females would tend to converge more when they rated the speaker as more attractive. The male participants on the other hand presented the opposite trend and would converge less when they rated the white speaker as more attractive. For the author, this is in keeping with behaviour as predicted by CAT for the female participants. They more they like the speaker, the more they wish to reduce social distance leading to greater levels of convergence. For the male participants, the authors suggests that the divergent behaviour may be the result of a threat response. Those males who rated the speaker as highly attractive may have felt socially threatened by that speaker and tried to distance themselves. Another explanation the author offers is that perhaps attractiveness ratings of male speakers by females and males do not constitute the same measure.

Ultimately, the author of this paper concludes that given the tendency of low vowels to be converged towards, that accommodation is phonetically selective. However, given that the sex of participants, the degree of liking and speaker voice all played roles in the accommodation seen in the participants, that it is also socially motivated. The conclusions for the three goals that the paper set out to achieve were as follows, (1) vowels are imitated in accommodation, (2) the presence of additional social information does impact on that imitation, albeit in a complex way, and (3) degree of liking also impacts on the imitation seen in accommodation although again, in a complex way.

#### Fundamental frequency (F0)

Before making any considerations about the link between fundamental frequency (F0) and accommodation, a brief explanation of the relationship between F0 and pitch will be provided. It is important to understand that *pitch* refers to a perceptual experience whereas F0 describes the lowest frequency of a periodic or quasi-periodic wave. In the case of speech, these values vary by individual and sex but tend to fall in the region of 60 to 300 Hz (Bulatov, 2009, pp.410). Pitch on the other hand is the perceptual result of hearing the fundamental frequency. Whilst pitch and F0 are related, the distinction between F0 as an acoustic measure and pitch as a perceptual measure is an important one. Here, it is the acoustic measure of F0 that is being considered.

Babel and Bulatov (2012) provide both perceptual and acoustic-phonetic measures of accommodation. The acoustic measure of choice for the authors was F0, this is what is discussed here. The perceptual measure that they took was related to the presence or absence of F0 but was not a measure of pitch per se.

The paper arranged for twenty-two native speakers of American English (7 male, 15 female) to take part in a shadowing task. The participants were required to take part in a pre-task to elicit baseline responses. This consisted of reading a word list of thirty-nine words. The data from the pre-task was only used for the perceptual assessment. The shadowing task had participants split into two groups. One group heard repetitions of the same thirty-nine words, produced by a pre-recorded model speaker, used in the pre-task without any modifications, the unfiltered condition. The other group heard words that had been high-pass filtered at 300 Hz to remove the frequencies in the speech signal that constitute potential F0s, filtered condition. Both groups were asked to repeat the word that they heard as naturally as possible immediately after hearing it. There were three shadowing blocks for each of the groups. The two groups then had to complete a post-task that consisted of the same content as the pre-task.

The results of the acoustic-phonetic analysis demonstrated that participants in the unfiltered condition, in general, tended to converge towards the model speaker, with their own F0 becoming more similar to that of the model speaker. However, the participants in the filtered condition had a tendency to demonstrate divergent behaviour, with their F0 becoming more dissimilar to that of the model speaker. Whilst the general trends seen here were statistically significant, the effect sizes were small. Further to this, the trends for convergence and divergence were modulated by a number of other factors including participant sex and task block. It was found that males would tend to converge more in terms of their F0 than women. In the filtered condition, greater divergence was found in the first and third shadowing blocks.

The authors note that the findings of the acoustic-phonetic analyses are somewhat difficult to interpret due to the fact that it is unclear if the participants in the filtered condition are accommodating in relation to F0 or rather to some other feature of the speech signal that was introduced as a result of the filtering (eg. unnatural sounding speech). Additionally, they note that it is possible for participants to infer the fundamental frequency from the remaining harmonics present in the speech signal. However, they point out that due to the differences between the filtered and unfiltered conditions, this is unlikely to have happened since if this were the case then the difference would have been less prominent. Having said this, the authors make it clear that their data provide evidence that F0 is used in accommodation but that there are a number of other factors that also influence it. They suggest that taking other acoustic-phonetic measures alongside F0 to get a more complete picture of accommodative effects.

Much like the studies presented in subsection 2.2.2, the key benefit of the way in which studies presented here make assessments of accommodation is that a good degree of control can be exercised over the stimuli used to elicit accommodation. The use of acoustic-phonetic measures allows for more accurate pinpointing of the

elements of speech that are used in accommodation. However, as has been evidenced in the work presented above, even with the benefit of being able to control the stimulus eliciting the phenomenon, the effects are often complex. For the most part, the use of a single acoustic-phonetic measure has been shown not to capture all the necessary information that makes up accommodation. In addition, the use of a non-interactive method for elicitation of accommodative behaviours only allows for the behaviour of one speaker to be evaluated. In order to understand accommodation as an joint phenomenon, an interactive approach needs to be taken.

### 2.2.4 Acoustic-phonetic interaction approaches

The final type of measurement approach for accommodation that will be covered in this thesis will be referred to as the 'acoustic-phonetic interaction approach'. The previous subsection, subsection 2.2.3, concerned approaches that measured accommodation using acoustic-phonetic measures and used stimuli drawn from non-interactional sources to elicit accommodation. This subsection deals with studies that perform the measurement of accommodation using acoustic-phonetic measures but that elicit accommodation from interactional stimuli. As with the previous subsections concerning the measurement approaches used to evaluate accommodation, the aims of this subsection are to demonstrate the key findings of studies using this approach and to highlight their benefits and drawbacks. Again, because the use of acoustic-phonetic measures can refer to a number of different acoustic-phonetic features drawn from the speech signal, this subsection is grouped by the features of interest. The acoustic-phonetic features discussed here are the same as those discussed in subsection 2.2.3.

#### Vowels

The complex interplay between the accommodation of different vowel types that was found in the studies presented in subsection 2.2.3 is also seen in studies that use an interactional source to elicit accommodation.

Purnell (2009) looked at the role that long-term versus short-term knowledge of the linguistic norms for a given speech community has on accommodative behaviour in vowels. In order to investigate this, the author turned to the linguistic diversity found in south-eastern Wisconsin in the USA. Specifically, the study drew on the differences between the African American community in the area and the white community in the area. The study made use of data that the author had collected from 18 participants (9 black, 9 white), although only data from three of the white speakers was presented in this particular paper. The participants were split into four groups. The first group consisted of three members of the African American English (AAE) speaking community that had not seen much exposure to the local white linguistic norms. In this group, the participants were interviewed by a member of

their own linguistic community, another AAE speaker. This group was coded the $B-B$ group. Another group consisted of three white speakers being interviewed by a white speaker. This group was coded as the $W-W$ group. Both the $B-B$ and the $W-W$ groups served as points of reference for the two test groups. The two test groups consisted of a core group and a peripheral group. The core group had three members of the AAE speaking community, with little exposure to the local white linguistic norms being interviewed by a member of the local white community. This group was coded as the $B_{core}-W$ group. The peripheral group had three members of the AAE speaking community, with high levels of exposure to the local white linguistic norms being interviewed by a member of the local white community. This group was coded as the $B_{periph}-W$ group. The experimental sessions for the participants consisted of free conversation, word list and sentence reading. For the analyses in this paper, the data from the free conversations was used with word list data used if the free conversation data was not of sufficient quality/did not exist.

The analyses focused on a number of different vowel monophthongs and diphthongs that had been shown to be suggestive of either AAE or local white norms. The analysis focused on the Lobanov normalised (Lobanov, 1971) F1 and F2 values at the head and tail (onset and offset) of the vowels produced. The author puts forward a number of possible hypotheses for what one might expect. The null hypothesis ($H_0$) is that none of the AAE speakers will show any difference in their vowel realisations in comparison to the white speakers. The first hypothesis that the author puts forward is what he calls the *ethnicity hypothesis* ($H_{eth}$) which posits that all AAE speakers will maintain their speech patterning irrespective of their speech partner. This hypothesis draws on the AAE speakers wanting to emphasise and identify with their group membership. The second hypothesis that the author puts forward is the *interlocutor hypothesis* ($H_{inter}$), this hypothesis suggests that conversations with the white speaker will demonstrate convergence. This hypothesis relies on the automatic need for speakers to be co-operative, thereby adjusting their speech patterning towards that of their interlocutor in order to achieve this. The final hypothesis put forward by the author is the *knowledge hypothesis* ($H_{know}$) which suggests that each group containing an AAE speaker will demonstrate a distinct pattern of accommodation. The reasoning behind this hypothesis is that the level of exposure to the local white norms that the AAE speakers in the $B_{core}-W$ group will have had is, for the most part, limited to that which they have experienced during the experiment. The AAE speakers in the $B_{periph}-W$ group on the other hand are known to have had a good deal of exposure to the local white norms and would therefore have a wider repository of phonetic norms which they know to accommodate towards.

The findings of the paper allowed the author to reject $H_0$, $H_{eth}$ and $H_{inter}$. It was found that there were different patterns of accommodation for each of the vowels evaluated and for each of the two test groups, $B_{core}-W$ and $B_{periph}-W$. Since the groups were found to differ from the white speakers, the $H_0$ can be rejected. All

of the members of the AAE community did not participate in maintenance of their AAE vowel features so the $H_{eth}$ hypothesis can be rejected. Differences were found between the $B_{core} - W$ and $B_{periph} - W$ groups. Because the identical co-operation, as observed through levels of accommodation, in relation to the ethnicity of the interviewer were not observed, the $H_{inter}$ can be rejected. The data from this paper support the final hypothesis put forward by the author. That the level of linguistic knowledge that a person has about the community that their interlocutor comes from impacts on how they accommodate. It was found that for the same vowel shifts the AAE speakers in the $B_{periph} - W$ group would converge to a high degree whilst the AAE speakers in the $B_{core} - W$ group would simply maintain the norms of the local community without putting emphasis on the AAE specific features. In a different set of vowel shifts it was found that the AAE speakers in the $B_{core} - W$ group would tend to diverge whilst for the same vowel shift the AAE speakers in the $B_{periph} - W$ group would demonstrate a maintenance of the AAE community norm features.

The work presented by Purnell (2009) adds to the complexity that is seen in the accommodation of vowels. His work demonstrates that there is an interplay between both the linguistic community in which one has been raised and the knowledge that one has of the linguistic norms (or perceived linguistic norms) of the community that the interlocutor comes from. The work points to a continuum of factors that impact on types of accommodation that can be observed in conversational speech and provides support for the notion of external social influences on accommodation as evidenced by other studies (eg. Babel, 2010).

The complex interaction of situational/social factors with the realisation of accommodative behaviours is found also by Pardo et al. (2010). This paper took a hybrid approach to measuring accommodation and conducted a perceptual analysis of accommodation along with drawing acoustic-phonetic measures. Only the acoustic-phonetic measures associated with the measurement of vowels will be reported here.

The experiment involved a set of 12 same sex speaker pairs (6 female pairs, 6 male pairs), each of which were paired such that each member of the pair were from different dialect regions of the USA. They were all unacquainted pairs. The task that the participants were asked to undertake was split into a pre-task, a conversational task and a post-task. The pre-task saw the participants produce two sets of baseline speech samples. One was produced by reading a list of the landmark phrases from the Map Task (Anderson et al., 1991) embedded in a carrier sentence. The other baseline speech sample was produced by reading another carrier sentence containing one of a number of items that contained nine vowels of American English. The conversational task consisted of the Map Task (see section 2.1.2 for details). The only variation on the traditional set-up of the Map Task was that in three of the pairs in both the male and female groups the instruction giver was instructed to imitate

the other member of the pair without them knowing. For the other three pairs in each group, it was the instruction receiver that was instructed to imitate. Finally, the participants took part in the post-task which was identical to the pre-task.

The results of the *AXB* perceptual tests demonstrated that convergence of the speakers persisted into the post-task so the authors proceeded to analyse the vowel spectra of the pre and post-task vowels. The authors note that it would have been useful to have sampled vowels from the conversational task but that the sampling of vowel tokens in the conversations was limited and could have led to potentially unreliable measures. The results of the vowel comparisons demonstrated no overall convergence in vowel spectra. In addition, it was found that instructing the instruction givers to imitate actually led to divergence, in general. When instruction receivers were instructed to imitate, no significant change in either direction was found.

The results of this paper are interesting and somewhat in contrast to those reported in other studies concerned with accommodation of vowels. The authors of the paper remark that because their results are limited in scope, the results should be taken as a starting point for further investigation rather than as evidence to draw conclusions from. Having said that, taking into consideration the findings from this paper and the paper from Purnell (2009), it would seem that general tendencies of vowel accommodation are hard to account for using traditional phonetic analyses. A different form of analysis is proposed by Bailly, Lelong, et al. (2010).

Bailly, Lelong, et al. (2010) present a study which demonstrates an accommodative mechanism that is somewhat more clear through the use of automatic detection and analysis techniques but using a highly restricted participant interaction paradigm. They developed a novel way of investigating accommodation which involved participants playing a game of *verbal dominoes* where interlocutors had to match the first syllable sound of their own word utterance to the final syllable sound of the word uttered by their interlocutor (as described in subsection 2.1.2). Participants were given a set word list so as to constrain their responses and to elicit the desired vowel pronunciations but they were balanced so that one could not guess what the next utterance that would be required. Participants needed to attend to the utterances of their interlocutor in order to complete the task. Prior to the interaction task all participants were recorded to provide baseline utterances of the words used in the verbal dominoes game prior to any engagement with another speaker.

Bailly, Lelong, et al. (2010) move away from more traditional forms of phonetic analysis of speech such as formant tracking of vowels in favour of a more holistic approach. They cite Delvaux and Soquet (2007) as grounds for the use of global automatic analysis of spectral distributions, or specifically in this case Mel Frequency Cepstral Coefficients (MFCCs, outlined in greater detail in subsection 3.6.1). MFCCs offer a broader characterisation of the speech signal than traditional measures such as formant values. However, there is a trade off in that they tend to lack the speci-

ficity that is gained through the use of more targeted measures like formant values. The authors used MFCCs to characterise the words used in the experiment at a phonemic level. Because the stimuli that were used in this experiment consist of either consonant-vowel (CV) and consonant-vowel-consonant (CVC), the vocalic portion of the words always constituted a vowel. It was the MFCCs of these vocalic portions that was used to determine the amount of accommodation between speakers. This was measured as the difference between the vocalic productions from the baseline utterances and the vocalic portions of the words uttered during the interactional task.

Whilst the authors did also perform other measures of accommodation (which all presented similar findings), it is only the results of the vowel data that are discussed here. Results demonstrated cases of strong convergence with some cases of modest divergence. The authors note that they do observe target specific behaviours where certain trends in accommodation are found for some vowels but not for others. The same is true for the behaviour of some speaker pairs where for some convergence on one vowel meant more maintenance was seen in other vowels. The authors point to individual differences in the way that speakers tend to fill their own vocalic space, especially between mid-vowels, although evidence for this is drawn only from French data and may not generalise to other languages (Ménard, Schwartz, & Aubin, 2008; Neagu, 1997).

Although this work helps demonstrate a link between interaction with another and accommodation, it is for a specific context. There was no overlap allowed during the experiment so that the automatic alignment of the phoneme level MFCCs wouldn't struggle and this is not representative of a real conversation. This paper was presented as the beginning of a series of experiments in which the context of the interaction will gradually become more similar to real world contexts. The experiment was designed to highlight those individuals which are the most likely to demonstrate accommodation and to use them going forward in the stream of experimentation. In addition to this, the introduction of regulations put in place to constrain the degree to which the participants can predict one another's upcoming interaction may be slightly at odds with the mechanisms which govern live accommodation. Taking a careful step-by-step approach, working from the bottom-up towards live conversation is a sensible and robust way to go about investigating accommodation and is much welcomed. However, currently it doesn't allow for a full assessment of accommodation in an interactional setting. Bailly, Lelong, et al. (2010) argue that their verbal dominoes paradigm encouraged participants to engage in 'active action-perception loops' (pp. 4). Considerations on the link between action and perception (or production and perception) are a key element in the interpretation of accommodation in a conversational setting (discussed in subsection 2.3.1) but it can also be helpful to look at how this interaction operates at a more imitative level, as presented here.

Where Bailly, Lelong, et al. (2010) are able to demonstrate the effects of accommodation through the use of more holistic measures such as MFCCs, Evans and Iverson (2007) are able to demonstrate it using phonetically selective measures. In their work Evans and Iverson (2007) recruited 27 Northern English (NE) speakers that were completing their secondary school education prior to attending university. They had the participants produce two repetitions of a list of target words and the phonetically balanced passage 'Arthur the Rat'. This was done at four time points, once before attending university (T1), once after having attended university for three months (T2), once upon completion of their first year at university (T3) and once upon completion of their second year at university (T4). Of the 27 participants initially recruited 23 completed the experiment up to T3 and 19 completed the experiment up to T4. Due to the fact that the participants developed accommodative behaviours in response to natural contact with others of a different linguistic community, this is classed here as being interactive. The accommodative behaviour is elicited from an interactive setting even though the data itself might not be collected in that setting.

For the acoustic analysis, the authors used the data collected from the word list and focused on vowel forms that they determined would be likely to vary, the vowels for *bud, cud, could* and *bath*, based on their previous work (Evans & Iverson, 2004). The measures that they took were changes in F1 and F2 over time and changes in vowel duration over time. For *bud* & *cud*, the authors demonstrated a centralization of the vowels away from their high-back forms as measured in T1. This change was detectable in both the F1 and F2 dimensions although the changes for *cud* occurred faster than for those in *bud*. Both vowels maintained their distinctions in their duration in relation to each other. For the *could* vowel a similar effect was seen. Where it had initially been produced with a high-back vowel, it became more centralized over time and this change was detectable in both F1 and F2 dimensions. As with *bud* and *cud* there was no change in vowel duration over time. Finally, for *bath* the same trend was found, participants began to pronounce the vowel with lower F1 and F2 values but the duration of the vowel remained the same. For all other vowels there was no effect of time. Much like other studies evaluating accommodation from an acoustic-phonetic point of view, the authors found consistent individual variations in the degree of change.

The authors interpret their results as evidence that speakers are able to adapt their speech repertoires at a relatively late stage, young adulthood, in order to better identify with new language communities. It should be noted that the authors point out that the changes made in the vowel productions do tend to be small. However, they are also significant. It is interesting that in this setting, the adaptation of vowels is able to be detected in a much clearer way than in Purnell (2009) and Pardo et al. (2010). One explanation for this is that there is added social imperative, more of a desire to belong for the participants that are involved in the Evans and

Iverson (2007) study since the accommodation is attributable to real-world exposure. This may well have provided enough impetus to drive the participants to make greater changes in their vowel pronunciations since the accommodative behaviour would have been driven by a social desire to engage with others and to maintain friendships. However, the other factor that is likely to play a role here is time.

The studies reviewed earlier in this subsection about vowels drew their data from comparatively short interactions between speakers whereas the Evans and Iverson (2007) study took place over a number of years. The longer time period may have allowed for the overall trends of accommodation to have become more prevalent. This type of effect is discussed in Sonderegger, Bane, and Graff (Accepted) where it is suggested that the short-term variability (ie. hours or even days) of phonetic forms may be somewhat mitigated by sampling over longer time frames (ie. months or years). If this is the case then perhaps more novel approaches to the evaluation of vowel change in accommodation, such as that used by Bailly, Lelong, et al. (2010), may prove to be a more fruitful approach.

#### Speech Rate

This particular part of the thesis deals specifically with studies of accommodation that make use of interactive stimuli to elicit accommodation, unlike subsections 2.2.2 and 2.2.3 which dealt with studies using non-interactive stimuli. In general, this means that the stimuli used in the studies tend to be somewhat less controlled than those using non-interactive stimuli. However, the following study by Casasanto et al. (2010) diverges from this trend somewhat. The authors make use of virtual reality to control the stimuli that participants are exposed to. This allows for every aspect of the environment in which the participant is engaged to be controlled by the experimenter. Whilst overt control of the stimuli that a participant is exposed to might be more akin to a non-interactive stimuli source, it is the fact that in virtual reality the environment responds to the participant's behaviours in real-time that allows it to be classed as an interactive source. Casasanto et al. (2010) utilized this functionality in order to investigate the degree to which convergence could be found to be due to variations in speech rate. In addition to this, the responses that participants were allowed to make were free of overt constraints, such that their responses can be classified as spontaneous. That is to say that the participants' responses were not scripted even if the speech content of the virtual interlocutor was.

In their experiment, Casasanto et al. (2010) invited 62 Dutch participants (30 male, 32 female) to take part in an experiment evaluating the impact of speech rate on accommodative behaviours. To do this, the authors constructed a virtual environment (VE) in which the participants would be immersed. This VE consisted of a single long aisle in a supermarket and a virtual interlocutor called VIRTUO. The supermarket aisle was stocked with products that one might generally expect to find.

The experiment consisted of a baseline recording block consisting of four trials and a test block. The participants entered the VE by putting on a virtual reality headset fitted with a microphone to capture their speech. Participants explored the VE by moving their heads to look around, participants did not need to physically move to explore the VE and were seated for the whole experiment. The baseline block consisted of four trials in which the participants were asked to look at a specific product on the shelves of the supermarket aisle and provide a description of it. Participants were alone in the VE for the baseline block. Following the baseline block, participants were introduced to VIRTUO and the test block began. During the test block VIRTUO escorted the participants along the shopping aisle, stopping at specific products and asking the participants questions about that product. VIRTUO's speech was a pre-recorded script that was produced by a male native Dutch speaker. VIRTUO had no ability to understand or interpret the speech of the participants and has a limited vocabulary. The responses that VIRTUO offered were the result of the experimenter pressing buttons to offer a pre-scripted response. A random delay of 150 ms - 400 ms accompanied the button press so that the participants' speech rate was not influenced by the turn taking behaviour of the experimenter. In order to test the influence of speech rate on accommodation, VIRTUO's speech rate was adjusted to two different speeds, fast and slow. This provided two conditions for which the participants would be split up into, 33 were randomly placed in the fast condition and 29 in the slow condition. In the fast condition, VIRTUO's speech was sped up by 12% whereas in the slow condition VIRTUO's speech was slowed down by 12%. The speech rate of the participants was measured by defining intervals based on the participant's utterances and then dividing the number of words in the transcript of the utterance by the number of seconds in the interval. To determine the change in participants' speech rates, if any, speech rate during the interaction with VIRTUO was compared against the participant's baseline recording.

It was found that those participants in the fast condition demonstrated a significantly higher speech rate during their interactions with VIRTUO whilst there was no significant change found in the speech rate of the participants in the slow condition. In addition, it was discovered that participants in the fast condition demonstrated convergence in their speech rate early on in the experiment, after only the first item interaction with VIRTUO. There was however, a non-significant trend of increased speech rate in the slow condition, the authors attribute this to the possible influence of additional factors beyond VIRTUO's speech rate. The most prominent of these external factors being 'immersion' in the VE or the degree to which a person feels as though they are actually present in the VE.

The key finding here was that VIRTUO's speech rate affected the speech rate of participants at all. This is because participants are accommodating to a virtual being with no way of interpreting the social factors and intentions carried across in accommodation as suggested by CAT. This may be supportive of an automatic theories

and views of accommodation. The authors suggest that the effects observed in their data may be the result of 'overlearning' where some aspects of human interactions become so automatic that they are overextended to contexts where the social purpose of the action no longer applies. However, the authors also found that the level of accommodation found in the participants was correlated with the degree to which the participants identified with VIRTUO, as measured by a post-task questionnaire. This suggests that even though participants knew that VIRTUO was virtual, they still attributed social traits to him. It may have been the case that participants were accommodating to their own perceived social traits of VIRTUO. Having said this, the speech that was used as VIRTUO's voice was pre-recorded. This pre-recorded voice may have had some subtle but perceptible markers that contributed to the accommodative effect but if this were the case then there would likely have also been some impact on the speech rate of the participants in the slow condition.

The finding that the speech rate of the participants in the fast condition changed rapidly upon engagement with VIRTUO is also interesting as traditional accounts of accommodation would suggest that greater accommodation should be found as time spent with the interlocutor increases. The data here do not show this, in fact they show quick adaptation followed by sustained maintenance. These results would seem to suggest that accommodation is somewhat more automatic than might have previously been assumed. Especially given that accommodation took place in relation to a virtual interlocutor, with no social imperatives, at an early stage in the interaction and was linked to degree of perceived affiliation. In addition, it is worth noting that in this experiment, the speech samples were taken from the interaction with VIRTUO rather than from a post-task. When this is coupled with the fact that the authors found convergence at an early stage in the interaction, it raises the question of whether interpreting accommodation across the course of an interaction, in relation to the conversational goals of the participants might demonstrate some of the short-term variation proposed by Sonderegger et al. (Accepted). The interpretation of acoustic features over time during a conversational interaction is not common practice in experiments considering accommodation. In this paper, the authors do make some attempt to interpret change over time, by commenting on the nature of the rapid accommodative effects found in their data. However, they offer their interpretations with a warning that they did not counterbalance the order of VIRTUO's questions and that the content of the questions might have influenced the degree of accommodation over time.

#### Fundamental Frequency (F0)

One study that explicitly set out to evaluate the degree of adaptation in an acoustic feature of speakers' speech over time was that of Collins (1998). This study assessed the degree of convergence in mean F0 between female participants in conversational pairs who were engaged in a free-flowing conversation. The use of a free-flowing,

unconstrained conversation is unusual in research on accommodation. Mainly this is because it is hard to guarantee repetition of specific acoustic-phonetic features for use in statistical testing. However, the use of free-flowing conversation in this study offers both a more ecological context for the investigation of accommodation and because the author is considering the supra-segmental feature of F0, she does not have to worry about obtaining sufficient repetitions.

The experiment itself consisted of 4 pairs of female participants (8 participants in total), each of which was paired based on demographic information. This was done so as to avoid unequal social relations and to ensure that participants had enough in common to maintain a fifteen minute conversation. Participants were first asked to read two lines of a poem so as to achieve a baseline measure. They were then sat in their pairs and asked to hold a conversation for fifteen minutes. The mean F0 for each speaker was extracted at nine times across the course of each fifteen minute conversation. Each sample consisted of roughly $1 - 1.5\,\mathrm{s}$ of uninterrupted speech extracted at approximately $1\,\mathrm{min}$ intervals.

Findings demonstrated that mean F0 does converge between participants and that the mean F0 will co-vary between conversational partners during the course of a conversation. However, the author notes that convergence was only found at some points during the interaction and that the convergence was not linearly consistent. Collins (1998) did not record the linguistic content of her data however and can therefore not distinguish any convergence found in mean F0 across the interactions from that which may be attributable to the pitch patterning linked to speech acts such as questions, demands or pleas. She notes that without this information to guide the analysis it cannot be concluded that the convergence identified is actually being used as an interactional resource by the participants rather than it just being a statistical coincidence.

Having said this, it may also be the case that pitch is being used by participants as both a linguistic cue and an accommodative cue. As participants interpret the information contained in F0 over time, there may be instances where the information pertains to linguistic meaning and other instances where F0 is linked to a form of social representation. The temporal aspect of the acoustic signal here was sampled as a categorical variable based on the mean F0 sampled over $1 - 1.5\,\mathrm{s}$ with roughly $1\,\mathrm{min}$ between each sample for each participant individually. This method would not have the temporal resolution or the statistical power to discern elements of the acoustic signal between speakers which co-vary or are mutually dependent on one another at a time-scale which would necessarily be linked to phonetic features of the interaction rather than any linguistic features that may be present at a less fine grained temporal resolution. What is meant by this is that there may be temporally fine grained elements of the F0 signal in both speakers which contribute to joint social representation or accommodation but which do not necessarily carry any linguistic content until a given amount of time has passed. The information in

the signal is cumulative and time dependent, there may be subtle adaptations that are masked by the more salient and overarching manipulations that are found in F0 due to linguistic factors such as the intonation patterning involved in questions, demands and pleas. Disentangling how different aspects of the F0 signal might influence joint social representation and accommodation versus how it influences linguistic constructs is a difficult problem. Especially considering that a third possibility is that both of these aspects may very well interact with each other in a number of ways to contribute to the overall effect of each other.

### 2.2.5 Summary

The main aim for this subsection is to summarise the information presented in this section (section 2.2), providing an outline of the areas that have been investigated and the problems that are still to be addressed.

This section has presented an evaluation of the methods that have been employed to investigate accommodation. In order to consider the work presented in this field in a systematic and organised way, the field was separated into four quadrants (as represented in Table 2.1). The quadrants were defined by two characteristics: (1) the way in which they measured accommodation, perceptual or acoustic-phonetic, and (2) they way that they elicited accommodation, interactive or non-interactive. This allowed studies to be categorised as one of four methodological approaches 'Perceptual Interaction', 'Perceptual Non-Interaction', 'Acoustic-Phonetic Non-Interaction' or 'Acoustic-Phonetic Interaction'. This subsection offers a summary of the material discussed over the course of this section.

A number of the merits and drawbacks of each approach overlap between different approaches. In light of this and for ease of interpretation, a general summary of the overlapping merits and drawbacks is presented in table 2.2. Each of these overlapping merits and drawbacks will be briefly discussed before moving on to provide a summary of the key points to take away from this section.

The main merits and drawbacks of the approaches considered in this section, as summarised in table 2.2, can be thought of as either providing or not providing the following in relation to accommodation:

- a measure that accounts for all perceptually relevant aspects of speech
- a measure that can identify key speech features
- a stimulus that has comparatively high ecological validity
- a stimulus with comparatively high experimental control
- a stimulus that allows for assessment of both speaker perspectives

Each of these listed features will now be briefly discussed in relation to the stimulus and measure groups in which they can be found in table 2.2

| | **Interactional** | **Non-Interactional** |
|---|---|---|
| **Perceptual** | + Accounts for all perceptually relevant aspects of speech<br>+ Comparatively high ecological validity<br>+ Can assess both speaker perspectives<br>− Cannot identify key speech features<br>− Comparatively low experimental control<br><br>(eg. Pardo, 2006; Alshangiti & Evans, 2011; Kim, Horton, & Bradlow, 2011; Pardo, Gibbons, Suppes, & Krauss, 2012) | + Accounts for all perceptually relevant aspects of speech<br>+ Comparatively high experimental control<br>− Cannot assess both speaker perspectives<br>− Cannot identify key speech features<br>− Comparatively low ecological validity<br><br>(eg. Namy, Nygaard, & Sauerteig, 2002; Shockley, Sabadini, & Fowler, 2004) |
| **Acoustic-Phonetic** | + Can identify key speech features<br>+ Comparatively high ecological validity<br>+ Can assess both speaker perspectives<br>− Cannot account for all perceptually relevant aspects of speech<br>− Comparatively low experimental control<br><br>(eg. Collins, 1998; Evans & Iverson, 2007; Purnell, 2009; Bailly, Lelong, et al., 2010; Casasanto, Jasmin, & Casasanto, 2010; Pardo, Jay, & Krauss, 2010; Manson, Bryant, Gervais, & Kline, 2013; Bailly & Martin, 2014) | + Can identify key speech features<br>+ Comparatively high experimental control<br>− Cannot assess both speaker perspectives<br>− Cannot account for all perceptually relevant aspects of speech<br>− Comparatively low ecological validity<br><br>(eg. Shockley, Sabadini, & Fowler, 2004; Nielsen, 2011; Yu et al., 2013; Babel, 2010, 2012; Babel & Bulatov, 2012) |

Table 2.2: Summary of the work discussed in section 2.2. Each quadrant represents one of the four conceptual approaches presented in the section. Each of these quadrants contains a brief summary of the merits and drawbacks of each of the approaches, along with the studies discussed in the section relating to that particular approach.

Those studies that use perceptual measures can be said to contain a measure that accounts for all perceptually relevant aspects of accommodation in speech. By making the choice to discern if accommodation had taken place by asking participants to rate the levels of similarity between two samples, experimenters are able to make full use of the natural ability of the human perceptual system to evaluate speech as a whole. The judgement as to whether accommodation has taken place is made based on the full range of speech features that are available to the person making the judgement. Doing so removes the issue that acoustic-phonetic approaches have in understanding if accommodation might be the result of multiple interacting phonetic features. However, the issue inherent in using perceptual measures is that there will be variability from person to person based on a variety of factors relating to the social and linguistic exposure of the person making the judgement of accommodation. Although this sort of variability is often controlled for in experiments by using large numbers of participants and averaging across them, the numbers of participants used in studies of accommodation tend to be rather small (see: Pardo et al., 2016a).

Whilst perceptual measures do offer a more complete view of accommodation than acoustic-phonetic measures, they cannot directly inform the experimenter about the speech features that participants are using to accommodate. This is a consequence of the underlying nature of the tool used to assess accommodation (ie. the human perceptual system). In order to determine the key speech features that are being used to accommodate, an acoustic-phonetic approach is required. Due to the ability of acoustic-phonetic approaches to isolate and evaluate specific aspects of the speech signal, it is possible to design experiments that can focus on determining if any given phonetic feature plays a significant role in accommodation. However, analyses of this kind are more time consuming to produce than their perceptual counterparts. Identifying the relevant parts of the speech signal can be a time consuming and tedious task, even with the aid of automated and semi-automated extraction tools (eg. Fromont & Hay, 2008; Rosenfelder et al., 2011; Sonderegger & Keshet, 2012). So performing analyses that encompass the whole of the phonetic repertoire of a speaker is somewhat infeasible. As a result, whilst acoustic-phonetic approaches do allow for identification of key phonetic features in accommodation, they do not allow for access to the information regarding the exact mix of phonetic features that are being used in accommodation (Pardo, 2013).

Studies taking an interactional approach allow for accommodation to be assessed from a more ecologically valid point of view in comparison to non-interactive designs. The process of accommodation is reliant on having a stimulus against which to adapt one's own vocalisations and the most common of these stimulus sources is other speakers. By allowing exposure to other speakers in an interactive format, these studies track somewhat closer to the likely sources of accommodation that participants come into contact with outside of the laboratory. Naturally, these stud-

ies are still completed in a laboratory setting so they will never fully replicate the accommodation that might be found in everyday speech but they do provide a closer approximation than studies using isolated and tightly controlled stimuli.

Those studies that take a non-interactive approach to eliciting accommodation can generally be said to have greater experimental control over their stimuli. Studies based on stimuli produced from speech samples can present carefully controlled examples of specific speech features for the participants to accommodate to. This allows for precise questions to be asked and for targeted hypotheses to be tested. Whilst there are means for eliciting higher incidences of specific speech features in an interactive setting (eg. DiapixUK - Baker & Hazan, 2011), they still cannot guarantee elicitation or control the specific form of the target speech feature. Interactional settings are inherently variable and will often produce speech features that adapt to the utterances of the speech partner.

The final main element of the presented works on accommodation is the ability to assess accommodation from the point of view of both speakers. In general, this is something that cannot be captured by non-interactive approaches to assessing accommodation. This is simply because of the lack of a direct speech partner in the design of the experiment. Whilst non-interactive approaches may have a baseline measurement with which to compare the speech of a post-exposure sample from another speaker, they cannot assess how a speaker varied in response to the ongoing speech of a speech partner. Given that accommodation is modulated by social factors (section 2.1.2) and that these factors can be represented in the speech signal, it would be reasonable to assume that over the course of an interaction speakers would converge and diverge in relation to the social information embedded in the speech of their interlocutor. That is to say that the relationship between accommodation and time is non-linear and varies dependent on contextual events and any given time-point. This is akin to the day-to-day variance suggested by Sonderegger et al. (Accepted) except over the course of a conversational interaction rather than a period of weeks or months. In order to capture this variance, interactional experiments would need to be conducted.

This section has focused entirely on the measurement techniques used to detect accommodation in the speech signal. Taking this section and section 2.1 together, it is clear that accommodation relies on the interaction between two speakers. It is also clear that accommodation is a complex phenomenon which is unlikely to follow a simple linear relationship with incoming speech stimuli. In order to fully understand accommodation, it would be of use to consider the cognitive mechanisms that might underpin the phenomenon. This would provide insight into how speakers are able to accommodate with such ease, given the apparent complexity of the phenomenon. The next section, section 2.3, explores the possible cognitive mechanisms that could underpin accommodation. Further to this, details are provided about how accommodation and brain activity might be linked. Following

this, section 2.4 discusses how machine learning approaches could help to provide a shared approach for measuring both speech accommodation and brain activity.

## 2.3 Why should accommodation be linked to joint brain activity?

The previous two sections, section 2.1 and 2.2, focused exclusively on the interpretation of accommodation from the perspective of speech. They presented the linguistic theories that underpin accommodation and the ways in which it is measured. Although the above sections did make use of the social-psychological and psycholinguistic concepts that underpinned theories such as CAT, no attempt was made to link that work to neural processing. This section aims to address that by providing the necessary background to link the literature on accommodation to that of brain activity.

As discussed in section 2.2, accommodation has been shown to be detectable in both situations with a social motive to accommodate and in situations where that social motive has been minimised. Detecting accommodation in both of these settings suggests either that the social imperative to accommodate is strong enough to persist into situations where most social motives have been stripped away or that accommodation plays somewhat of a more fundamental and automatic role in communication.

In general, the studies reported in section 2.1 and 2.2 do not set out to evaluate accommodation as a phenomenon in its own right. There is often a use of the phenomenon to provide insight into a broader topics aiming to elucidate the possible links between perception and production (eg. Shockley et al., 2004; Pardo, 2006; Bailly, Lelong, et al., 2010; Alshangiti & Evans, 2011). Most of these make use of or attempt to provide support for a particular theory or set of theories that aim to describe the perception and production of speech. In this section the links between accommodation and the most prominent theories that tend to be used to explain it will be considered.

Subsection 2.3.1 outlines the core theories that will be considered and provides the reasoning behind why they relate to accommodation. This is done to provide an exploration of some of the cognitive mechanisms that have been proposed to drive speech production. However, the subsection does not aim to cover all theories of speech production, rather just those that suggest or predict accommodation. This subsection can be thought of as providing the information regarding the internal mechanisms required for an individual to produce accommodative acts.

The following subsection, subsection 2.3.2 explains the concept of neural entrainment and how it relates to the processing of the speech signal. There is a short outline of how neural oscillations in the brain are generated and then evidence is presented for the role of neural oscillations in the processing of information in the brain. This is then further extended to the role that oscillatory activity might play in the processing of speech in the brain by way of coupling to critical events in the speech signal.

The penultimate subsection, subsection 2.3.3 explores research surrounding oscillatory activity in the brains of pairs of interacting participants and how this might link accommodation and neural activity. First, studies concerning joint interaction and oscillatory activity are presented and the method of *hyperscanning* is presented. Following this, a selection of studies that have looked at joint neural activity in relation to speech are presented. Finally, a proposition is made as for how this could relate to accommodation.

Finally, subsection 2.3.4 provides an overview of what was discussed in this section, draws together findings and relates them to accommodation.

It should be noted that this section does not aim to outline all of the theories and proposals put forward that account for the relationship between the production and perception of speech. It simply aims to present and evaluate the theories that pertain to the interpretation of accommodation.

### 2.3.1 Accommodation - looking under the hood

Pardo et al. (2016a) notes that whilst the literature on accommodation provides a great deal of insight into the social and cultural settings surrounding accommodation, it is 'mute regarding cognitive mechanisms that support convergence and divergence during speech production' (pp. 2). In order to fill this gap in knowledge, additional literature needs to be drawn upon. It is the aim of this subsection to provide an overview of work that presents cognitive mechanisms that support or predict accommodation in speech production.

According to Pardo et al. (2016a) the key theories that present cognitive mechanisms that suggest or predict accommodation are the motor theory of speech (Liberman & Mattingly, 1985), the direct-realist approach (Fowler, 1986), the storage of phonetic detail in episodic memory (Goldinger, 1998) and the mechanistic approach of language use in dialogue (Pickering & Garrod, 2004; Pickering & Garrod, 2013). Here, each of these theories will be outlined and an explanation of how and why they relate to or impact upon theories of accommodation will be provided.

#### The motor theory of speech

This theory was proposed mainly to account for speech perception rather than speech production but one of its key claims is for an intimate link between speech perception and speech production. As a result of this intimate link, it presents some pertinent considerations for accommodation. Here it is considered in its revised form (Liberman & Mattingly, 1985).

The motor theory of speech has two key claims. The first of these that the objects of speech perception are not the speech signals themselves but rather the intended phonetic gestures of the speaker. That is to say that it is the movement of the articulators themselves (eg. the tongue, lips, jaw, pharynx etc.) that is perceived when

interpreting speech. The theory posits that these intended phonetic gestures are represented in the brain as invariant motor commands. It is argued that to perceive an utterance is to 'perceive a specific pattern of intended [phonetic] gestures' (pp. 3). Referring to the phonetic gestures as 'intended phonetic gestures' is deliberate since there are a number of reasons (such as co-articulation) that specific gestures may not be directly relatable to the content of the speech signal.

The second key claim of the theory is based around the invariant motor commands that underpin the first claim. It suggests that if speech perception requires the listener to decode speech gestures using the same set of motor commands that they use to produce speech, then both speech production and speech perception must be intimately linked. The theory argues that the link between speech perception and speech production isnon-trivial, arguing that 'perception and production are only different sides of the same coin' (pp. 30). The reason that this theory takes such a strong stance on the link between perception and production is because it considers speech stimuli and speech responses to resemble each other, which is something that other modular response systems do not do. It is argued that the perception and productions targets are the neural representations of the motor commands and are thus, the same. Further to this, it also remarks that both perception and production are regulated by the same structural and grammatical constraints. Thus, if the systems were regarded to be separate, an explanation for how the same structures evolved for both perception and production would need to be provided. It is more likely that the systems are linked rather than having developed separately.

The example that Liberman and Mattingly (1985) provide is that of a simple reflex system. They draw on the work of Lee and Reddish (1981) that considered the mechanisms involved in the timing of body movements for diving gannets. The point is made that the system for automatically converting the optical stimuli regarding the location of the water's surface into a motor response to adjust the gannet's posture does not demonstrate parity between stimulus and response. The optical stimulus is very different from the response provided by the motor system. It is this lack of parity that suggests different components for the module that governs a gannet's automatic diving reflex. One component must interpret the optical signal before passing that information to a motor response component. For the motor theory of speech, language retains a parity between stimulus and response.

In the motor theory of speech, the listener recruits the motor system in order to perceive the neural representation of the stored invariant motor commands. This is performed through an analysis by synthesis which retrieves a speaker's intended gestures from their acoustic output (Galantucci, Fowler, & Turvey, 2006). This tightly coupled relationship between speech perception and speech production provides indirect support for accommodation.

If the perceptual system is not only recruiting the motor system but is inherently coupled with it, then when a social drive to accommodate is presented, accommo-

dation could be an automatic outcome. If the listener is perceiving the speech of a speaker in terms of their intended string of co-articulated motor commands with the listener's own motor commands being used for analysis by synthesis, this could lead to carry over into the listener's speech production.

### The direct-realist approach

Where the motor theory of speech relied on the motor system to account for speech perception and by extension speech production, the direct realist approach suggests more of a shared mechanism. The work of Fowler (1986) suggests that effective communication requires the perceiver to evaluate an information source in order to recover the 'distal event' that encoded that source. This element is similar to the motor theory of speech. However, where the motor theory of speech abstracted towards the neural representations of motor commands, the direct-realist approach considers the 'distal event' to be the articulatory actions of the vocal tract. This theory argues for 'direct perception of the environmental source of its [the acoustic signal] structure'(pp. 6). That is to say that the vocal tract articulations are what is being perceived by a listener when the acoustic speech signal is being interpreted.

It is proposed that the ability to imitate speech is easier to understand under a theory where vocal tract gestures are perceived rather than from a theory in which the speech signal is mapped onto phonological categories (Sancier & Fowler, 1997). The reasoning behind this claim uses the tendency of infants to imitate facial expressions to exemplify its claim. It is noted that infants can imitate facial expressions without seeing their own face. In order to achieve this, infants must be able to interpret the facial gesture of another and convert that into a facial gesture of their own. To be able to do so at such a young age suggests that this ability is automatic. In the case of communicatively important bio-signals, this is interpreted as receiving instructions for an imitative response. Communicatively important information received by the perceptual system is considered to serve as a goad for imitation. It has even been explicitly stated that 'perceiving speech is, effectively, receiving instructions for its imitation' (Sancier & Fowler, 1997, pp. 431).

Again, this theory assumes a tight link between speech perception and speech production. It is considered that the direct perception of linguistically relevant vocal tract gestures provides an inherent impetus to imitate. This is assumed to work through a similar mechanism that allows for the imitation of facial mechanisms, where vocal tract gestures are imitated rather than facial expressions. This theory, much like the motor theory of speech, was developed from a functionalist, biological perspective of language and as such, might have too much of a narrow focus on phonetic structure (Studdert-Kennedy, 1986). Like the motor theory of speech, it does not explicitly address the relationship between the speaker and the listener nor the need for communicative acts to contain an intent or goal. For an articulation to be considered as a speech act, there should be an intent behind the utterance as well

as an acoustic signal that is structured by the speaker's articulations. In addition, it requires an intended audience (even if that audience is oneself). The above two theories go some way to explaining the cognitive mechanisms of accommodation but lack a robust explanation of the interactive nature of speech.

### Phonetic detail and episodic memory

The previous two theories that have been discussed focus on the acts of perceiving and producing speech. They provide an account of speech communication that is driven by stimulus and response, they do not explicitly consider the impact of prior information or memory on the perception/production system. In so far as memory is concerned, the motor theory of speech and the direct-realist approach may be said to contain representations of the speech motor commands and articulatory gestures fo the vocal tract, respectively. They do not make suggestions about the relationship between speaker and listener, preferring to focus on the individual rather than interacting speakers. They also make the assumption that the mechanisms used in speech perception and production are innate rather than learned. The role of speaker specific adaptations and factors such as word frequency must then be considered to be encoded in the motor signals during speech. Another mechanism, involving episodic memory (where episodic memory is the result of experiences as opposed to semantic memory which is the memorisation of specific facts) is highlighted by Goldinger (1998).

In the episodic memory model presented by Goldinger (1998) the phonetic detail required to accommodate is learned rather than being present in the speaker by default. It is argued that each exposure to a word that a person hears leaves a trace in their episodic memory. As such, words that are heard more frequently will build up a greater number of traces. Encountering a new realisation of a word will activate all traces linked to that word and they are averaged to produce an 'echo' which allows for word recognition. When an echo of a high frequency word is produced, it contains fewer of the specific attributes of that particular realisation since it is the result of an average across the traces that are contained in episodic memory. This leads to higher frequency words being less likely to provide the necessary phonetic detail required in order to accommodate. Lower frequency words on the other hand, have fewer traces in episodic memory so the echoes produced of them will contain more of the detail that is specific to that particular realisation. Thus, lower frequency words would be more likely to be converged towards, under this episodic memory model. Allowing for the mechanisms of speech perception and production to be updated in response to previous exposure also means that adaptation in response to specific individuals can also be accounted for. Traces specific to a particular speaker or social group can be more strongly activated during perception, leading to a stronger weighting towards some phonetic features during echo production. This in turn leads to audience specific adaptation in speech production.

The work presented by Goldinger (1998) was produced in order to test predictions from an exemplar based model (Hintzman, 1986). It builds upon previous work on exemplar based theories of speech perception and production (eg. Klatt, 1979). It therefore assumes that stimulus variability is stored in memory but suggests that this need not necessitate data reduction. Unlike some theories that suggest speaker specific details such as voice should be considered as noise during phonetic perception in order to normalise for ease of lexical recognition (eg. Pisoni, 1993), this episodic memory model posits that normalisation and speaker specific voice memory can co-exist. The notion of traces allows for both lexical items and speaker specific features to be stored as traces that can be concurrently activated during perception, going on to influence production. When speaker specific access is linked to specific memory traces, general lexical and phonetic form is accessed through more global averages. This model is not necessarily an alternative to the motor theory of speech and the direct-realist approach but perhaps more of a complimentary mechanism. Importantly though, it allows for the consideration of audience specific accommodation, where speech production is influenced by stored representations of speaker and speech community norms.

#### Mechanistic language use in dialogue

One theory of speech perception and production which explicitly predicts phonetic accommodation is that proposed by Pickering and Garrod (2004). In their mechanistic account of language use in dialogue, they propose a theory centred around speech as an interactive process between speakers. They approach speech perception and production as an act involving two or more speakers from the outset.

The basic proposal is centred around a simple priming mechanism linked to a shared representation of the topic at hand. It is argued that during a speech interaction with another speaker, an alignment in the mental representations that each speaker holds of the topic at hand is required for successful communication. This alignment of mental representations takes place at multiple linguistic levels (eg. semantic, syntactic and phonological). Further to this, alignment at one level promotes alignment at other levels. The mechanism that underpins this ability to align mental representations is described in a more recent paper (Pickering & Garrod, 2013) and involves covert imitation of both one's own and one's partners speech utterances. In order to situate this mechanism within language use, concepts from computational neuroscience literature concerning action perception (Davidson & Wolpert, 2005; Wolpert, 1997) were translated from these fields into a linguistic framework. Essentially, it is proposed that along with the command to produce the intended speech utterance, the brain also produces what is referred to as an efference copy of that command. This efference copy is a prediction of what should be produced by the command. As one comprehends what one is saying, a monitor compares the produced utterance to the prediction made by the efference copy. If

there is no mismatch or if the mismatch is within a given tolerance, no adjustments are needed. However, if a difference is found, adjustments can be made to bring the utterance into line with what is predicted.

Within this theory, alignment of situation models is required for successful communication. As mentioned above, the process of aligning situation models involves alignment at multiple linguistic levels since alignment at one level promotes alignment at other levels. Maintenance of this alignment is not thought to be an overt mechanism but is rather, a mechanism that is invoked through priming. The predictions made by the efference copy are produced through a priming driven by the incoming speech of the interlocutor. The constant monitoring of one's own speech whilst automatically integrating information from an interlocutor helps to maintain alignment. As such, the efference copy can be considered to contain predictions that help to maintain alignment across these multiple levels through constant monitoring of perceived output.

Again, this theory maintains that there is a close link between speech perception and production. The efference copy that is produced is dependent on both the information that has been stored through the past perception of speech and the incoming speech signal of the interlocutor yet it acts as a control mechanism for speech production. The fact that this mechanism is driven by a priming that originates from the incoming speech signal of the interlocutor actually predicts the presence of accommodation. The incoming speech signal is used to continually update utterances in relation to the joint goal of situational alignment.

Within the four theoretical frameworks that are presented above, one key theme that permeates through all of them is the notion that speech perception and production are linked. Whether the mechanism that underpins the theory is reliant on matching the neural components of motor commands (Liberman & Mattingly, 1985), interpreting articulatory gestures of the vocal tract (Fowler, 1986), activating the echoes of stored speech forms (Goldinger, 1998) or aligning situation models (Pickering & Garrod, 2004; Pickering & Garrod, 2013), each theory relies on speech perception and production being linked. It is this inherent link between perception and production that allows for accommodation to be considered as a plausible component of speech communication.

Accommodation, as defined in section 2.1, can be broadly thought of as the tendency of a speaker to adjust their production of speech sounds in relation to their speech partner. In order to accomplish this, a speaker must necessarily recruit both their perception and production systems. Indeed, the process of speech communication in general requires co-operation between these systems. However, the theories presented in this subsection begin to suggest that the perception and production systems do more than co-operate. They suggest that the systems at the least, share their mechanisms or, perhaps more arguably, suggest that the mechanism could be similar for both systems. It is this move towards considering perception and production

as a shared mechanism that provides the most promising underpinnings for accommodation in terms of a cognitive mechanism. Irrespective of the exact workings, a shared cognitive mechanism between speech perception and production would allow for rapid and relatively automatic adaptation of speech sounds in relation to a speech partner.

### 2.3.2 Neural entrainment

In the previous subsection, the theories that were likely to underpin accommodation were outlined. It was noted that a key element present in all of the theories was that the perception and production of speech was thought to be linked. The proposed links between perception and production were mainly presented as being contained within the individual. It is the purpose of this subsection to explore how the brain might interpret environmental stimuli to maximise efficiency of speech processing. The concept of *neural entrainment* is presented as a likely candidate in maximising speech processing efficiency which could play a role in linking accommodation to brain activity.

Neural entrainment refers to the coupling of neural activity to environmental stimuli. The concept underpinning this notion is that in order to interpret and track dynamically evolving environmental stimuli, the neuronal mechanisms within the brain must come into line with temporally expected critical events (Henry & Obleser, 2012). That is to say that, in order to process incoming information, the brain must continually adapt relative to the form of that information so that it is able to sample it in an efficient and appropriate manner. Whilst the entrainment between neural activity and the activity of other humans applies to any interpretable human response (eg. gesture imitation and gaze following: Dumas et al., 2011), speech is of particular interest.

Before discussing the work detailing the links between neural entrainment and speech, it is worth providing a brief (and basic) outline of how the neural entrainment process is assumed to work. As the name suggests, neural entrainment is based around the way in which neurons operate. Neurons are cells that are 'specialized for the reception, conduction, and transmission of signals' (Pannese, 2015, Ch.2, pp.9). They are contained within large, interconnected networks of neurons which all work together. Any individual neuron will '(a) react to various physical and chemical stimuli giving rise to signals (excitability), (b) convey these signals at high speed (conductivity), and (c) transmit them to other neurons or to nonneuronal cells (e.g., muscle or gland cells) thereby influencing their activity'(Pannese, 2015, Ch.2, pp.11). Neurons react to stimuli by producing an electrical discharge which carries a signal to a target location. These electrical signals are what is detected by brain activity measurement techniques such as EEG (or the magnetic component of the electrical signals in the case of MEG). Neurons are continually active but can show

increased activity in response to particular events (Perrault, Vaughan, Stein, & Wallace, 2003; Fox & Raichle, 2007). Following activation, neurons have a recovery period before an activation can again be produced. It is this process of electrical discharge and recovery that gives rise to what have been come to be known as brainwaves or neuronal oscillations. It is the ebb and flow of electrical activity in the brain. Experiments linking this electrical activity in the brain with environmental stimuli are at the heart of functional neuroimaging studies (studies aiming to establish the functional role of neural activity). Although techniques for capturing and interpreting this information vary (see: Savoy, 2001; Bandettini, 2009), the core aim of all studies is to determine how the brain deals with the stimuli with which it is presented.

As stated above, neural entrainment refers to the coupling of neural activity to environmental stimuli. The stimuli to which the neural activity is coupled can come from many sources including musical beats (Nozaradan, Zerouali, Peretz, & Mouraux, 2015), artificially induced electromagnetic forces (Thut et al., 2011), visual stimuli (Spaak, de Lange, & Jensen, 2014) and speech (Giraud & Poeppel, 2012). Importantly, in order for entrainment to occur, there must be some temporal predictability in the stimulus (Lakatos, Karmos, Mehta, Ulbert, & Schroeder, 2008). The reasoning for this relates to the excitability of local neuronal populations. As Peelle and Davis (2012) explains, neuronal oscillations characterise the shifting excitability of local neuronal populations. At some points in time, these populations are highly excited whilst at other times, they are less excited. This gain and loss of excitation is what produces neuronal oscillations. Given that these neuronal populations will have periods where they are approaching high excitation, it is thought that information arriving at this time will be processed more efficiently than information arriving at periods of low excitation, as illustrated in figure 2.3.

What this suggests is that temporally structured stimuli are most efficiently processed when the neural oscillatory activity in the receiving region of the brain is aligned such that the relevant information arrives at a time of high excitability. As such, neuronal oscillatory activity that is phase-locked with the temporal structure of the stimuli can be considered to be being efficiently processed. Naturally, the stimuli that the brain receives are highly complex and as a result, phase-locking is constantly being realigned based on the expected temporal occurrence of the stimuli. By doing this, what is effectively happening is that the oscillatory activity forms predictions of the temporal location of upcoming critical stimuli (Engel, Fries, & Singer, 2001). The form that the critical stimuli driving neural entrainment take however, is still under investigation, although suggestions have been made regarding the stimuli linked to neural entrainment to speech.

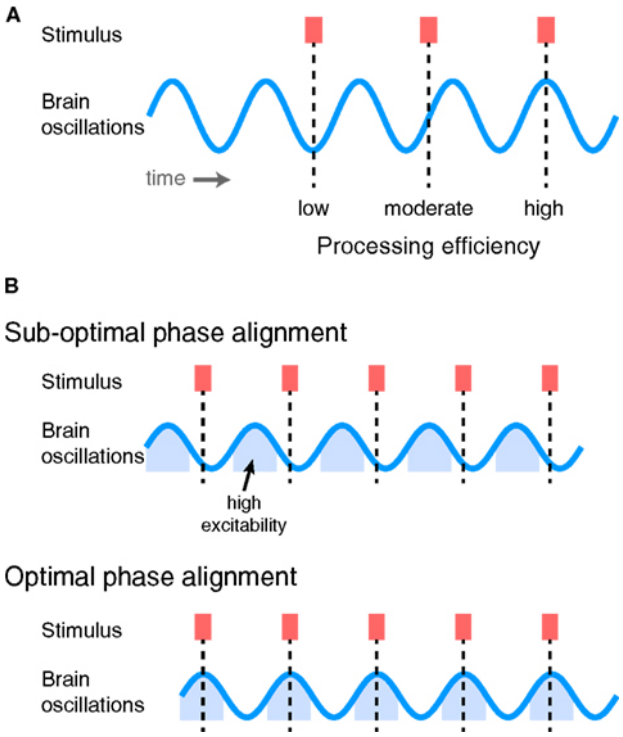



Figure 2.3: In this figure taken from Peelle and Davis (2012), panel A represents the efficiency of processing for sensory stimuli given the current phase of ongoing oscillatory activity. Stimuli that arrive during a low-excitability phase are processed with less efficiency than stimuli arriving during a high-excitability phase. Panel B demonstrates how the processing of stimuli with temporal regularity can be improved by shifting the phase of ongoing neural oscillations to match that of the stimuli. The top of panel B shows a stimulus with temporal regularity arriving at sub-optimal phases of neural oscillations. In this case the incoming stimuli would be processed less efficiently. The bottom of panel B demonstrates how by shifting the phase of brain oscillations to match the temporal regularity of the incoming stimulus, processing efficiency can be improved

### Neural entrainment to speech

A great deal of work has been performed regarding both where and how speech is processed in the brain. This work has a wide scope including identifying areas in the brain that selectively respond to human voices (Belin, Zatorre, Lafaille, Ahad, & Pike, 2000), elucidating how the brain encodes voices (Latinus, McAleer, Bestelmeyer, & Belin, 2013), determining how the brain processes phonetic structure and pitch (Zatorre, Evans, Meyer, & Gjedde, 1992), suggesting how the cortex might be organised to process speech (Hickok & Poeppel, 2007), demonstrating the role of motor areas in speech perception (Wilson, Saygin, Sereno, & Iacoboni, 2004) and investigating the cortical representations of linguistic structure (Ding, Melloni, Zhang, Tian, & Poeppel, 2016). Within work performed to uncover the mechanisms of speech processing in general there are a number of studies that aim to determine how continuous speech is interpreted by the brain. It is in these works that the link between neural entrainment and speech can be found.

Giraud and Poeppel (2012) provide a summary of how speech processing can be

driven by neural entrainment. They base their summary on earlier studies (Poeppel, 2003; Giraud et al., 2007; Ghitza, 2011) and propose that the phase of neuronal oscillations is reset in relation to the *speech envelope* of the speech signal. The speech envelope can be considered as a smooth curve outlining the extremes of a sample of speech. An example of the relationship between the speech waveform, the speech envelope (amplitude envelope in this case) and syllable boundaries adapted from Cummins (2012), can be found in figure 2.4. In figure 2.4, it can be seen that the

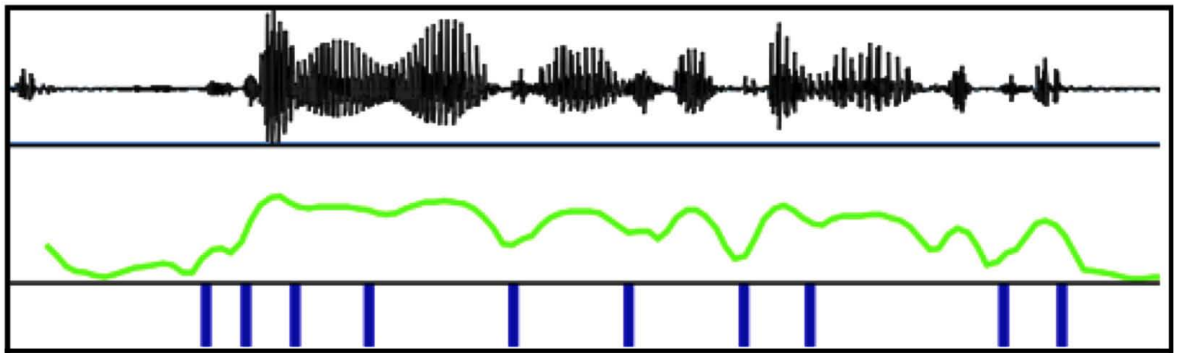


Figure 2.4: In this adapted image from Cummins (2012), the top panel shows the speech waveform of a Slovak sentence. The middle panel shows the amplitude envelope of the same sentence and the bottom panel shows the approximate syllable boundaries for that sentence.

amplitude envelope of the example sentence broadly tracks with the upper extremes of the speech waveform. Information about the fine detail contained in the signal is not captured. Although there are a number of different ways to extract speech envelopes (some of which might be better suited to studying neural entrainment than others (Biesmans et al., 2015)), an explanation of the methods for extraction is not necessary here. It will suffice to say that the speech envelope captures only the global trends in a speech sample.

The relationship that Giraud and Poeppel (2012) suggest between the speech envelope and neural activity is localised to the auditory cortex (Zatorre, Belin, & Penhune, 2002). They suggest that ongoing oscillatory activity in these areas interacts with the neuronal activity of the incoming auditory speech signal. The salient 'edges' of this signal, as represented by the speech envelope, then proceed to reset the phase of the ongoing oscillatory activity in the auditory cortex. Specifically, this phase-resetting is thought to occur in the theta band (4 to 8 Hz) of oscillatory activity, which entrains to track the speech envelope. The gamma band (30 to 70 Hz) of oscillatory activity, which is also proposed to track with the speech stimulus, is in a nested relationship with theta activity. This relationship is such that the phase of theta dictates the amplitude and possibly the phase of gamma. Working together, these oscillations transform the auditory stimulus into a discrete, temporally organised neuronal activation (spike) train which tracks with the incoming stimulus. Theta dictates the chunking of the activation train whilst gamma dictates the fine detail. This is graphically represented in figure 2.5, drawn from Giraud and Poeppel

(2012).

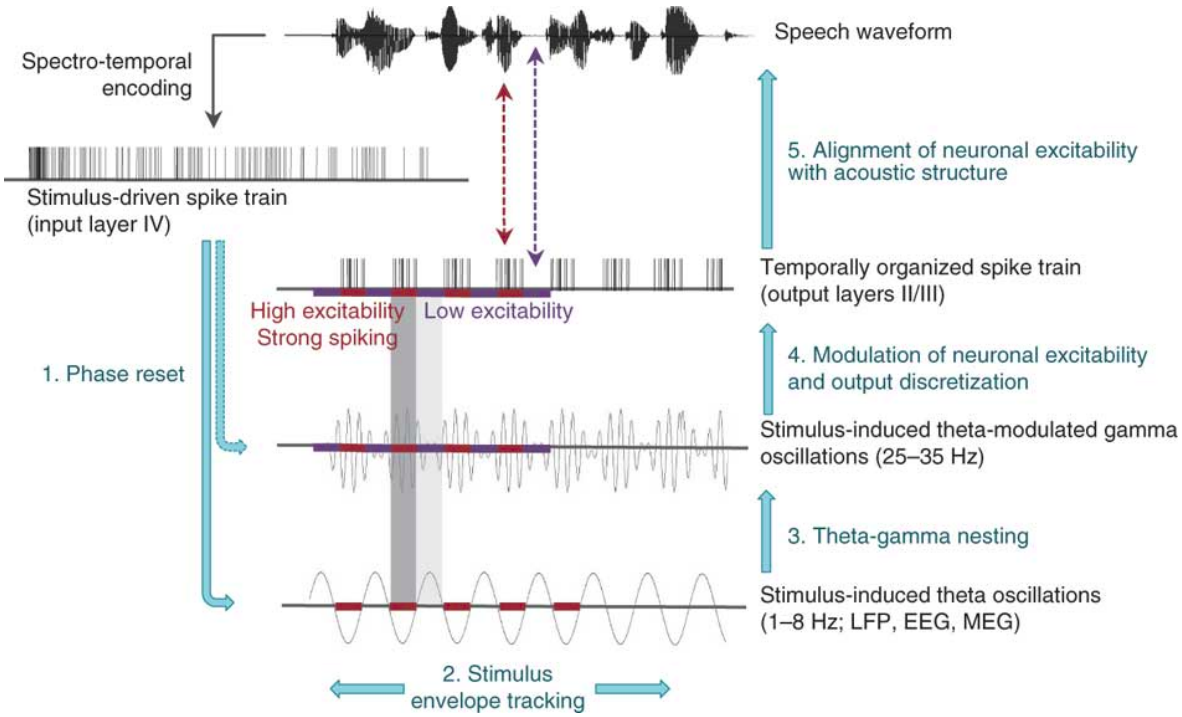



Figure 2.5: Graphical representation of neural entrainment to speech in the auditory cortex as proposed by Giraud and Poeppel (2012). Image taken from Giraud and Poeppel (2012)

.

Giraud and Poeppel (2012) go on to offer a neuro-biological account of the neural structures that would facilitate this model of speech processing. Whilst the specifics of the biological considerations of neuronal structure are not wholly pertinent to this thesis, the functional purpose of the structuring is. Their account of the structuring of neurons makes it possible for the information encoded from the stimulus to be interpreted in discrete chunks that allow for further analyses to be performed. It is suggested that chunking of the incoming speech information into discrete units, with nested temporal sensitivity, allows for phonological abstraction that helps to encode phonemic and syllabic information. It is important to note that thus far in the model proposed by the authors, all oscillatory activity has been constrained to the auditory cortex. It is thought that the auditory cortex provides the initial processing of the speech signal (and other auditory signals) which can then be passed onto additional neuronal structures for further processing. Evidence in support of this view is also presented as the authors find that oscillatory activity in frequencies relating to those believed to track with speech stimuli are found elsewhere in the cortex (the somatosensory cortex and articulatory motor cortex). The key observation in the case of linked oscillatory activity outside of the auditory cortex is that this activity is not found during rest, it was only found during speech processing. Since activity in the auditory cortex and the motor cortex is triggered during processing, this provides some neurobiological support for the link between perception and production discussed in subsection 2.3.1.

Additional work has taken this view of speech processing further. Doelling, Arnal, Ghitza, and Poeppel (2014) suggest that the 'sharpness' or the clearness of syllables plays a key role in the intelligibility of speech. They performed an MEG study looking at the neural coherence in response to speech samples that had been edited to vary their sharpness. Participants were presented with auditory stimuli that had either been stripped of the temporal cues (no syllables, just fine acoustic-phonetic detail), left with only the temporal cues (no fine acoustic-phonetic detail, only syllables) or an amalgam of both edited stimuli and were asked to rate intelligibility. They theorised that the conditions with no temporal cues would demonstrate less neural coherence and would also show less intelligibility when compared against a control condition. Indeed, they showed that neural coherence and intelligibility was poorest in the condition stripped of temporal cues. Interestingly though, they also found that intelligibility (although not neural coherence) was poor for the condition containing only temporal cues. The authors suggest that this is attributable to a lack of information being present after processing, due to the removal of acoustic-phonetic detail, rather than inefficient processing. Their findings support the view that speech information is more efficiently processed when coupled with oscillatory neural activity but that coupling to the speech envelope alone is not sufficient for comprehension.

Recent research from Kösem, Bosker, Meyer, Jensen, and Hagoort (2016) has attempted to reconcile the idea of neural entrainment being predictive of upcoming stimuli with the models of oscillatory speech processing put forward by Giraud and Poeppel (2012). One of the predictions of the oscillatory models of speech processing is that the entrainment to stimuli should be invariant to speech rate, that is to say that they function irrespective of speech rate. Kösem et al. (2016) performed an MEG experiment to test this prediction. Participants were presented with a series of sentences where the opening of the sentences, the carrier window, had varying speech rates (fast or slow). The final three words of these sentences, the target window, had an intermediate speech rate in all cases. The final word was ambiguous with regards to its vowel duration (long or short). The participants were asked to identify the final word of the sentences. Behavioural results demonstrated greater perception of long vowels when preceded by fast speech and greater perception of short vowels when preceded by slow speech. For the neural activity measured during the carrier window, the data replicated existing findings with oscillatory activity entraining to the frequency of the speech rate. In addition, traces of the oscillatory activity entrained to the carrier window frequency were detected in the target window. Further to this, the strength of the oscillatory carry-over into the target with was significantly correlated with the reported perceptual effects. That is to say that those participants whose perceptions were more strongly influenced by the speech rate in the carrier window demonstrated greater carry-over into the target window. These findings provide further support for the proposal that neural entrainment both

predicts features of upcoming speech and that it also influences processing of speech stimuli.

Naturally, no theoretical model goes without criticism. Obleser, Herrmann, and Henry (2012) and Cummins (2012) offer two notable criticisms of the proposals of Giraud and Poeppel (2012). Obleser et al. (2012) take a neuropsychologist's perspective on the potential issues with the oscillatory models of speech processing. Whilst they applaud Giraud and Poeppel (2012) for their 'visionary' and 'synergistic' perspective, a number of issues are raised. Obleser et al. (2012) argue that the proposed class of models:

1. Lack specificity in defining the oscillatory frequencies and the relationship between them.

2. Provide little consideration for top-down mechanisms that could modulate entrainment.

3. Rely too heavily on the role of the temporal speech envelope.

Each of these points are addressed in a response to Obleser et al. (2012) by Ghitza, Giraud, and Poeppel (2013) and as such, they will be considered together here. The criticism put forward by Obleser et al. (2012) regarding the lack of specificity for the oscillatory frequencies suggests that without a clear delineation of frequency bands for interpretation, researchers run the risk of overlooking important functional differences between activity bands. In response to this, Ghitza et al. (2013) acknowledge that there is a lack of specificity in the neurophysiological version of the model but go on to point out that greater specificity can be found in the phenomenological model (Ghitza, 2011). There is also a suggestion that further research is needed to neurophysiologically validate the model but that the ability of the model to explain behavioural findings provides context for future such experiments.

The term 'top-down' in the context of the Obleser et al. (2012) criticism refers to the influence of linguistic information. They cite Peelle, Gross, and Davis (2013) where it was demonstrated that phase-locking between neural activity and speech was found to be greater in the presence of linguistic information for stimuli with identical amplitude envelope characteristics. This work is taken to suggest that the information contained in the amplitude envelope alone is not sufficient to speech comprehension via neural entrainment. Again, Ghitza et al. (2013) respond to this criticism by linking it with a response about frequency band communication. They are in agreement with Obleser et al. (2012) that intra-band communication should be incorporated into their model. It is stated that they imagine delta oscillation to play a key role in the prosodic parsing of syllables and that delta oscillation interacts in a top-down fashion with the driving theta oscillation. However, they note that their model is not a 'one-size-fits-all solution for all of speech comprehension' (pp.1) but rather is restricted to syllable recognition without context.

The final criticism from Obleser et al. (2012) is that there is an over reliance on the speech envelope. They presented evidence that demonstrated neural entrainment to frequency modulations that held a consistent amplitude and therefore had no variance in the amplitude envelope. If entrainment is to the amplitude envelope of the speech waveform then there should not be entrainment when the amplitude envelope is flat. Ghitza et al. (2013) responds with a theorem from the field of communications that allows for this, given that a filter has been applied (Rice, 1973). It is then suggested that the human cochlea is a type of filter and that if the oscillatory model is considered with cochlear filtered speech (which is the brain's main source of speech input), then the model still holds. However, the reliance on the speech envelope has additional issues that relate to concerns from a phonetician's point of view.

The other notable criticism of Giraud and Poeppel (2012) comes from Cummins (2012). In this critique, the main points of contention are that:

1. Syllables are not necessarily recoverable from the speech envelope.

2. Oscillatory models do not have the necessary structure to encode the rich spectral information in speech.

3. There is a lack of representation for the complexity of both speech and speakers.

As with the criticism from Obleser et al. (2012), a response to Cummins (2012) was offered by Ghitza (2013). In that response a case is made for the 'theta-syllable', conceptualised as a unit of speech information defined by neural function.

Cummins (2012) makes the point that whilst syllables are an easily perceived and produced by speakers but the mapping of speakers' percepts onto the acoustic signal is not a trivial task. He notes that the the amplitude envelope is not necessarily an adequate representation of syllable phase. Indeed, in figure 2.4, which is drawn from Cummins (2012), it can be seen that the syllable boundaries aren't necessarily consistent with the amplitude envelope. Ghitza (2013) concedes that syllables are inherently ambiguous when drawn from the acoustics of speech. What is proposed is that the information source used to entrain is actually the output of the cochlea and that this drives the creation of a theta-syllable. The theta-syllable is defined as 'a theta-cycle long speech segment located between two successive vocalic nuclei' (pp.5). This is then used in addition to an ongoing process at an idling frequency to improve processing through entrainment and prediction.

Where neural oscillations are assumed to be periodic in nature, the speech signal is quasi-periodic at best. A mis-match between the two systems in terms of their oscillatory behaviour is suggestive of loss in or sub-optimal information transfer. Given the rich amount of spectro-temporal information in the speech signal, this could present a problem. Cummins (2012) picks up on this issue and suggests

that the oscillatory models of entrainment are not sufficient to decode all of the information contained in the speech signal. To tackle the criticism that oscillatory models lack the ability to decode the rich spectro-temporal content of speech, Ghitza (2013) provides an extended explanation of the mechanics of the oscillators used in the model. It is explained that the types of oscillators assumed in the model are able to gradually change frequency to adapt to changes in the incoming cochlear signal. Further to this, the point is made that the oscillator model of speech processing aims only to account for some components of the parsing and decoding of speech and that it is not a one-size-fits-all solution.

Finally, Cummins (2012) presents work that suggests that speaker and listener are both different parts of the same dynamical system (Cummins, 2009) where intelligibility is core to synchronisation. It is argued that the neural oscillation models do not take into account the intimate knowledge that speakers have of their own language and that this knowledge must be taken into account when considering synchronisation. The reply to this was that within the neural oscillation literature, the use of the term synchronisation is in a less restrictive sense than that employed by Cummins (2009). Further to this, it is suggested that the mechanisms that underlie the neural oscillation account of speech processing cannot be disentangled from intelligibility.

The debate surrounding the exact mechanisms of oscillatory activity in neural entrainment to speech continues. However, the evidence for a relationship between oscillatory activity in the brain tracking with incoming continuous speech is rather compelling. In addition, these theories have seen success in improving the performance of speech recognition systems (Lee & Cho, 2016), providing further support for a role in the computational contribution to speech processing. Having said that, the opening of the closing statement in Cummins (2012) which reads:

> There seems to be a need here for the development of formal models that can capture the reciprocal coupling of speaker and listener, taking into account their implicit but hugely constraining practical knowledge of what it is to speak. (pp.2)

holds a particular resonance for studies viewing speech as a fundamentally interactional process. Ultimately, it is likely that neuronal oscillatory behaviour does play a role in the encoding and processing of speech information but considerable further investigation is needed (Ding & Simon, 2014).

### 2.3.3 Linking accommodation and brain activity

Subsection 2.3.1 discussed how theories of the cognitive mechanisms that underlie accommodation all assume, in one form or another, a link between the perception and production of speech. Subsection 2.3.2 outlined possible neural mechanisms for how the brain might entrain to the speech signal in order to improve processing

efficiency. This subsection aims to draw those two areas together to explore how accommodation and brain activity in interacting speakers might be linked. In order to do this, literature on brain activity during joint activities will be drawn upon.

Social interactions, which spoken engagements can be considered to be, are a fundamental aspect of human behaviour and play a critical role in human development. As such, they have attracted a good deal of research in the field of neuroscience in order to understand the neural activity associated with it (eg. Iacoboni et al., 1999; Rilling et al., 2002; Schilbach et al., 2006; Prehn et al., 2015). However, studies such as these are often restricted to the investigation of a single person in response to a pre-recorded stimulus due to the constraints of the neuroimaging tools. Because of this, the studies cannot be said to be truly interactive and therefore do not evaluate social interaction in a socially relevant environment. Recently, there has been a move towards viewing brain activity during social interactions within a dual or multi-person environment (see: Hari & Kujala, 2009; Babiloni & Astolfi, 2014). This move has been made possible through advances in imaging methodologies, allowing for the joint measurement of brain activity in interacting participants and has been termed *hyperscanning* (Montague et al., 2002; Dumas et al., 2011; Konvalinka & Roepstorff, 2012). The use of these hyperscanning techniques allows for the study of the human brain during joint processes between individuals and to uncover the ways in which behaviour impacts joint neural activity.

Of particular interest to this thesis is the possibility that neural activity between participants can become entrained when engaged in a joint process such as taking part in a spoken engagement. Dumas et al. (2011) provide a review of the hyperscanning literature suggesting investigating behavioural synchronization and neural entrainment. They offer the following statement about the link between joint processes and neural entrainment:

> We believe that engaging in joint attention processes prompt coordinated actions between the participants, which might lead to interindividual neural synchronization. (pp. 49)

To back up this claim they present findings from a number of studies that have each made contributions to understanding the link that is made between brains when engaged in a social or joint interaction. They provide Montague et al. (2002) and King-Casas et al. (2005) as examples of hyperscanning techniques using fMRI to provide evidence for functional correlations between specific areas of the brains of two people engaged in a social interaction. However, the main focus of the review is around the use of EEG in hyperscanning contexts. Primarily, these EEG hyperscanning studies used decision making games to investigate social interaction. The games used included the prisoner's dilemma (Babiloni et al., 2007; Astolfi et al., 2009), card games (Babiloni et al., 2007) and the ultimatum game (Yun, Chung, & Jeong, 2008). These studies helped to show that specific regions in the brain are

activated in all participants in response to specific behavioural stimuli. In the case of Yun et al. (2008), it was found that certain frequencies of neural oscillations in a specific region of the interacting brains was closely related to the social interaction. Lindenberger, Li, Gruber, and Müller (2009) demonstrated that during synchronised guitar playing or joint listening to a metronome, the oscillatory activity of the participants becomes coupled. Further to this, in experiments conducted by Dumas, Nadel, Soussignan, Martinerie, and Garnero (2010) it was demonstrated that during a bi-manual (two handed) movement task where participants had to mimic hand movements of a partner, there was a coupling between the oscillatory activity of the participants brains across a range of frequencies. It is suggested that these findings demonstrate coupling between participants at a number of levels and that the activity at different frequencies codes for this. The theta band was suggested to code for broad aspects of the interaction such as hand position and speed since it is continuously in synchronization between participants. Higher frequencies such as beta and gamma demonstrated an asynchrony between participants, the authors interpreted this as a demonstrating a differentiation between information processing levels, where higher frequencies are involved in top-down modulation of oscillatory behaviour, transient motor movements and attentional capacity. Overall, the review highlights the importance of hyperscanning approaches in the understanding of neural activity during joint attention and the papers discussed provide a grounding for the investigation of neural entrainment between interacting participants.

When considered in the context of a spoken interaction a main behavioural correlate to neural activity is the speech signal. Given the aforementioned theories and studies exploring entrainment of the oscillatory activity in the brain to the speech signal (subsection 2.3.2), it could be theorised that ongoing interactions between speakers might lead to inter-speaker neural entrainment. Indeed, studies looking at the inter-brain synchrony of individuals engaged in a complex joint activity (flight simulation) were shown to have dense patterns of inter-brain connectivity during the most cooperatively intensive sections of the task (Astolfi et al., 2012; Toppi et al., 2016). However, studies directly investigating this effect in conversational interactions are few. This is because of the inherent issues of measurement when jointly recording brain activity and speech. Of what can be considered the three main methods of recording brain activity, fMRI, EEG and MEG, all suffer from sensitivity to head movement, meaning that recording whilst speaking produces highly contaminated data. In addition to this, since the temporal resolution of fMRI is on the order of seconds, it is too low to evaluate the ongoing activity related to the speech signal. Having said this, a few select examples can be found.

Whilst not strictly a hyperscanning study and although they don't address the speech signal explicitly, Kuhlen, Allefeld, and Haynes (2012) provide an example of how the EEG signals of two interacting speakers can become coupled. Participants were either classed as speakers or as listeners, the speakers were required to

produce a recital of a story whist wearing an EEG cap and having their faces and speech recorded. Once all the speakers had been recorded, the video and audio recording of two speakers were overlayed on each other. This produced a composite recording such that one could attend to either of the speakers. This was then presented to the listeners who were tasked with attending to either one or the other of the stories whilst wearing an EEG cap. Results demonstrated a significant correlation between the brain activity of the listener and the speaker to which they were attending. No such correlation was found between the listener and the other, superimposed speaker. Unlike studies that explicitly looked at linked neural behaviour in homologous brain areas (eg. Stephens, Silbert, & Hasson, 2010), here the analysis was not restricted to any one brain area. As a result of this, a relationship was found between neural activities in different brain areas. The relationships that were found in the data mainly pertained to activity in the low frequency bands ($< 3$ Hz). This is in keeping with models of oscillatory entrainment (Giraud & Poeppel, 2012) and hierarchical models of language processing (Pickering & Garrod, 2013) and is interpreted as such by the authors. Additionally, it should be noted that this study provides some precedent for the use of EEG during speech production, which has traditionally been seen as a limiting factor in EEG study design (discussed further in subsection 2.4.2).

One study that aimed at evaluating the relationship between behavioural signals (speech) and neural activity was performed by Kawasaki, Yamada, Ushiku, Miyauchi, and Yamaguchi (2013). In this study the speech feature of interest was speech rhythm, here taken to be the duration and interval of pronunciation for English letters. The authors had pairs of participants take part in an alternating speech task. The task itself asked the two participants in each pair to alternately and sequentially produce the alphabet (in English) whilst both wearing EEG caps. Participants performed this task in pairs (human-human), then with a machine (machine-human), then again in pairs (human-human). This allowed for the experimenters to observe if the same type of behavioural synchronisation, in the speech rhythm, is found when interacting with a non-human entity. Results demonstrated that the behavioural response, speech rhythms, were more likely to become synchronised (ie. have smaller differences between the duration and intervals used by each of the speakers) in human-human compared to human-machine interactions. The authors interpret this as evidence that synchronisation in the speech signal is specific to human interactions since the interval used by the machine in the machine-human interactions was fixed and would therefore have been easy to synchronise to. Further to the results of the behavioural data, the authors also found that brain activity in the theta and alpha bands (6 to 12 Hz) to become synchronised between participants. These are frequencies that have previously been associated with both working memory during joint tasks (Dumas et al., 2010) in the case of alpha rhythms and speech tracking (Giraud & Poeppel, 2012) in the case of theta rhythms. These find-

ings suggest a link between oscillatory neural activity and the speech signal, at least at a very coarse level such as speech rhythm.

Another study of interest, although yet to be published, is that of Jensen, Borrie, Studenka, and Gillam (2016). In this study a group of ten participant pairs were asked to sit facing each other and were then given pictures which were slightly different. They then had to complete a spot-the-difference task using the pictures they were given and through verbal interaction only (ie. they could not see each other's pictures). Participants took it in turns to describe their picture to their partner, each turn lasted for ten seconds. Upon the end of a turn a beep sounded and the other participant spoke for ten seconds, the experiment proceeded in this fashion for five minutes for each pair. Brain activity was tracked using a functional near infra-red spectroscopy (fNiRS) hyperscanning set-up. Prior to engaging in the task, participants' brain activity was measured at rest in order to form a baseline measure. Neural coherence between the participants was calculated both during rest and during interaction, this was done across a frequency band of 7 to 50 Hz. Results demonstrated greater neural coherence during interaction when compared to rest at both the group and individual dyad levels. The authors take this as evidence that participants' neural activation patterns align to a greater extent when engaged in spoken dialogue. This study focused on the frontal cortex and had participants face each other during the task; as the authors point out, this could be interpreted as an increase in the social cognition between the speakers. The authors suggest that this allows the conversational partners to 'mentalize, make judgements, internalize perceptions and to monitor the success of the conversations' (pp. 9).

Taken as a whole, the literature presented in this section can be taken as evidence for the possibility of inter-brain synchronisation during a joint interaction such as a spoken conversation (Lindenberger et al., 2009; Kawasaki et al., 2013; Jensen et al., 2016). Moreover, it provides evidence for increased neural coherence between participants during more demanding tasks (Lindenberger et al., 2009; Astolfi et al., 2012; Toppi et al., 2016). Thus, it might be reasoned that in situations requiring high levels of coordination, a higher degree of neural coherence could lead to greater production of more similar speech characteristics via entrainment to a partner's speech properties. Such an interpretation would be consistent with the theories underpinning the cognitive mechanisms of accommodation, especially the view put forward by Pickering and Garrod (2013). It would also be consistent with theories involving the entrainment of oscillatory activity to the speech signal and theories that propose an inter-brain synchrony during social interaction.

### 2.3.4 Summary

This section has aimed to do three things, (1) explore the possible cognitive mechanisms underlying accommodation, (2) explain the concept of neural entrainment

and how it relates to speech processing and (3) link accommodation and brain activity. Here, a brief summary of each of these aims is provided.

Subsection 2.3.1 reviewed four key theories that provide schemas for the cognitive mechanisms that might underlie a process such as accommodation. These were, the motor theory of speech, the direct-realist approach, the episodic memory model and the mechanistic approach to dialogue. Each of these theories was distinct in the way that they proposed the cognitive mechanisms to work. Some believed that the target of perception was linked to mental representations of articulatory commands (Liberman & Mattingly, 1985) whilst others believe the speech is fundamentally interactive and that it therefore must involve alignment at multiple levels of representation (Pickering & Garrod, 2004; Pickering & Garrod, 2013). Irrespective of which theory is closer to the truth of the cognitive mechanisms underlying accommodation (or speech processing in general) one thing that they all have in common is a belief that perception and production are linked in some non-trivial way. This is important when considering accommodation because it allows for a flow of information from the incoming speech signal of a partner to impact on one's own speech productions. Moreover, if perception and production mechanisms are linked, then it would provide both backing and a functional route to explain the apparent automaticity of accommodation.

Subsection 2.3.2 explored how neural entrainment suggests the coupling of oscillatory activity in the brain to environmental stimuli. It demonstrated how the brain can capitalise upon regular and semi-regular critical events in ongoing signals generated by the environment to improve processing and to allow for predictions to be made. Further to this, work was presented that suggested a specific mechanism for the processing and encoding of speech in the auditory cortex (Giraud & Poeppel, 2012). This suggested that long term aspects of the speech signal were used by the brain to reset the phase alignment of a 'master' oscillatory theta band which then had nested gamma band activity that allowed for processing of finer detail. Whilst there were a good number of studies demonstrating effects such as this, there is still a great deal of debate surrounding the topic and more work is needed to gain a full understanding of the finer mechanisms of neural entrainment to the speech signal. Whether the brain truly entrains to syllables, speech rhythms or another feature of the speech feature during speech processing is still under investigation. That said, it seems reasonably clear that the brain does use the oscillatory activity to entrain to environmental stimuli.

In subsection 2.3.3 the concept of neural entrainment between participants engaged in a joint activity was explored. The methodology of recording the brain activity of two participants in tandem, known as *hyperscanning*, was introduced and studies employing it were presented. Of the studies that make use of hyperscanning to investigate the neural correlates of social engagement, the majority did not consider speech. However, it was consistently shown that when engaged in a joint

activity, there are links between both the areas of the brain that are activated in each participant (eg. Montague et al., 2002; King-Casas et al., 2005) and the oscillatory activity between the participants (eg. Lindenberger et al., 2009; Dumas et al., 2010). Of the studies that have tried to evaluate the link between speech and neural activity during vocal interaction (Kuhlen et al., 2012; Kawasaki et al., 2013; Jensen et al., 2016), all have demonstrated a link between the brain activity of participants but only one has also shown an effect in the speech signal (Kawasaki et al., 2013). Work in this area is still in its infancy and a great deal more research is still needed to both strengthen current findings and to refine the technique.

Taking all of this together, it could be theorised that neural entrainment to the speech signal of another speaker could produce a degree of inter-speaker entrainment. This would be especially true in situations that require high levels of coordination between participants since studies suggest that greater entrainment may result in improved processing. This greater degree of speaker entrainment and improved processing might then result in an automatic transference of the partner's speech characteristics to the production system, leading to increased convergence. Of course the inverse would be true as well, where in instances of poor entrainment, this loop gets disrupted and leads to greater variation in the speech features of the speakers. In addition to all of this, there would also be the influence of top-down speaker knowledge and biases which could impact the system, but this would likely only take place in comparatively extreme situations where conflict arises. If entrainment such as this is a phenomenon of speech communication, then periods of speech convergence would involve greater neural entrainment to the interlocutor's speech whilst speech divergence would demonstrate less entrainment.

Having said all of this, certain issues around the recoding and interpretation of speech and brain signals in tandem still remain. For instance, there is still no consensus on which bands of oscillatory activity in the brain are most critical for the evaluation of speech. This is especially true considering the lack of interpretation of top-down information in models of oscillatory neural entrainment. In addition, if attempting to record speech and brain activity in tandem considerations need to be made as to how the influence of movement artefacts in the neural signal will be reduced/eliminated. Further to this, there is also no consensus on the phonetic correlates of accommodation. In fact, the evidence points to the use of different phonetic features given the context and background of the speakers (see sections 2.1 and 2.2). Such subjects are what will be discussed in the following section, section 2.4.

## 2.4 How can accommodation and brain activity be measured in tandem?

So far in this chapter the phenomenon of accommodation has been described in section 2.1, the methods used to measure accommodation have been reviewed in section 2.2 and the possible links between accommodation and joint brain activity have been discussed in section 2.3. The subject that has not yet been addressed is how one would go about producing a measure of accommodation that can be used in interactional speech which avoids some of the shortcomings associated with more traditional methods. In addition to this, the measure should also be able to be used in tandem with neural data to make some assessment of the link between accommodation and joint brain activity. This section presents an approach, drawn from machine learning and automatic speech recognition literature, that could be used as a starting point for development of such measures.

### 2.4.1 Machine Learning - a helping hand(?)

Alpaydin (2014) describes machine learning as 'programming computers to optimize a performance criterion using example data or past experience' (pp. 3). This description of machine learning can be unpacked and considered in terms of a real-world example by specifying the performance criterion. If it is specified generally as 'speaker recognition' the description of machine learning becomes more clear and pertinent to the work discussed in this thesis. Now we have the phrase 'programming computers to optimize speaker recognition using example data or past experience'. Indeed, machine learning has been used for a number of years in automatic speech recognition systems for both speech recognition in general (Hinton et al., 2012; Deng & Li, 2013) and for the recognition of specific speakers (Wan & Campbell, 2000; Lan, Hu, Soh, & Huang, 2013). The challenge for this thesis is to apply machine learning not to optimize speaker recognition but to optimize the recognition of accommodation. This subsection aims to explain how machine learning can be used to provide an alternative tool to traditional methods for detecting accommodation.

As Alpaydin (2014) points out, machine learning, like all learning, requires example data or past experience. If the example of speech recognition is carried forward, an intuitive example of what is meant by this can be provided. If asked to describe the voice of one's mother to a stranger, it would be difficult to provide precise criteria that could be used by the stranger to identify her voice. If, on the other hand, the stranger was allowed to listen to a number of extracts of her voice, they would unconsciously pick out structural features that help to identify her voice from other voices. This is possible because there are regularities in the speech signal that are specific to one's mother's voice. Whilst it may not be possible to explain exactly

what features the stranger uses to differentiate her voice from that of other speakers, it is something that humans are able to do with a high degree of accuracy and in light of considerable noise and variation. Machine learning approaches work in a similar way to that of a stranger learning to recognise the voice of a novel speaker. Each speaker's voice is composed of a pattern of acoustic features specific to them. In this example, a machine learning approach would analyse example speech samples in order to determine the patterns that pertain to a particular speaker. Recognition would then be performed by comparison of these identifying patterns against new speech samples. The way in which machine learning goes about determining these patterns is dependent on the purpose for which it is being used.

In the case of accommodation detection, traditional acoustic-phonetic methods rely on the experimenter selecting the phonetic variables of interest. This is acceptable if the questions being asked about accommodation pertain to the acoustic-phonetic features in question but if the focus is on accommodation itself, this poses issues. As discussed in section 2.2, there are a number of factors that can impact on the phonetic variables that are used in accommodation. These include individual differences (Yu & Abrego-collier, 2011), local linguistic norms (Evans & Iverson, 2007), social settings (Pardo et al., 2012), speaker sex (Namy et al., 2002) and dominance (Pardo, 2006) to name but a few. Given all of the variation in accommodative behaviours, it would be useful to be able to characterise the general form of a participant's speech during a single experiment. This could then be used as a benchmark against which to compare any given speech sample for deviations (ie. accommodation). Characterisation of the patterns underlying structured signals, such as speech, is something that machine learning is designed for.

The use of machine learning to provide a general characterisation of a participant's speech in a given experiment would provide the underlying pattern of a participant's speech based on statistical regularities in the data itself. Where traditional methods of measuring accommodation would extract specific segments of speech (eg. vowels or stops), machine learning uses all available speech observations to produce a model that approximates the participant's speech. The benefit of this is that, a priori assumptions about the acoustic make-up of a participant's speech are not made. The patterns are determined by the data and this means that accommodation can be posed as a classification problem. The types of accommodation that were originally outlined in CAT (see section 2.1) of convergence, divergence, complementarity and maintenance could serve as the groups into which speech is classified. The method through which this is accomplished, as pointed out above, depends on the task at hand. Thankfully, machine learning is a vast field with many applications and there are pre-established methods to accomplish this classification task.

Machine learning is used in a vast array of different fields such as Statistics and Statistical Learning, Pattern Recognition, Signal and Image Processing and Analysis,

Computer Science, Data Mining, Machine Vision, Bioinformatics, Industrial Automation, and Computer-Aided Medical Diagnosis (Theodoridis, 2015). Within each of these fields there are a variety of different tasks that machine learning is employed for. For instance, machine learning based pattern recognition has been employed to detect subtle and complex disease induced changes in the brain (Sotiras et al., 2016). In contrast to this, machine learning based classification techniques have been employed to determine authorship of certain texts (Jockers & Witten, 2010). With the advent of the age of 'Big Data' many global companies are making use of machine learning in order to extract trends, enhance the consumer experience and predict consumer behaviour (Lin & Kolcz, 2012; Cui, Wong, & Lui, 2006). Machine learning has also been used in conjunction with imagery captured from both unmanned aerial vehicles and satellites to aid in disaster response, wildlife protection, human rights and archaeological exploration (Ofli et al., 2016). With such a broad range of applications, it is well beyond the scope of this thesis to present a full evaluation of machine learning as a whole. Rather, what will be focused on is a specific area of machine learning that is targeted at and has a history in speech recognition. More specifically, the focus will be on Hidden Markov Models (HMMs).

The reason for focusing on HMMs is that they have a proven track record in speech recognition (Gales & Young, 2008) and they provide a method of pattern detection that is comparatively simple in contrast to some more advanced techniques for speech recognition such as recurrent neural networks (Chan, Jaitly, Le, & Vinyals, 2016; Deng, Hinton, & Kingsbury, 2013). Given that the goal of this thesis is to offer an alternative to traditional methods of accommodation detection, it is sensible to begin with a reasonably simple implementation before more sophisticated techniques are employed. Further to this, HMMs have been used in speaker verification technologies to detect fraudulent attempts to imitate a speaker's voice (James, Hutter, & Bimbot, 1996; Reynolds, Quatieri, & Dunn, 2000). Although such systems can now be circumvented by synthesised speech (De Leon, Pucher, Yamagishi, Hernaez, & Saratxaga, 2012), the fact that they are able to identify individual speakers is of use to accommodation detection. If HMM based approaches can differentiate between speakers, that is to say that they can recognise patterns in the speech signal that differentiate one speaker from another, then it stands to reason that HMMs could be used to identify shifts towards or away from the patterns identified for another speaker.

What now follows is an elaboration of the way in which HMMs are used in speech recognition. First, a conceptual overview of HMMs is provided before concentrating on the detail.

### Hidden Markov Models

HMMs are able to characterise the general form of a continuous signal. When implemented for linguistic purposes, a HMM can be used to estimate the probability

of a given speech sound having been uttered by a particular speaker. They can be thought of as a way of generating a series of probabilities that provide the likelihood of a particular sequence of observations being produced given the previous sequence. The concept that underpins HMMs is something known as a Markov chain, named after the Russian mathematician Andrey Markov who studied them in the early 20$^{th}$ century (Seneta, 1996). Before moving on to consider HMMs, an explanation of Markov chains will be provided. A Markov chain is a probabilistic process that makes use of the current information to predict upcoming information. An important feature of Markov chains is that the upcoming information should be dependent on the previous information. These sets of information that are involved in the Markov chain can be better described as 'states'.

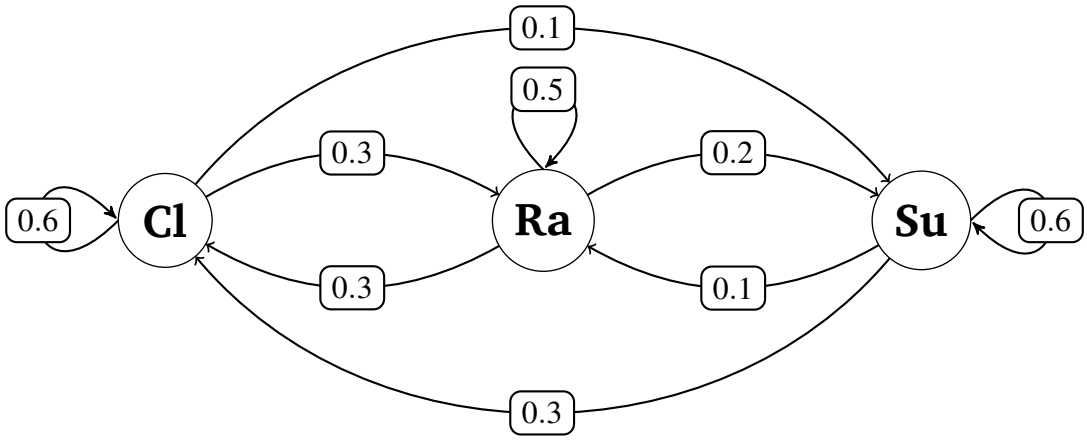



Figure 2.6: Example Markov state diagram for weather states. Circles represent the three weather states: cloudy (Cl), rainy (Ra) and sunny (Su). The possible transitions between states are represented by arrows, the probabilities for these transitions are found in the boxes associated with each arrow.

A widely used example of Markov chains is the prediction of the upcoming weather state based on the current weather state. If it is assumed that there are three possible states that the weather can be in, cloudy (Cl), rainy (Ra) and sunny (Su) a simple Markov chain can be described. In order to construct a Markov chain, data about the state of the weather must be collected for both the current state and the previous state. Assuming that each day in the year contains only one weather state and if data is collected over the course of the year, the probabilities for any given state being produced given the current state are able to be calculated. An example of this is given in figure 2.6. For each state there is a probability associated with transitioning from that current state to any other state, including remaining in the same state. The sum of the transition probabilities for any given state must be 1 because it is not possible for the following day to be neither cloudy, nor rainy, nor sunny, it must be one of these states. Taking all of these probabilities together allows for construction of a transition probability matrix. An example of this based on the probabilities in figure 2.6 is provided in table 2.3.

This transition matrix can then be used to determine the probability of the weather state for the next day given the current state. This can be done using a state probability vector. If the current state is rain then the probability vector, will be [0 1 0] in relation to the state sequence [Cl Ra Su]. This is so because the current state is rain, thus the probability of it being cloudy or sunny is currently 0 whereas the probability that it is raining is 1. This vector can then be multiplied by the transition matrix to provide the probability for the next state (ie. the probability of transition to another state). The result of this multiplication can be found in table 2.4 and is performed as follows, where the current state $= S1$ and the transition matrix $= M$:

$$S1 \times M$$

$$[010] \times \begin{bmatrix} 0.6 & 0.3 & 0.1 \\ 0.3 & 0.5 & 0.2 \\ 0.3 & 0.1 & 0.6 \end{bmatrix}$$

$$[(0.6\times 0+0.3\times 1+0.3\times 0);(0.3\times 0+0.5\times 1+0.1\times 0);(0.1\times 0+0.2\times 1+0.6\times 0)]$$

$$\begin{bmatrix} 0.0 & 0.0 & 0.0 \\ 0.3 & 0.5 & 0.2 \\ 0.0 & 0.0 & 0.0 \end{bmatrix}$$

[0.3 0.5 0.2]

In this example, the most probable state to follow rainy is again, rainy. Of course, in real world examples the probabilities are rarely this straightforward and are often far smaller, especially when considering Markov chains with a large number of state sequences. Further to this, what has been described here is known as a $1^{st}$ order Markov chain, meaning that only the state of the $1^{st}$ previous state is taken into consideration. It is possible to expand a Markov chain such that it considers the previous $n$ states, but this is beyond the purpose of this descriptive subsection.

If this Markov approach is extended to a speech example, it could be imagined that the states of interest might be different words in a sentence such as 'the fat cat'. At this point, HMMs can begin to be considered within a speech recognition context. If the goal is to recognise words then one could infer the states from the observations, provided that the probabilities of a particular observation occurring with a particular state can be estimated. In the weather example, this would be

| | | *Next State* | | |
|---|---|---|---|---|
| | | Cl | Ra | Su |
| *Current State* | Cl | 0.6 | 0.3 | 0.1 |
| | Ra | 0.3 | 0.5 | 0.2 |
| | Su | 0.3 | 0.1 | 0.6 |

Table 2.3: Example transition matrix for weather states.

| | | *Next State* | | |
|---|---|---|---|---|
| | | Cl | Ra | Su |
| *Vector* | Ra | 0.3 | 0.5 | 0.2 |

Table 2.4: Example multiplication of state probability vector with transition matrix for weather states.

akin to something such as determining the state of the weather given a set of observations about the condition of the ground (ie. is the ground wet or dry? Does the ground look dark due to the sky being overcast?). In this situation, one would not be able to detect the state of the weather directly, the only information available would be regarding the condition of the ground. This can be considered as a Markov chain with the additional constraint that there is now no way of explicitly determining the state. Instead, the states must be inferred from a sequence of related observations. In the case of word recognition, this sequence of observations is the speech waveform of the sentence and the states are the words themselves. This is why the approach is called a 'Hidden' Markov Model, because the states that are being estimated are not directly observable, they are 'hidden'. In this example, the only possible states are 'the', 'fat' and 'cat' although the order in which the states are presented is not known since the states are hidden. In order to build a HMM, as with the example for the weather, data must be collected. Here, orthographically transcribed data must be provided. The collected data is used to train a HMM such that the general features of each word are captured, allowing for a degree of variability (a Gaussian distribution). In order to capture these features, it is common practice to sample the speech signal in short windows over which the signal can be considered to be static, this is generally between $20$ to $40$ ms. A commonly used method for representing speech waveforms is to transform windows of speech data into Mel-Frequency Cepstral Coefficients (MFCCs, specifics found in subsection 3.6), which provides a cross-section of the spectral properties of the speech signal. Each new sample is considered in relation to the last to update the probability of that particular set of observations belonging to that particular word. This allows for a probability distribution to be built using the data contained in the speech waveform relating to each word. These probability distributions can then be used to predict the probability of a new set of observations belonging to any of the three states in this example ('the','fat' or 'cat'), as demonstrated in figure 2.7. This is done by determining the probability of any state occurring given the current observation given the parameters set by the probability distributions trained for the HMMs. Thus, it is possible to calculate the probability of the current observation being produced by any of the states, given the previous state.

If this is extended to the assessment of accommodation, it is possible to build a HMM for the general speech characteristics of two speakers, $A$ and $B$ based on samples of their speech, respectively. Given an HMM for each of the speakers, it is

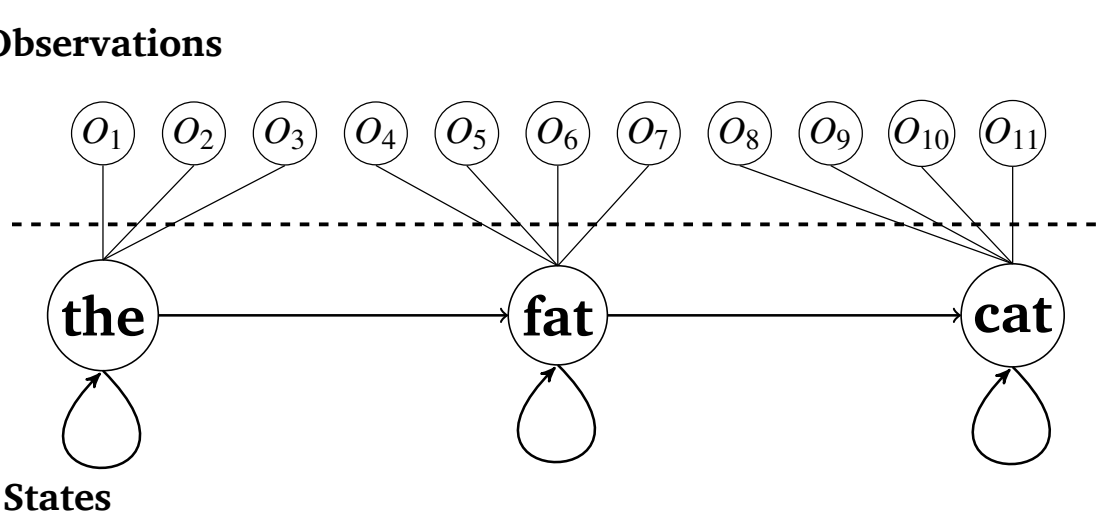



Figure 2.7: Example HMM diagram for speech states.

then possible to predict the probability that any given speech sample produced by $A$ could also have been produced by $B$. This would be done by submitting $A$'s speech data to $B$'s HMM for classification (and vice versa for $B$'s similarity to $A$). When this is done continuously across an interaction, it is possible to get a continuous string of probabilities detailing the similarity of the speakers. This can be considered as an ongoing measure of accommodation between the two speakers.

More specifically, HMMs are probability distributions defined over joint sequences of symbols (eg. phoneme categories) and observations (eg. MFCC coefficients). For speech signals, they can be characterised when considering a sequence $S = (s_1, s_2, \ldots, s_N)$ of states, where every $s_i$ belongs to a predefined set of symbols $V = \{v_1, \ldots, v_D\}$ , and a sequence $X = (\vec{x}_1, \ldots, \vec{x}_N)$ of observations, where $\vec{x}_i$ is a vector of physical measurements extracted from a speech signal at time $t_i$ ($t_j > t_i$ if $j > i$). A HMM is the joint probability $p(X, S|\Lambda)$ of observing $X$ and $S$ to occur together, where $\Lambda$ is the set of the parameters. The parameters' set $\Lambda$ can be characterized by taking into account the actual expression of the probability:

$$p(X, S|\Lambda) = \pi_{s_1} b_{s_1}(\vec{x}_1) \cdot a_{s_1 s_2} b_{s_2}(\vec{x}_2) \ldots a_{s_{N-1} s_N} b_{s_N}(\vec{x}_N) \tag{2.1}$$

where $\pi_{s_1}$ is the probability of the sequence $S$ starting with state $s_1$ (there are $D$ parameters $\pi_{v_i}$, one per element of $V$), the $a_{s_i s_j}$ are the probabilities of a transition between state $s_i$ and state $s_j$ (there are $D \times D$ parameters arranged in a matrix $A$ where element $i, j$ corresponds to the probability of a transition between $v_i$ and $v_j$), and $b_{s_i}(\vec{x})$ is the emission probability density function, i.e. the probability of observing $\vec{x}$ when the state is $s_i$ (there are $D$ distributions, one for each element of $V$, and each of them has parameters that are included in $\Lambda$).

In general, HMMs are used as follows: first a vector of physical measurements is extracted at regular time steps from a speech signal, resulting in a sequence $X$. Then, the sequence of states $S^*$, most likely to underlie the sequence of observations

is found by:

$$S^* = \arg \max_{S \in \mathscr{S}_N^{(V)}} p(X, S|\Lambda) \tag{2.2}$$

Where $\mathscr{S}_N^{(V)}$ is the set of all possible sequences of $N$ symbols each belonging to $V$. When $V$ contains $D$ symbols, there are $D^N$ possible sequences. This provides the probability distribution of any given speech sound being uttered by the speaker who produced sequence $X$.

When considered in relation to the detection of accommodation, HMMs allow for utterances produced by a speaker $A$ – whilst in interaction with a speech partner $B$ (eg. in a conversation) – to be tested against $A$'s general speech characteristics to determine if $A$'s speech changes holistically during interaction. This then provides a measure of speech accommodation across multiple acoustic features as represented by the observation vectorisation .

The use of HMMs in this way allows for all of the available data produced by a speaker to be taken into consideration when measuring accommodation. In addition, by sampling the speech data and extracting spectral cross-sections at each sample point, the process can be said to be accounting for a broad range of features of the speech signal. Further to this, the approach can also model temporal dependency depending on the level of specificity required. For instance, the states in the HMMs can relate to the general speech characteristics of a speaker without any temporal dependence (single state HMM or a Gaussian Mixture Model, GMM) or they can relate to the general speech characteristics of a speaker with temporal dependence (left-right HMM) or they can relate to the word specific characteristics of a speaker with temporal dependence (word-dependent HMM). A further description of the exact methodology employed in this thesis can be found in subsection 3.6. Taken together, these three aspects of an HMM (or machine learning) based approach to accommodation detection allows for more of the data to be utilised in comparison to traditional techniques, for a holistic measure of speech to be employed and for time to be considered at a local level within the speech signal.

### 2.4.2 Extending to brain data

Having described HMMs and how they can be employed in speech recognition and more specifically to accommodation detection in subsection 2.4.1, it is the aim of this current subsection to describe how they can be applied to brain data. The reasons why it is necessary to consider brain data in a non-standard way when assessing a phenomenon such as accommodation will also be explored.

In subsection 2.3.3 the potential links between brain activity and accommodation were explored. It was noted that the most evidence for a relationship between the speech signal produced by another person and on going brain activity in a listener

was found in the literature detailing neural oscillations (Giraud & Poeppel, 2012; Kawasaki et al., 2013; Jensen et al., 2016). Given that neural oscillations appear to be the most promising link between speech and brain activity, the method used to assess brain activity must be able to measure neural oscillations. Considering that the possible brain activity measurement tools available for this thesis are fMRI, MEG and EEG, this rules out fMRI. Although fMRI techniques have been used in many studies investigating speech (eg. Belin et al., 2000; Latinus et al., 2013; Blank & Davis, 2016) this technique generally indicates where activity is taking place and does not offer the ability to measure neural oscillations. Instead, one of the remaining two methods will need to be used.

Subsection 2.3.3 also noted that in order to determine the link between interacting speakers, a hyperscanning methodology would need to be employed (Montague et al., 2002; King-Casas et al., 2005; Astolfi et al., 2009; Dumas et al., 2011). Whilst hyperscanning has been demonstrated to be possible using MEG (Hirata et al., 2014; Zhdanov et al., 2015), the set-up of such a system is complex. Due to the size and expense of MEG machines, it is uncommon for any given lab to have more than one machine. This would mean that access to two machines that are geographically separated (possibly by hundreds of miles) would be needed and that they would have to be linked over the internet. It would also be necessary to split an experimental team across both sites to arrange participant recruitment and scanning. For this reason, MEG was ruled out as a possible tool. This leaves EEG as the remaining method for measurement of neural oscillations in an interacting dyad. As such, the remainder of this subsection will focus exclusively on EEG.

Despite the advantages of EEG, the potential hurdles for its implementation should be discussed. There are three key points that will need to be considered in order to implement EEG in an experiment that is investigating accommodation. These are:

- How to assess a continuous EEG signal for effects related to accommodation.
- How to eliminate muscular artefacts associated with speech from the EEG signal.
- How to perform an EEG hyperscanning experiment whilst recording speech.

A brief explanation of how EEG works and what it is measuring will precede explorations of each of these points.

As explained in subsection 2.3.2, the brain is made up of cells called neurons that produce electrical activity . EEG measures brain activity by detecting these electrical signals produced by the brain. EEG records the change in voltage across the surface of the scalp. It has been described as ‘a record of the oscillations of brain electrical potential recorded from electrodes on the human scalp’ (Nunez & Srinivasan, 2006, pp.3). The signals that are recorded are the net summation of the tiny electrical

signals generated by large populations of electrically active, individual neurons. The recording of ongoing electrical activity in the brain means that EEG has a high temporal resolution, being able to record changes at a millisecond scale. However, between the brain and the electrode on the scalp lies the dura mater, the skull, the musculature of the head and the skin. The electrical signal is impeded by all of this matter and is therefore weak by the time it reaches the electrodes on the surface of the scalp. To compensate for this, the electrodes are connected to an amplifier to boost the available signal. Each of these electrodes are kept on separate channels and EEG systems typically consist of $32, 64$ or $128$ channels (although higher density systems are also available, eg: Riedner et al., 2007; Trendafilov et al., 2016). In general, higher density electrode arrays allow for more accurate spatial localisation of brain activity (Laarne, Tenhunen-Eskelinen, Hyttinen, & Eskola, 2000). Whilst higher electrode densities allow for better spatial resolution, the inference of the source of electrical activity in the brain is impeded by the 'inverse problem' which has been compared to reconstructing an object from its shadow (Grech et al., 2008). The 'inverse problem' can be seen in many scientific disciplines and can be simply described as using the results of an experiment or test to calculate or infer the causes. In the case of EEG, it is using the electrical signal captured at the scalp (results) to infer the location in the brain that generated it (cause). Taken together, this means that EEG can be used to track the broad ongoing electrical activity in the brain but is not the most effective way to isolate the source of the activity.

The continuous monitoring of EEG signals has been useful in clinical environments for diagnosing conditions such as epilepsy (Young, Jordan, & Doig, 1996) and in intensive care (Scheuer, 2002). However, studies of perceptual or cognitive processes in the brain have often utilised techniques that average over many trials. These types of studies rely on the resulting event-related potentials (ERPs) that are detectable in response to a stimulus. An ERP is characterised by a small deflection from the the ongoing time course at multiple electrode sources in response to a presented stimulus (eg. visual image, auditory tone). The reason that these must be elicited from multiple trials is because of the inherent sources of noise in the EEG signal. These sources of noise can be generated from the electrical activity of muscles in the body (eg. eye or neck movements), the electrical activity of the heart and from the AC electrical signal generated by electrical machinery (Grech et al., 2008). Whilst the noise generated by electrical machinery is maintained at a constant frequency and is present throughout the EEG data, which therefore makes it relatively easy to remove, the noise generated by muscular activity is more inconsistent and harder to remove. This interference of muscular electrical activity has led to most EEG studies keeping the movement of participants to a minimum. It is often the case that participants will be asked to rest their head on a chin rest in order to avoid neck movement and to avoid blinking during stimulus presentation portions of an experiment (eg. Bieniek, Pernet, & Rousselet, 2012; Rousselet et al., 2014). Whilst

methods do exist for separating, non-neurophysiologically generated electrical activity from sources of noise (see: Jung et al., 2000; Delorme et al., 2007; Piazza et al., 2016), they can be complex to implement and have only been developed comparatively recently. The process of averaging across multiple trials makes it possible to increase the signal to noise ratio such that effects related to the presented stimuli become clearer and those relating to other sources are minimised. This is possible because the noise during each trial can be considered to be mostly random and that it will be likely to cancel out when averaged over many trials.

ERP studies have been used throughout research focusing on cognitive and perceptual processes (Campanella, Quinet, Bruyer, Crommelinck, & Guerit, 2002; Koelsch et al., 2005; Davis, Winkielman, & Coulson, 2015). Indeed, specific deviations from the ongoing time course of the EEG signals (N400, P600) have been linked to specific roles in the processing of language (Swaab et al., 2012). However, ERP studies require the presentation of tightly controlled, time locked stimuli in order to reliably infer that the electrical response measured is definitely in response to the presented stimulus. When evaluating a continuous spoken interaction, the stimulus onsets are not tightly controlled nor time locked. Speech is perceived as it is produced, the stimuli are not edited to investigate one specific element of speech or one particular feature of language. This means that there is no clear stimulus onset or clear differentiation between trials. For standard ERP approaches to EEG, which requires averaging over a large number of trials, this is a problem.

However, as explained in subsection 2.4.1, HMMs are able to characterise the general form of a continuous signal. The source of the signal does not matter, it can be data from the stock market (Hassan & Nath, 2005), amino acid strings in DNA (Wheeler & Eddy, 2013) or in this case, an EEG data stream. Recall that the construction of an HMM requires a string of observations in order to estimate the underlying states. Given that in this thesis, the phenomenon of interest is accommodation in the speech signal, the underlying states can be assumed to be the words and/or phonemes that are being produced during interaction. This allows for the data from the EEG signal of participants to be used as the sequence of observations that allow for estimation of the states. The only additional step for the application of HMMs to an EEG data source is to select an appropriate vectorisation parameter. In the speech example from subsection 2.4.1, it was mentioned that a commonly used vectorisation approach for the speech signal was to convert to MFCCs; this is specific to the speech signal and is filtered to better represent the scaling in the human cochlea.

In order to vectorise the EEG signal appropriately, a measure will need to be selected that better represents that activity of interest. There are a number of possible methods for accomplishing this (for a comprehensive review see: Gross, 2014) and specifics of the choice made for this thesis is provided in subsection 4.2.6. Once a characterisation of the signal has been made, provided that the EEG signal is time

locked to the speech signal, it is possible to compare the adaptation of the EEG signals produced by each speaker in relation to the speech being produced, the incoming speech of the partner and of both perceived and produced speech as a whole. As such, the application of HMMs addresses the problem of how to assess a continuous EEG signal for effects related to accommodation.

Although HMMs are able to filter out a good deal of noise from signal data by pruning in the probability matrices (minimising the probability of certain identified brain activity patterns occurring in relation to the target behavioural stimulus), the old adage of 'garbage in, garbage out' still applies. In this case, although the issue of not having multiple trials to average over may have been circumvented through the application of HMMs, the issue of noise from muscular activity is still present.

This brings the discussion to the second key point for consideration, how to eliminate muscular artefacts associated with speech from the EEG signal. Given that accommodation as investigated in this thesis requires a continuous verbal interaction between two speakers, it will be necessary to have participants speaking whilst wearing EEG caps. What this means is that there is likely to be a good degree of muscular noise in the EEG data collected. However, in recent years there has been a peak in interest surrounding the capture and analysis of EEG data during complex motor tasks such as walking (Gwin, Gramann, Makeig, & Ferris, 2010; Debener, Minow, Emkes, Gandras, & Vos, 2012) and talking (Tran, Craig, Boord, & Craig, 2004; Ganushchak, Christoffels, & Schiller, 2011; Porcaro et al., 2015). This work is particularly driven by advances in the field of brain-computer-interfaces (BCI) (Zhang et al., 2016; Minguillon, Lopez-Gordo, & Pelayo, 2017). The results of efforts to make the processing of EEG signals more robust to muscular artefacts are toolboxes for computational processing that are able to efficiently filter out signal elements that are not generated by the brain.

One widely used toolbox for the pre-processing and analysis of EEG data is EEGLAB (Delorme & Makeig, 2004) which is implemented under the MATLAB® computing environment and is continually being updated (Delorme et al., 2011). Currently this toolbox contains tools that allow for the processing of EEG for BCI, for real-time interactive EEG experiments and for EEG experiments involving physical movements (Delorme et al., 2011; Ojeda, Bigdely-Shamlo, & Makeig, 2014). Due to the fact that EEGLAB is widely used in the EEG community and because it has integrated tools allow for the removal of muscular artefacts in EEG data, it is a fitting analysis toolbox for the work in this thesis. More specifically, EEGLAB has a signal cleaning plug-in that utilises artifact subspace reconstruction (Mullen et al., 2013) to eliminate data that appear to be non-neurophysiological. This process relies on performing a Principal Components Analysis across sliding windows of EEG data to interpolate high-variance signal components. The interpolated time point in the EEG data is then reconstructed from the signal subspace. The signal generated by the process of speaking tends to produce high-variance components in the EEG

signal. Because of this feature, this approach using artifact subspace reconstruction should be suitable for removal of most speech artefacts.

Further to this, signal cleaning process, a standardised method for pre-processing data across participants has also been proposed and has been titled PREP (Bigdely-Shamlo et al., 2015). The PREP pipeline for EEG pre-processing aims to standardise early stage EEG processing. This is important because lack of attention at the early stages of EEG processing can lead to reductions in the signal to noise ratio at later stages (Bigdely-Shamlo et al., 2015) and this effect could be magnified when dealing with speech contaminated data. What the PREP pipeline offers is a standardised approach to the early stages of EEG processing such that any remaining artefacts further down the processing pipeline are minimised and that the process is standardised across all participants. After cleaning of the data using artefact subspace reconstruction and early stage pre-processing of data using PREP has been completed, the EEG data should be relatively artefact free and able to be submitted to other standard artefact removal processes such as independent components analysis (ICA). Utilising the recent advances in artefact removal should answer the second key point for the application of EEG, how to eliminate muscular artefacts associated with speech from the EEG signal.

The final point to consider in extending the investigation of accommodation to EEG data is how to perform an EEG hyperscanning experiment. There are already a number of studies that have employed an EEG hyperscanning methodology (eg. Babiloni et al., 2006; Dumas et al., 2011; Astolfi et al., 2011). However, it is still a relatively new technique for capturing brain activity. For the most part, the development of brain imaging methods has mostly focused on non-interactional tasks (Hari & Kujala, 2009). Further to this, if an analysis of the speech produced during an EEG session is to also be evaluated, it is important that the latency between the collection of EEG and speech data is as low as possible. This is to ensure that the data collected from the EEG caps can be accurately correlated with the ongoing speech signals. When it is then considered that the EEG data that is being collected is being generated from two different sources and that the speech data is also being generated comes from two different sources, accurate integration of the separate data sources becomes essential. For the synchronisation of EEG signals, this is relatively simple. The number of channels that EEG systems tend to have are multiples of 32 (ie. 32, 64, 128 etc.), this means that the amplifiers that are attached to the EEG channels must accept a number of channels that are a multiple of 32. Thus, the solution to synchronisation of EEG signals between two interactants is to feed a 64 channel amplifier two 32 channel EEG signals. The resulting output can then be treated as a standard 64 channel output and separated into two 32 channel signals at the processing stage. This is the method employed by Dumas et al. (2010). With the synchronisation of EEG signals solved, the remaining task is to synchronise the two speech signals with the two EEG signals; for this it is possible to once again turn

to EEGLAB.

When performing EEG experiments that take place during physical activity, a number of different streams of data must be processed. This includes the EEG data, data regarding the physical location of participants in space, the video stream of participant movement, data regarding event markers and also auditory data, if used. A plug-in for EEGLAB called MoBILAB (Ojeda et al., 2014) that aims to make data acquisition across a number of modalities more accessible. In order to do this, there needs to be an effective method for data source synchronisation. The path that MoBILAB takes to synchronize different data sources is to make use of an application called Lab Streaming Layer (LSL, Delorme et al., 2011). LSL allows for the accurate time synchronisation of multiple streams of time-series data sources. It achieves this through the use of computer networking such that each stream of data is allocated on a separate port within the same network. Through the use of LSL along with the EEG synchronisation method outlined above, it is possible to conduct a hyperscanning experiment that makes use of time synchronised EEG and speech data from two interacting participants. By using the same amplifier for both EEG data streams from participants in conjunction with LSL for integration of the audio data, this answers the final point of how to perform an EEG hyperscanning experiment whilst recording speech.

### 2.4.3 Summary

In order to sensibly interpret data drawn from the speech and neural signals of interacting participants, standard methodologies will not suffice. The use of approaches from machine learning may provide an alternative. Standard approaches in the assessment of both speech and neural data rely on averaging across many trials or examples in order to uncover small effects that might be attributable to behavioural stimuli. However, when investigating phenomena that unfold in real-time in response to an interactional partner, producing tightly controlled samples over which to produce an average is not generally possible. The use of machine learning approaches, such as HMMs, allow for the characterisation of the general properties of a signal through modelling the signal input as a series of joint probabilities. This allows for the uncovering of specific and nuanced patterns that can be utilised in further comparative analyses, thus eliminating the need for averages drawn from controlled stimuli.

In this section, a potential approach for the application of HMMs to the detection of accommodation in speech has been proposed, this was outlined in subsection 2.4.1. It is suggested that the speech signal be modelled as a sequence of shifting probabilities that any given sample was produced by either speaker in an interaction. By evaluating the speech signal in this way, it is possible to develop a view of accommodation that considers the speech produced by speakers in a more holistic

manner whilst also accounting for the temporal relationships in the speech signal.

Further to the application of HMMs to the speech signal, subsection 2.4.2 suggests that a similar approach can be taken for the evaluation of the EEG signal during an interaction. However, in order for this to be feasible, the ongoing brain activity of two participants would need to be monitored. This necessitates the use of a hyperscanning methodology, where two interacting participants have their brain data collected simultaneously. Considerations for some of the technical hurdles that could be experienced by this type of experiment were presented and potential solutions have been suggested.

## 2.5 General summary

This chapter has aimed to provide the necessary pre-requisite background information for the studies presented in this thesis. The core aim of the chapter was to answer a number of 'What, How and Why' questions pertaining to the research goals of this thesis. Those questions were:

- *What* is accommodation?
- *How* is accommodation measured?
- *Why* should accommodation be linked to joint brain activity?
- *How* can accommodation and brain activity be measured in tandem?

The answers to each of these are briefly summarised in this final subsection before moving on to the presentation of the experiments in chapters 3 and 4.

Section 2.1 provided the content to answer the first question, 'What is accommodation?'. It opened with a brief overview and discussion of the nature of accommodation, providing a definition for its use in this thesis. It then offered a recounting of the core theories (SAT and CAT) that were constructed to account for accommodative behaviour and how they helped to progress research in this area. Following this, the relationship between accommodation and a number of social factors was explored. These factors included speaker sex, dialect, familiarity and dominance. Whilst additional social factors may also play a role in accommodation, the work presented on the above factors demonstrated that the relationship between accommodation and social factors is complex. In order to clearly evaluate these effects, the methodologies used to investigate this phenomenon needed to be explored.

Section 2.2 offered a consideration of the many ways in which accommodation has been measured, dealing with the second core question 'How is accommodation measured?'. In order to do this, it broke down the types of method used to evaluate accommodation into four different categories: 'perceptual interaction approaches','perceptual non-interaction approaches','acoustic-phonetic non-interaction approaches' and 'acoustic-phonetic interaction approaches'. Before making an assessment of the work in each of these categories it provided definitions for what each of the terms used in the construction of the category names meant and how that related to the techniques used to assess accommodation. This section concludes that whilst each of the methods used to assess accommodation has its merits, there are also a number of drawbacks associated with each method, all of which are summarised in table 2.2. It suggests that for accommodation to emerge, there must be an interaction between speakers and that accommodation makes use of the full repertoire of phonetic features available to a speaker. However, it notes that the relationship between accommodation and the speech material being accommodated towards is likely to be a complex and non-linear.

Section 2.3 explored the cognitive mechanisms that are thought to underpin accommodation, introduces the concept of neural entrainment and draws a link between accommodation and joint brain activity. The core cognitive mechanisms that are explored in this section are the motor theory of speech (Liberman & Mattingly, 1985), the direct-realist approach (Fowler, 1986), episodic memory for phonetic detail (Goldinger, 1998) and the mechanistic approach to language use in dialogue (Pickering & Garrod, 2004; Pickering & Garrod, 2013). The key factor running through each of these theories was that the speech production and perception systems must be linked in some way. This linking of perception and production allows for the postulation of a pathway for accommodation to be automatically produced since perception could theoretically have a direct impact on production. Following this discussion of cognitive mechanisms, the notion of neural entrainment to environmental stimuli was introduced. It was explained that the brain has shown evidence of altering its activity in order to align with rhythmic signals in the environment, allowing it to track ongoing events and to predict upcoming events (Henry & Obleser, 2012; Peelle & Davis, 2012). A discussion then followed regarding the potential links between neural entrainment and speech, concluding that whilst the specifics of the role that neural entrainment plays in speech processing may still be unclear, it may be generally concluded that neural entrainment does play some role in speech processing. The final part of this section aimed to suggest a possible role that accommodation might play in the improvement of speech processing during an interaction through the joint entrainment of speakers' brain activity to the speech signal. As a whole, this section provides an answer to why accommodation could be linked to joint brain activity. The section concludes by noting that whilst a good amount of work has been completed, there are still a number of issues that need to be addressed if a link between accommodation and neural activity is to be assessed.

Section 2.4 discusses the issues surrounding the devising of a measure that is able to detect accommodation from continuous interactive speech and how to implement an experiment that tracks both speech accommodation and brain activity in tandem. The first part of this section suggests that a potential solution to the measurement of accommodation in continuous interactive speech may lie in the field of machine learning. A brief explanation of machine learning is provided before focusing on a specific type of machine learning, HMMs. The reasoning behind choosing HMMs as a method for assessing accommodation lie in their long-term employment in speech and speaker recognition systems which are used to process naturally occurring speech. The concepts underlying HMMs are then presented using some examples, this was then followed by a mathematical definition of their usage. The section then turns to consider the application of HMMs to brain data and the potential hurdles that an experiment attempting to measure accommodation and brain activity in tandem might face. Three key points for consideration were presented which cover applying HMMs to EEG data, the elimination of move-

ment artefacts from EEG data during speech and how to record EEG and speech data from two speakers simultaneously, known as hyperscanning. Taken together, the use of machine learning and hyperscanning were suggested to be a potential answer to the question of how accommodation and brain activity can be measured in tandem.

What this thesis proposes is that accommodation should be seen as an interactive process between speakers and as such, it should be measured in an interactive setting. Further to this, since speakers are able to use the full remit of phonetic features to accommodate, it is suggested that a more holistic view of accommodation be taken so that this multi-feature usage can be accounted for. In addition, it is proposed that accommodation might play a role in improving speech processing during a joint activity, such as a speech interaction, by way of inter-speaker neural entrainment. It is acknowledged that there are a number of issues surrounding measurement and testing of these proposals and a machine learning, HMM based approach is offered as a potential solution.

# Chapter 3

# Behavioural Experiment

The function of this chapter is to build on what has been presented in chapters 1 and 2 by addressing the first of the two main tasks that this thesis aims to perform. This was outlined in chapter 1 as the creation of an appropriate measure for accommodation in speech. Chapter 2 presented the literature surrounding the current approaches for the detection of accommodation and highlighted the main drawbacks associated with them. It also presented the potential advantages that the inclusion of HMMs offer for the development of a measure for assessing accommodation that is holistic but remains rooted in the acoustic signal.

Here, an assessment of accommodation is offered first using a traditional phonetic approach and then with the application of HMMs. This is done by performing an experiment that elicits speech from a dyadic interaction focused on a collaborative task. The two approaches presented here address two sub-questions related to the overall goal of creating a continuous measure of accommodation.

The first approach presented here is based on an adaptation of current analysis techniques from the field of phonetics. The second approach integrates an HMM-based approach to the same problem. The key research questions, in relation to the previously discussed literature, are as follows:

- Can a standard phonetic analysis approach be used to detect accommodation across a continuous interaction?
- Can an HMM-based analysis approach be used to detect accommodation in a continuous interaction?

Asking if traditional phonetic measures can be used to characterise accommodation in a continuous interaction allows for a demonstration of a traditional phonetic approach to the assessment of these kinds of phenomena. The applicability of an HMM-based approach can then be contrasted against this, allowing for a comparison to be made.

As is outlined in Chapter 2, previous analyses of accommodation have tended to focus on evaluating the difference between a speaker's speech prior to an interaction and their speech after an interaction (eg. Pardo et al., 2016b). These ap-

proaches allow for an assessment to be made regarding the overall degree to which that speaker's speech shifts towards or away from their speech partner. However, they measure accommodation as a global phenomenon or an end product and tend to treat the actual ongoing process of accommodation as an unknown. That is to say that, for these types of studies, whilst accommodation can be said have occurred, investigation of the process itself is not possible. This behavioural experiment was designed to assess if an HMM-based approach can provide an alternative tool for the investigation of accommodation based on a classification from the speech signal as it unfolds.

Whilst the main goals for this experiment are to ascertain if traditional and HMM-based approaches are able to detect accommodation from the ongoing speech signal, an additional goal is to evaluate a potential experimental paradigm aimed at increasing the potential for accommodative behaviour.

Although the phenomenon of accommodation is robust, having been observed in many settings (eg. Babel, 2009a; Casasanto et al., 2010; Bailly, Lelong, et al., 2010) and across a variety of time-scales (Pardo, 2006; Pardo et al., 2012; Sonderegger, 2015), it can be somewhat subtle and automatic (Aguilar et al., 2015), whilst also being impacted by a large number of social factors (Babel, 2009b; Schweitzer & Lewandowski, 2014). Due to the range and variety of social factors that can modulate the presence of accommodation and its inherently subtle nature, controlling for the impact of social factors was necessary. This is implemented here through a 'self-selection' protocol that was designed to provide optimal conditions for accommodation to emerge. The goal of introducing this was to aid in providing a clear signal for detection and ultimately for relation to the neural signal in the next experiment.

The following section provides an overview of the methodology developed to address the first key task of the thesis and the methodology used to mitigate the impact of social factors. It also presents the results of the experiment along with a discussion of the findings and their implications for the subsequent EEG experiment (detailed in Chapter 4).

## 3.1 Participant Recruitment and Selection

Given the clear impact that social factors have on accommodation, a number of steps were taken to constrain their influence. This was done in two ways, the first was to target the recruitment of participants to only include a specific subset of the population. The second was by applying a 'self-selection' protocol that was designed to group the participants such that the personality types of each dyad were similar and the interpersonal attraction within each dyad was high. Both of these are detailed in this section.

### 3.1.1 Sex and dialect

Based on literature concerning the role of sex on language change (Trudgill, 1972) and the impact that dialect differences can have on speech production (Wells, 1982; Foulkes & Docherty, 1999), restrictions were placed in the types of participant recruited. These restrictions meant that only female participants from the city of Glasgow conurbation were recruited to take part in the experiment.

Whilst the literature still demonstrates some uncertainty about the exact role of sex in accommodation (Pardo et al., 2016b) a case still remains for the use of single sex, female only dyads. Women have long demonstrated a tendency to be at the forefront of linguistic change (Trudgill, 1972; Milroy & Milroy, 1985) and also demonstrate a greater attunement to social cues (Milroy & Milroy, 1985). Given that we know social factors play a role in accommodative behaviours, it would seem reasonable to conclude that females would show a greater linguistic diversity as a result of greater sensitivity to social cues. Additionally, omitting male participants removes the effects of physiological factors associated with sex. Coupled with the evidence that suggests that females are more likely to produce accommodative behaviours (eg. Namy et al., 2002), the use of female only dyads presents a sensible route for the maximisation of accommodative behaviours.

Further to controlling for sex, recruitment of participants was also restricted to those having been born and raised in the city of Glasgow conurbation. Because dialect also plays a role in accommodation (Bigham, 2010; Campbell-Kibler, Walker, Elward, & Carmichael, 2014), the potential for large variations in the dialects of the participants was restricted. However, over-restriction could lead to an over-constriction of the potential accommodative space for the speakers. Since Glaswegian has been shown to be linguistically diverse (Macaulay, 1976; Stuart-Smith, 1999; MacFarlane & Stuart-Smith, 2012; Lawson, Scobbie, & Stuart-Smith, 2013; Stuart-Smith, Rathcke, Sonderegger, & Macdonald, 2015) but with broad uniformity (Macafee, 1983), recruitment from only the city of Glasgow conurbation allows for adequate restriction of dialect to avoid large variations whilst also providing a large enough accommodative space for accommodation to take place relatively unhindered.

### 3.1.2 Self-selection protocol

The goal of the self-selection protocol was to group the participants into dyads with similar personality traits and with a high degree of interpersonal attraction. By grouping participants into dyads with similar personalities and that are more engaged with one another through greater interpersonal attraction, it was theorised that they would express a greater degree of accommodative behaviour. The reasoning behind this stemmed from the literature detailing the social factors that have an impact on accommodation (eg. Babel, 2012; Yu et al., 2013; Beňuš, 2014, , see also section 2.1.2). Combining this literature with that surrounding the tendency of desired personality to be linked to face preference in a socially organised way (Little, Burt, & Perrett, 2006; Bronstad & Russell, 2007) allowed for the development of the self-selection protocol.

The protocol itself required participants to 'self-select' their conversational partner for the experiment based on photographs of the other participants in the study. Each participant was presented with a matrix of 20 photographs (all of the initially recruited participants, including themselves). These photograph matrices were randomly generated such that no two matrices had the participants presented in the same order. This was done to eliminate order effects, it might have been the case that participants consistently focused on the top-left of the matrix for instance, if this had contained the same picture it would have led to an over representation of that one participant. The photographs were then overlaid with an alphanumeric grid numbering system, located in the top left of each individual photograph within the matrix. The grid numbering system saw the rows assigned letters ('A', 'B', 'C' and 'D') and the columns assigned numeric values ('1', '2', '3', '4' and '5'). So, 'A5' was found in the top right hand corner of the matrix and 'D1' in the bottom left of the matrix. Participants were sent one of the randomly generated matrices along with instructions to do the following:

> Open the attached image, view it in full screen (if possible) and select 5 people that you would like to complete the final portion of the experiment with. Then, make a note of the associated grid co-ordinates and e-mail them back to me.
>
> So, a response might be something like this:
>
> A3 - B2 - B4 - D1 - C2
>
> Where A3 is the most preferred partner, B2 is the second most preferred and so on...

These responses were then coded from first choice (1, most preferred) to fifth choice (5, least preferred).

Once all of the pairing responses had been received, they served as input for an in-house MATLAB® script which identified all the participants who mutually selected each other. The script then paired the participants in a progressive fashion

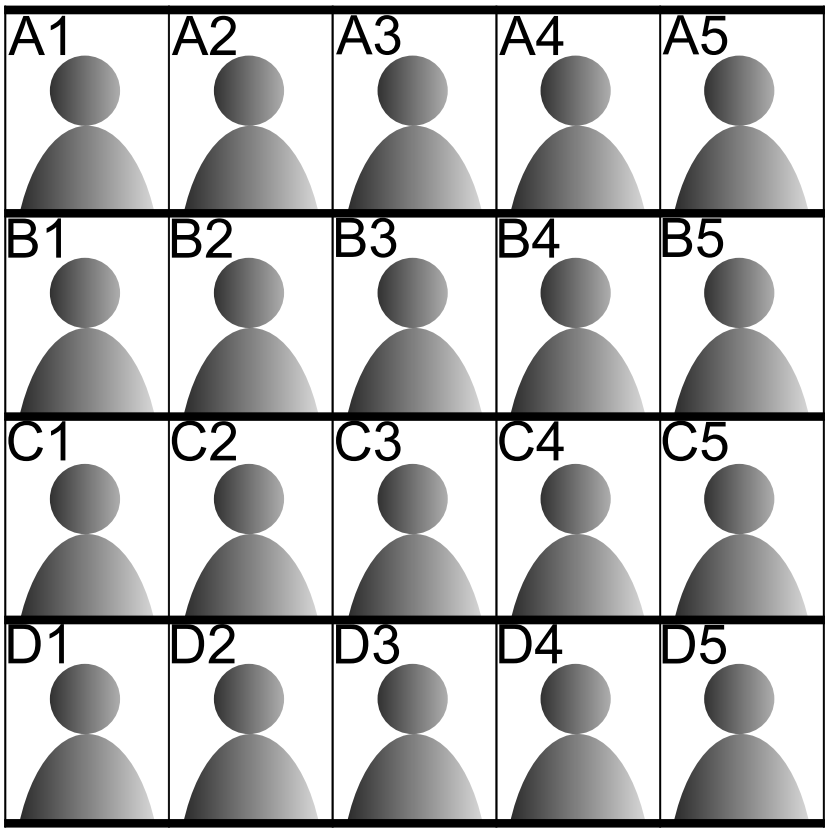



Figure 3.1: Example of 'Self-Selection' Photograph Matrix.

based on the rankings that the participants had assigned each other. Whenever it was identified that two participants had selected each other, the rankings that they allocated one another were taken into consideration and averaged. If a participant was selected by more than one other participant, they would be allocated to the pair where each member had allocated the highest ranking. For example, if participant 'X' and participant 'Y' selected each other where 'X' rated 'Y' as a 3 and 'Y' rated 'X' as a 1 then the average score for that pair would be 2. Now, let's assume that participant 'Z' and participant 'X' also selected each other, where 'X' rated 'Z' as a 1 and 'Z' rated 'X' as a 2 then the mean for that pair would be 1.5. In this instance 'X' would have been paired with 'Z' rather than 'Y' as they rated each other more highly (recall that in this instance $1 > 5$) and that is theorised to lead to greater phonetic accommodation. 'Y' would be allocated to a different pair. The rankings were only seen by the experimenter and participants were never told what rating they received from their conversational partner.

### 3.1.3 Participants

20 female participants were initially recruited to take part in the study. However, the self-reporting required for the self-selection protocol (see section 3.1.2) coupled with the experiment requiring two sessions led to an attrition rate of 40%. Ultimately, the study consisted of 12 participants in total.

The attrition rate also had an impact on the outcome of the self-selection proto-

col. Since the protocol was reliant on all of the participants returning their ratings for who they wanted to perform the experiment with, losing 40% of the participants meant that the number of people that could potentially select each other was reduced. It emerged that of the 12 participants who remained in the experiment only 6 selected each other in the rankings. This meant that the remaining 6 participants had to be randomly paired with each other. As a result, these pairs could effectively be used as a control group to test the ability of the self-selection protocol to group by personality and interpersonal attraction. The participants can therefore be considered as belonging to one of two groups, 'self-selected' or 'randomly paired'.

The participants' ages ranged from 19 to 65 (mean 30.92 years, standard deviation 14.38 years). Since accommodation is an inherent part of communication in general, there was no reason to apply an upper age limit, providing that hearing and sight were normal. Participants were all recruited from the city of Glasgow conurbation. All participants were native speakers of the Scottish English dialect. All participants were screened to ensure normal hearing and normal or corrected to normal eyesight. Participants were compensated with a monetary payment of £6 per hour. All participants were assigned a five character 'participant code' which was used in place of their names in order to maintain their anonymity.

### Demographic information

Demographic information about the participants that took part in the experiment can be found in Table 3.1. The participants' demographic information contains their age, name of their home town and the location of their home town in relation to Glasgow city centre. The participants' pair numbers, self-selection status (if they self-selected or were randomly paired) and participant codes are also presented. Members from each of the major areas within Glasgow were present in the study although there was a bias in turn out towards those who grew up in the south of the city ($n = 6$).

| Pair Number | Self Selected | Age | Home Town | Location | Participant Code |
|---|---|---|---|---|---|
| Pair 1 | Yes | 21 | Stamperland | South | GJN14 |
| | | 19 | Busby | South | HLH30 |
| Pair 2 | No | 65 | Giffnock | South | JSE11 |
| | | 36 | Ibrox | South | TMY30 |
| Pair 3 | Yes | 22 | Bearsden | North-West | JTN20 |
| | | 19 | Langside | South | ARA14 |
| Pair 4 | No | 50 | Hillhead | West | JYN22 |
| | | 36 | Pollock | South | ZSE07 |
| Pair 5 | Yes | 28 | Bearsden | North-West | SCA01 |
| | | 34 | Greenock | North-West | KBN30 |
| Pair 6 | No | 21 | Blantyre | South-East | SKN03 |
| | | 20 | Coatbridge | East | SHA13 |

Table 3.1: Demographic and Pairing Information. If the participants self-selected each other as a result of the self-selection protocol then the 'Self Selected' column will contain a 'Yes', otherwise they were randomly paired and the column will contain a 'No'. The location of the participants' hometowns is in relation to the city centre of Glasgow.

## 3.2 Task Materials

This section lists the task materials used in the experiment. It lays out the three tasks that the participants were asked to complete during the experiment, namely the DiapixUK task, the Big Five Personality Inventory and McCroskey's Interpersonal attraction questionnaire.

### 3.2.1 DiapixUK task

The DiapixUk task was selected in order to elicit free flowing conversational speech during a collaborative task, whilst also controlling for the effect of speaker role (eg. Giver vs. Receiver in the map task) which has been shown to impact levels of accommodation (Pardo et al., 2010; Krauss & Pardo, 2006). Initially developed at North-Western University, the Diapix task was refined for use in the UK at UCL by Baker et al., 2011. It consists of a set of twelve images that are separated into three 'scenes' (a farm scene, a street scene and a beach scene), it therefore has four images in each scene category. Each of these images has a counterpart that is exactly the same apart from twelve slight differences (see: Figure 3.2). These differences were engineered to elicit the use of a number of keywords which were specific to the scene that was presented to the participants. The keywords selected, formed minimal pairs with each other and allow for the assessment of changes in the phonetic features of these words whilst offering a form of control for the surrounding phonetic context.

The task consisted of presenting two participants, who could not see each other, with images of the same scene where one participant saw the original and the other saw the counterpart. The task itself was simply for the participants to communicate through verbal interaction only, in order to find all twelve differences between the images. Baker et al. (2011) performed extensive testing of the DiapixUK task and designed the task such that the images are balanced for difficulty and assessed for adequate speech production material. They have also assessed the level of speech contribution from both speakers, learning effects and keyword production frequency. They found no significant difference in difficulty between the images, elicitation of sufficient speech material to perform acoustic-phonetic and linguistic analyses, an equal speech contribution from each speaker and no significant learning effect from performing the task on multiple images. Although the authors found that keyword repetition was the least robust feature of the task it consistently produced enough speech material to perform analyses of the keywords as well as longer speech stretches.


Figure 3.2: An example of the DiapixUK stimuli. The above image pair is from the 'Beach' scene category.

### 3.2.2 Big Five personality inventory

Prior to being paired, participants were asked to complete the Big Five personality Inventory (BFI), developed at the University of Berkeley Personality Lab (Benet-Martínez & John, 1998; John, Donahue, & Kentle, 1991; John, Naumann, & Soto, 2008). As discussed in subsection 3.1.2, it was theorised that those pairs with more similar personality traits would be more likely to converge to one another. The self-selection protocol was used as a way to group participants into pairs with a high degree of similarity in their personalities. The BFI was used to ascertain whether the self-selection protocol was able to pair participants with similar personality traits together.

The BFI interprets personality as a construct of five broad personality dimensions that are defined as follows (adapted from John et al., 2008, pp.120, Table 4.2):

**Extraversion**

Implies an energetic approach towards the social and material world. Traits include sociability, assertiveness and positive emotionality. Those scoring highly for extraversion may be more likely to engage in behaviours such as introducing oneself to a stranger at a party or organizing a team project. Those with low scores for extraversion may be less likely to voluntarily engage with others at a party or may stay quiet when disagreeing with others.

**Agreeableness**

Observes a prosocial and communal orientation. Traits include altruism, trust and modesty. High scores for agreeableness are associated with emphasizing good qualities in others and consoling those in pain. Those with low scores for agreeableness may demonstrate antagonistic tendencies such as focusing on the poorer qualities of others and a lack of empathy with those in pain.

**Conscientiousness**

Describes a form of impulse control that facilitates task and goal directed behaviour. Traits include following norms/rules and planning/organising tasks. Those high in conscientiousness may tend to arrive early or on time for appointments and double check essays for spelling errors. Those scoring on the lower end of the scale may be less inclined to pro-actively engage in tasks.

**Neuroticism**

Contrasts emotional stability with negative emotionality. Neurotic traits include feeling anxious, nervous, sad and tense. High levels of neuroticism may lead to an individual becoming unduly upset when someone is angry with them. Those low in neuroticism have more of a tendency to accept the good with the bad without complaining or bragging.

**Openness**

Describes the breadth and depth of one's mental and experiential life. Traits include being open to new experiences, trying new things. Someone high in openness may look for stimulating activities to break up their routine or learn something simply for the joy of learning. Low scores in openness may indicate less willingness to engage in new activities and a preference for a set routine.

The BFI has been used to asses the personality of participants for a considerable amount of time and it has undergone a wide variety of tests of validity and reliability (Rammstedt & John, 2007). The inventory itself consists of a series of 44 statements, which assess the participant on five broad personality dimensions: Extraversion, Agreeableness, Conscientiousness, Neuroticism and Openness. The task is to respond to the statements on a five point Likert scale, where 1 = 'Strongly

Disagree', 2 = 'Disagree', 3 = 'Neither Agree or Disagree', 4 = 'Agree' and 5 = 'Strongly Agree'. Each statement is associated with a particular personality dimension and in some instances the statement is worded such that a high response will translate into a low score for that dimension in order to minimise the chances of participants identifying which questions assess which personality dimensions. The statements themselves can be worded such that they refer to a particular individual (eg. a celebrity or acquaintance) or to the person taking the test. The responses are then collected, coded (including reverse coding for some items) and summed based on the dimension that they are associated with. The mean of the question responses that grouped with a particular personality dimension provide the overall rating for the dimension in question.

### 3.2.3 McCroskey interpersonal attraction questionnaire

Whilst the self-selection protocol was designed in order to capture the degree to which similarity influences phonetic accommodation, it may not be an appropriate manner in which to assess this due to liking/attraction impacting on the participants' selections. The inclusion of the McCroskey interpersonal attraction questionnaire (IA) was used to try and isolate the degree to which the participants' selections of their potential conversational partner was influenced by how much they were attracted to the image rather than considering the person depicted to be similar.

McCroskey and McCain (1974) originally developed the interpersonal attraction questionnaire from factor analytic work that they performed on a group of two hundred and fifteen participants who were asked to respond to thirty Likert type statements about an acquaintance. They found that the factor analysis demonstrated three key dimensions of interpersonal attraction which they termed, 'Social Attraction','Physical Attraction' and 'Task Attraction'.

**Social Attraction**

A socio-emotional aspect, closely related to what one might ordinarily call 'liking'. Can be considered as a social or personal liking property.

**Physical Attraction**

A physical dimension based on dress and physical features. Can be considered as a measure of physical or material attraction.

**Task Attraction**

A category of interpersonal attraction related to what one might ordinarily call 'respect'. Can be considered as a task-orientation dimension related to how easy or worthwhile working with someone would be.

It was also shown that the test could be reduced to a fifteen item questionnaire. Since then, work has continued to refine the questionnaire which has seen it grow

to an eighteen item task although the application of the task remains the same (McCroskey, McCroskey, & Richmond, 2006). Participants are asked to complete a five point Likert scale in response to the presented statements, much like the method described in section 3.2.2. This is done in relation to another person, a friend, an acquaintance, a celebrity or in this instance, a conversational partner. The score range for each of the three interpersonal attraction dimensions is 5 to 35, with 5 representing low interpersonal attraction and 35 representing high interpersonal attraction.

It should be noted that the way in which this questionnaire is applied here is non-standard. In a standard setting the IA would be undertaken by a number of individuals rating the same person. For instance, in Weiss and Houser (2007) they had a number of students complete the IA in relation to each of their tutors. As such, they had multiple reports for each of the IA dimensions that came from different people. Due to the design of this study, it would not have been possible to gather multiple completions of the IA for each participant. As such, each member of each participant pair generated one report for each of the IA dimensions in relation to their experimental partner. By applying the IA in the way that it has been, the ability to make assessments about the interpersonal attraction towards any given individual is effectively lost. However, statements about general levels of interpersonal attraction across the whole experimental group and between pairing conditions can still be made.

## 3.3 Procedure

### 3.3.1 Pre-screening

A pre-screening stage saw the BFI administered to participants who were sat behind a screen and were out of the sight of the experimenter. The participants were presented with statements and were asked to evaluate how much each of these statements applied to themselves. This was presented to the participants on a computer screen using a MATLAB® script which first presented instructions concerning how to complete the task to the participants. It then presented the stimuli on a plain black background with the text presented in white and recorded their responses via button press. The required response was to indicate how much they agreed with the statement on a five point Likert scale. There was a legend at the bottom of every presented screen which reminded the participants of which button corresponded to which response (see: Figure 3.3 for an example screen-shot). Once the BFI-44 was completed the program processed the responses based on the method provided by the Berkley personality lab. This gave the final score for each personality trait for each participant. Following completion of the BFI-44 the participants were re-


Figure 3.3: An example of the stimuli used for the BFI.

quired to have a photograph taken of their faces. Care was taken to standardise the photographs as much as possible without interfering with the manner in which the participants wished to present themselves. As such, all photographs were taken in the same place, with the same distance between participant and camera with a plain white background. Once all 20 participants had completed this section of the experiment, a five-by-four photo matrix of their images was created, e-mailed out

to them and they were required to select their conversational partner for the next section of the experiment, as outlined in section 3.1.2.

### 3.3.2 Recording

Participants were asked to come back to complete the experiment at a time which was convenient for both members of their conversational pair. At no point before meeting did the participants know the name of their conversational partner; their participant codes were used at all times. They were brought to the recording studio and were briefed about the nature of the task they were to undertake. They were then invited into a sound attenuated booth to begin the experiment. They were sat opposite each other but in different corners of the booth, with a divider between them such that they could not see each other but could still hear one another. A graphical representation of the physical set-up is presented in figure 3.4 Each participant had an AKG SE 300B pre-amp equipped with an AKG CK91 condenser capsule serving as a mono microphone. These are designed to suppress off-axis sound and served to minimise the amount of speech captured from the conversational partner. They were positioned 20cm away from the participants' mouths. These microphones were fed into two separate channels which were then combined by the mixing desk into a stereo signal with the left channel assigned to one speaker and the right channel assigned to the other. In this way, we captured an audio signal which could be easily separated into the two individual speakers whilst maintaining a low latency time-lock between the utterances of the speakers. The audio was recorded at a sampling rate of $44\,100$ Hz.

Participants were seated approximately 30cm away from a flat screen computer monitor which was adjusted to be at their respective eye levels. The signal to both monitors was provided by the same computer but because the images presented were unique the signal could not simply be split. Instead the secondary DVI output port on the computer was used. This led to a slight delay of $\sim$100ms between stimulus onset on the two screens. The stimuli were presented using the Psych-Toolbox in an in-house produced MATLAB® presentation script. The participants were instructed that they did not need to provide any responses (eg. via button press), just to complete the task as outlined in section 3.2.1. It was made clear that once they had found all of the differences between the images or after fifteen minutes had passed (whichever came first), the screen would auto advance. In fact, the screen was advanced by the experimenter who was listening to the conversation and recording the differences that were being found. The participants were not aware that the experimenter was listening to their conversation; a cover story was offered about testing a speech recognition software in order to reduce any observer effects. The participants also performed a short trial run of the task which asked them to find just three differences between a pair of highly reduced stimuli images before

the main element of the experiment was run, to ensure they understood the task.

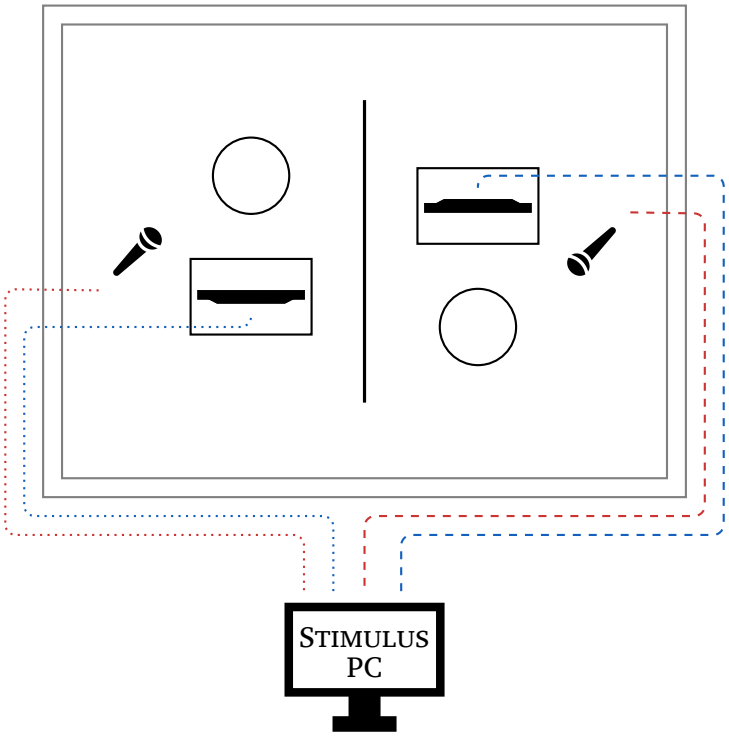



Figure 3.4: Diagram of physical experimental set-up. Circles represent the participants, dotted lines indicate the input and output connections for the participant on the left, dashed lines are used in the same way for the speaker on the right. Red lines indicate incoming data from the participants to the computer and blue lines indicate outgoing information (stimuli) from the computer to the participants.

The stimuli were presented in runs of three image sets with four blocks overall (not including the trial run). Participants were offered breaks in between each block. Each run consisted of one of each of the scene categories (ie. each run had a 'beach scene', a 'street scene' and a 'farm scene') so that two of the same scene category never appeared in succession. This was done to minimise the chance of participants spotting the differences associated with each scene as the differences were tied to the keywords which accompanied each scene. The order in which these scenes were presented within each block was randomised to minimise learning and order effects. To make conversations across pairs more comparable, the participants were instructed to begin in the top left corner of the scene and to proceed to look for the differences in a clockwise manner around the scene. Both participants were encouraged to contribute to finding the differences in order to make it less likely for one speaker to dominate the conversation.

Following the completion of the DiapixUK task participants were separated and asked to complete the McCroskey interpersonal attraction questionnaire, as outlined in section 3.2.3, about their conversational partner. This was done away from their conversational partner and they were not informed of the purpose of the questionnaire until they were adequately separated and could not hear one another.

### 3.3.3 Transcription and data management

An orthographic transcription was conducted on the collected data, primarily by the author but with additional help from an employed transcriber. All transcription was conducted in PRAAT (Boersma & Weenink, 2016) with separate transcriptions for each member of the speaker pair. The transcription conventions that were followed closely resembled the conventions that were outlined for a larger project within the department (See Appendix A) with a few minor changes.

- Where a speaker has contracted a word (eg. "the man is" becomes "the man's") this was written in the orthographically contracted form rather than in the full form. Due to the spontaneous nature of the speech collected, it was important to retain a transcription which closely matched what was actually said in order to avoid potential mis-matches when phonetically aligned with the forced aligner used by an application called LaBB-CAT. LaBB-CAT is a repository for time-aligned transcripts of speech data that allows for easy databasing and extraction of acoustic features.

- For compound words such as "sandcastle", "beehive" etc. they were written as two separate words ie. "sand castle", "bee hive" etc. The primary reason for this was so that all of the keywords could be found without having to also search the transcription for them in their compound forms.

- If someone begins a word but does not complete it, it was written with a tilde to indicate that it was cut off (eg. "messa∼") . This also applied to words which were finished but not started (eg. "∼ssage").

- Capital letters were only used for proper nouns and acronyms (eg. If a speaker says "at 3PM" for post meridian or afternoon). This was done purely for identification purposes when reading the transcript.

- For descriptions of colours and novel words, the following transcription convention was used "red-y", "orange-y", Simply adding a dash then the letter "y". There was a surprisingly high instance of descriptive words using the /i/ ending to signify that an object was in that semantic region but could not be described exactly by the word that the /i/ had been suffixed to.

In addition to these minor changes to the transcription protocol, there were also some specific requirements of the PRAAT tiers used to transcribe the speech data. There were four tiers in total, with the first tier containing the orthographic transcription. The second tier contained a note of the type of scene (ie. Beach, Farm, Street) that the transcription belonged to. This was done in order to be able to identify and make comparisons between the responses to different scenes. The third tier contained a note on the position that the particular stimulus was presented in the

experiment (ie. 1st, 2nd, 3rd etc.). This allows us to investigate the degree of phonetic accommodation across the course of the experiment as a whole. Finally, the fourth tier contained markers which identified the points at which the participants discovered each difference. This tier allows the assessment of whether phonetic accommodation increases as the participants converge on a task goal.

The data, along with the transcriptions were uploaded to a networked server running LaBB-CAT. Each individual speaker was the primary speaker in the transcription associated with their audio file. After a full orthographic transcription had been completed and uploaded to LaBB-CAT, the transcriptions were checked for errors and a forced alignment of the phonetic features was conducted. The forced alignment was completed with the utility built into LaBB-CAT which utilizes HTK to complete this task. Because each of the participants was recorded on a separate channel and was uploaded with individual sound files and transcriptions, any noise due to overlapping speech was minimised.

## 3.4 Evaluating the self-selection protocol

The goal of this section is to determine if the self-selection protocol was effective in producing a group of people with similar personality types and with a high degree of interpersonal attraction. By having a grouping of people with similar personalities and that are more engaged with one another through greater interpersonal attraction, it was theorised that they would express a greater degree of accommodative behaviour. The reasoning behind this stemmed from the wealth of literature detailing the factors that have an impact on accommodation (see sections 2.1 and 2.2). The literature suggests that accommodation is a subtle phenomenon that can be modulated by a good number of different and interacting factors. This study has controlled for as many of these as possible by selecting participants from a similar geographic region, by restricting the gender of the participants and the method by which this study aimed to control for the impact of social factors such as personality and interpersonal attraction was through the implementation of the self-selection protocol.

This protocol was used as an alternative to pre-screening the participants for personality traits and matching by their responses. Asking participants to self-select based on images of potential partners was a way to get pairings that best represented what the participants implicitly felt resembled someone that was similar to themselves. If participant matching had been based on the results of the BFI or IA then a priori assumptions about the similarities of personalities and attractions would have had to have been made by the experimenter. By asking the participants to self-select their potential partner based on their picture, it was hoped that a priori bias would be minimised.

If the self-selection protocol allows participants to select partners in the intended manner then the results of both the BFI and the IA should demonstrate a higher degree of affinity between partners paired by the self-selection protocol than by those that were randomly paired. The following section aims to evaluate this by assessing the BFI and IA results when separated by selection condition (self-selected or randomly paired). However, no major trends were detected and as such the reporting of the findings is somewhat reduced.

Since the main aim of this section is to determine if the self-selection protocol was able to group participants by personality and by interpersonal attraction, the descriptive statistics for the BFI and the IA will not be reported here. The BFI and the IA will only be considered in relation to the impact that the self-selection protocol had on the personality and interpersonal attraction factors reported by them. However, the descriptive statistics for the BFI and IA data can be found in appendix B.

The remainder of this section is structured as follows. Section 3.4.1 provides the results for the impact of the self-selection protocol on the personality dimensions reported by the BFI. Section 3.4.2 reports the results for the impact of the

self-selection protocol on interpersonal attraction as measured by the IA. Finally, Section 3.4.3 provides an interpretation of the suitability of the self-selection protocol for grouping participants by personality and/or interpersonal attraction. Additionally, it offers some considerations on possible issues with the protocol and suggests some potential improvements.

### 3.4.1 Ability to group by personality

The main aim of this subsection is to ascertain if the self-selection protocol, as outlined in subsection 3.1.2, is able to group participants by personality. This is assessed through the participants' responses to the BFI. Having pairs of participants with similar personalities, as determined by the BFI, would allow for an assessment to be made of the impact that similar personalities vs. randomly paired personalities has on accommodation. If the personalities of the participants in the self-selected pairs are found to be more similar than those that were randomly paired then the self-selection protocol can be said to be successful in pairing by personality.

In this study, the main use of the BFI is to determine if the self-selection protocol was able to group participants based on their personality traits. For this reason, the main interpretations of the BFI will be made in relation to the self-selection protocol rather than providing a commentary on the types of personality trait found in this sample of participants. Having said that, it is still advisable to understand what the results of the BFI represent. Details of the BFI can be found in subsection 3.2.2, but the main details are briefly recapped here.

The BFI provides a measure of five personality dimensions, these are Extraversion, Agreeableness, Conscientiousness, Neuroticism and Openness. The BFI scores each of these dimensions on a scale from $1$ to $5$ through a series of likert scale based questions about the participant. A score of $1$ indicates that the participant demonstrates little of that particular dimension whist a score of $5$ indicates that the participant demonstrates a large amount of that dimension. Each of these dimensions represent distinct personality dimensions and it is therefore meaningless to aggregate them into a 'global' measure of personality. The BFI has been rigorously tested and holds a high degree of validity and reliability.

#### Results

The difference between BFI score, for each dimension, within the pairs serves as a measure of how similar the two participants in a pair are. For instance, if both participants had similar levels of agreeableness, whether high or low, the difference between the BFI scores for that dimension would be $0$ or close to $0$. However, if both participants had dissimilar levels of agreeableness, the difference between the BFI scores for that dimension would be greater. Figure 3.5 presents the mean of these differences within each pairing condition, self-selected or randomly paired. The

BFI dimensions are Extraversion (Extra), Agreeableness (Agree), Conscientiousness (Consc), Neuroticism (Neuro) and Openness (Open).

If the self-selection protocol had been successful it would be expected that the difference between BFI dimensions within pairs would be smaller in the self-selected condition. This appears to be the case in the extraversion, neuroticism and openness dimensions but not for agreeableness or conscientiousness. However, there does not appear to be a general trend for the impact of the pairing condition on the degree of difference in BFI within pairs. In addition, there is a fair amount of overlap between the standard errors for each BFI dimension except for neuroticism. This suggests that any potential differences between the pairing conditions may not be real.

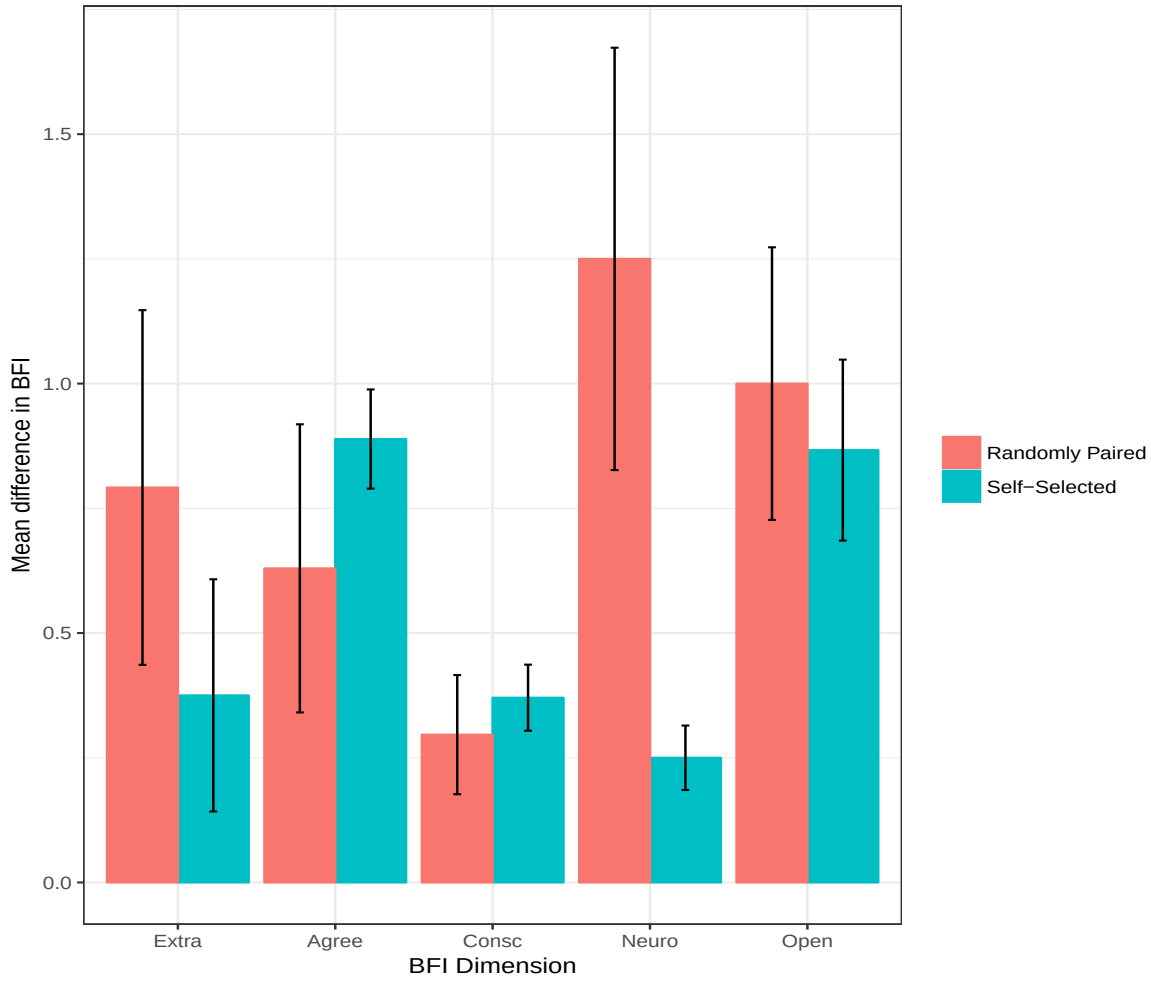



Figure 3.5: Barplot of the difference in BFI scores for each of the BFI dimensions, separated by pairing condition. Error bars represent the standard error.

The raw results of the BFI for each of the participants are presented in table 3.2. Results are presented with the participants grouped into their self-selected and randomly paired conditions. The scores represent any given participant's rating for each BFI dimension. The possible values for each of the personality dimensions range from 1 to 5.

As a general comment about the data, it can be seen that those participants in the self-selected condition tend to be younger than those in the randomly paired con-

| | Participant | Age | Extra. | Agree. | Consc. | Neuro. | Open. |
|---|---|---|---|---|---|---|---|
| Self-Selected | GJN14 | 21 | 2.88 | 4.78 | 3.00 | 2.75 | 3.40 |
| | HLH30 | 19 | 3.75 | 4.11 | 3.44 | 2.50 | 3.90 |
| | JTN20 | 22 | 4.88 | 4.89 | 4.33 | 2.63 | 4.00 |
| | ARA14 | 19 | 4.88 | 3.89 | 4.11 | 2.25 | 3.10 |
| | SCA01 | 28 | 3.38 | 3.89 | 4.11 | 2.38 | 4.40 |
| | KBN30 | 34 | 3.63 | 4.89 | 4.56 | 2.50 | 3.20 |
| Randomly Paired | JSE11 | 65 | 3.25 | 4.00 | 4.44 | 1.50 | 4.40 |
| | TMY30 | 36 | 3.38 | 3.89 | 4.33 | 2.88 | 3.80 |
| | JYN22 | 50 | 3.13 | 3.56 | 4.44 | 4.13 | 4.50 |
| | ZSE07 | 36 | 4.63 | 4.78 | 5.00 | 2.13 | 2.90 |
| | SKN03 | 21 | 3.13 | 4.33 | 3.00 | 2.88 | 3.90 |
| | SHA13 | 20 | 3.88 | 3.78 | 3.22 | 3.25 | 3.10 |

Table 3.2: Results of the Big Five Inventory personality questionnaire. The participants are presented in their respective self-selected or randomly paired conditions. The participants are further separated into their speaker pairs. Columns correspond to the Big Five personality traits: Extra. = Extraversion, Agree. = Agreeableness, Consc. = Conscientiousness, Neuro. = Neuroticism, Open. = Openness. All result are rounded to 2 significant figures.

dition. The mean age for the self-selected condition is $23.8$ years whereas the mean age of the randomly paired condition is $38$ years. Additionally, the differences in the ages between members of each pair tend to be smaller for the self-selected condition. The mean age difference between participants in their pairs for the self-selected condition is $4$ years whereas for the randomly paired condition it is $14.7$ years. This may suggest that those that self-selected their partner may have done so using the apparent age of the potential partners as a factor in their choice.

Taking these findings as a whole, there is little support for the self-selection protocol providing an appropriate method for creating participant pairs that have similar personalities.

### Discussion

The results of the BFI data do not appear to show any appreciable trends. There is little consistency in the self-selected condition having smaller differences between their personality dimension results than the randomly-paired condition. The broad outcome of the results for the self-selection protocol would therefore seem to be that it is not able to group participants by personality.

The fact that there is so much overlap in the standard errors of the conditions for most of the personality dimensions is not surprising given the small sample size ($n = 3$). However, this could not be helped, due to the multi-part nature of the experiment and the difficulties encountered in encouraging participants to return.

However, if a larger sample had been acquired it might have been possible to say more about the relationship between the pairing conditions and personality dimensions. Whilst there may be little overall pattern across the personality dimensions, there may be some personality dimension specific trends that are present but that cannot be captured with the small sample size used here. For instance, the more closely matching levels of neuroticism for the self-selected participants in comparison to the randomly paired participants may have an impact on accommodation. It may be that those with dissimilar levels of neuroticism could find it hard to work together due to differences in their emotional state; where one participant tends towards more positive emotional states and the other toward more negative states (ie. high neuroticism being associated with more negative emotional states: guilt, envy, anxiety etc.). This could, in turn impact on accommodation by influencing the degree of liking between participants. However, from this data it cannot be determined if there is a real difference in this personality dimension or not. Further data is required to make a judgement on that.

Another consideration that must be made regards the age of the participants in each of the pairing conditions. It would seem that by allowing participants to select their partners, those that demonstrated highest affinity for one another tended to be of roughly the same age (see table 3.2). Those that were randomly paired however, tended to show greater differences in their ages. It is difficult to tell if this is as a result of the self-selection protocol itself or is a result of the age range of participants in this particular sample. The fact that the two oldest participants, JSE11 and JYN22, were assigned to the randomly paired group and to the two next oldest participants may just be the result of chance. It could have been the case that the oldest participants were paired with SKN03 and SHA13 from the randomly paired condition, who were both in their 20s. However, assuming that the allocation is the result of the self-selection protocol, the age difference may be having an effect on the BFI results. It may be the case that as one ages, the general tendency for some personality traits may differ from that of younger people. For example, although JSE11 and JYN22 may not quite be at this age yet, it is often said that people can feel more lonely as they age. One way of combating this may be to develop a more extraverted and open personality so as to maximise the possibility of human engagement. If something like this were consistent across an age group, it may impact on results such that effects found were more attributable to age differences rather than pairing condition. However, it is unclear from this data if such an effect is taking place.

In sum, there is no real evidence, given this sample, that there is a relationship between pairing condition and similarity of personality within pairs.

### 3.4.2 Ability to group by interpersonal attraction

The main aim of this subsection is to ascertain if the self-selection protocol, as outlined in subsection 3.1.2, is able to group participants by interpersonal attraction. This is assessed through the participants' responses to the IA. Having pairs of participants with greater interpersonal attraction, as determined by the IA, would allow for an assessment to be made of the impact that greater interpersonal attraction vs. randomly paired interpersonal attraction has on accommodation. If the levels of interpersonal attraction in the participants in the self-selected pairs are found to be greater than those that were randomly paired then the self-selection protocol can be said to be successful in pairing by interpersonal attraction. As with findings for the BFI, little evidence is found to support this claim and as such the reporting is again reduced.

In this study, the main use of the IA is to determine if the self-selection protocol was able to group participants based on their interpersonal attraction. For this reason, the main interpretations of the IA will be made in relation to the self-selection protocol rather than providing a commentary on the scales of interpersonal attraction found in the participants. Having said that, it is still advisable to understand what the results of the IA represent. Details of the IA can be found in subsection 3.2.3, but the main details are briefly recapped here.

The IA provides a measure of three interpersonal attraction dimensions, these are Social attraction, Physical attraction and Task attraction. The IA scores each of these dimensions on a scale from 5 to 35 through a series of likert scale based questions about a specific person, here it is the person that the participant has just completed the task with. A score of 5 indicates that the participant feels little attraction towards their partner in that particular dimension. A score of 35 indicates that the participant demonstrates a large amount of attraction towards their partner in that dimension. Each of these dimensions represent distinct interpersonal attraction dimensions and it is therefore meaningless to aggregate them. The IA has been rigorously tested and holds a high degree of validity and reliability. However, it should be noted that the way in which this questionnaire was applied was non-standard. In a standard setting, the IA would be undertaken by a number of individuals rating the same person. For instance, in Weiss and Houser (2007) a number of students completed the IA in relation to each of their tutors. As such, they had multiple reports for each of the IA dimensions that came from different people. In the study presented in this thesis, it would not have been possible to gather multiple completions of the IA for each participant. By applying the IA in the way that it has been, the ability to make assessments about the interpersonal attraction towards any given individual is effectively lost. However, statements about group trends such as overall interpersonal attraction within pairing conditions can still be made.

### Results

One participant was excluded from analysis because they returned the same response for all of the IA questions. Participant ARA14 had the same value for all IA dimensions, upon inspection of her individual responses it was found that she provided the same response for all questions. Because the questions for the IA are counterbalanced, answering with the same response for all questions is an indication that this participant did not answer truthfully. All other participant responses were also checked to ensure that the responses given were credible. No further issues were found with the participant responses. This exclusion of participant ARA14 means that for the self-selected condition $n = 5$ whilst for the randomly paired condition $n = 6$.

Figure 3.6 presents the mean results of the IA for each of the pairing conditions. The three dimensions of the IA, Physical attraction, Social attraction and Task attraction are listed along the x-axis. The heights of the bars represent the mean IA value for each particular IA dimension within either the randomly paired of self-selected conditions. Error bars represent the standard error of the mean. The scores represent mean rating of interpersonal attraction within each pairing condition for each of the IA dimensions. The possible values for each of the interpersonal attraction dimensions ranges from 5 to 35. High IA scores represent a higher degree of interpersonal attraction within a given dimension If the self-selection protocol has grouped participant pairs by their degree of interpersonal attraction, then participants in the self-selected condition should show greater interpersonal attraction than the randomly paired condition.

In all of the IA dimensions presented in figure 3.6, the participants in the self-selected condition appear to have greater interpersonal attraction towards one another than the participants in the randomly-paired condition. However, the difference between conditions for the Task dimension may not be real due to overlapping standard errors. Given the $n$ for the groups ($n = 5$ for self selected, $n = 6$ for randomly paired), it would be inappropriate to conduct statistical testing. As such, only the trends can be interpreted and any differences cannot be demonstrated to be more than the effect of chance.

### Discussion

Results of the IA demonstrated that those participants in the self-selected condition were more likely to provide a greater interpersonal attraction rating for their partner on both Physical and Social attraction dimensions than those in the randomly paired condition. However, this effect was not seen for Task attraction.

Looking at the Physical attraction results, it may have been the case that the age range of the participants in each condition could have played a role. Those in the self-selected condition tended to be younger than those in the randomly paired

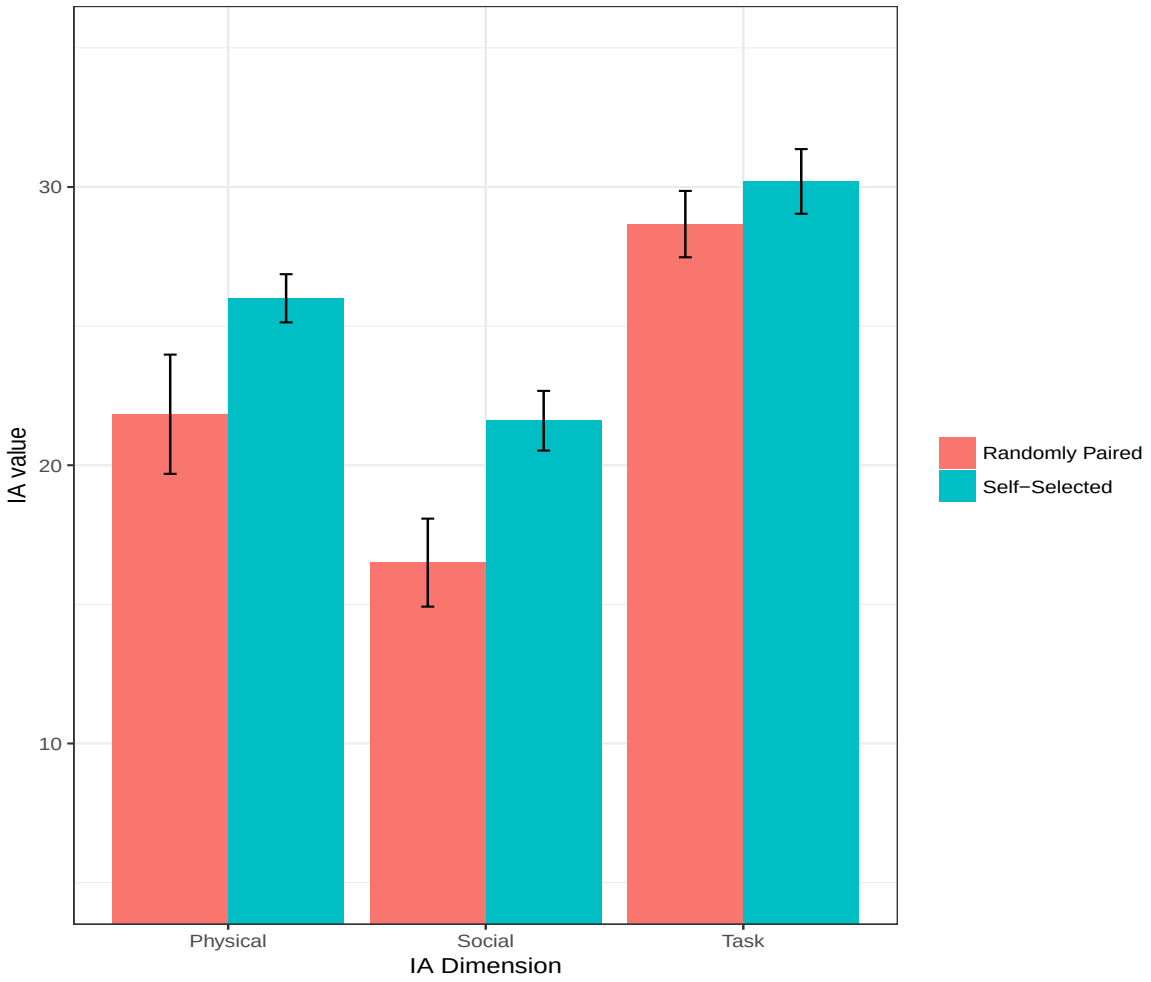



Figure 3.6: Barplot of IA scores for each of the IA dimensions, separated by pairing condition. Error bars represent the standard error.

group (as discussed in subsection 3.4.1). Since the Physical attraction questions in the IA contains some questions that relate to sexual attractiveness, it may have been easier for the younger participants to answer more openly. What is meant by this is that it may have been the case that the gender norms imposed by society in general have been gradually relaxing, thus allowing younger participants to offer more open answers to questions regarding sexual attractiveness of a member of the same sex. Older participants on the other hand, may have a more binary view of gender and as such, may offer more conservative responses that are akin to responding that a partner is not physically attractive at all simply because they are not of the opposite sex. Indeed, upon inspection of the data, the two oldest participants provided the lowest overall ratings for Physical attraction. Although this is speculation, such an effect is possible and might be contributing to the effect seen for Physical attraction. If this is present then it could be masking the true effect of pairing condition for Physical attraction.

The difference between the two pairing groups in Social attraction suggests that the self-selection protocol may have provided some form of grouping based on the social values of the participants. This would fit with the literature surrounding facial

preference being socially mediated (Bronstad & Russell, 2007). It is reasonable to assume that if one tends to select images of people that resemble oneself to a certain degree then they may have similar social values. It may be the case that by offering participants the opportunity to self-select their partners, they are picking up on the outward image that the person in the picture is trying to convey. For those that were randomly paired, this opportunity to select based on the outward image of the person in the picture was not offered and therefore, the values reported for Social attraction are lower. The ability to determine the amount of Physical or Task attraction towards a person may be hindered when making an assessment from a photograph. Social attraction on the other hand is easier to determine as a number of social factors can be identified from a photograph. These include factors like age, ethnic background, perceived class and perceived friendliness, amongst others (Berry, 1991; Todorov, Said, Engell, & Oosterhof, 2008; Ewing, Caulfield, Read, & Rhodes, 2015). Having said this, the issue of the spread of ages in the data persists. It may be the case that Social attraction is higher in the self-selected condition because those that are younger, have more closely matched social values than those with a larger age difference between them.

Although there appears to be no difference between the two pairing conditions for Task attraction, it is worth noting that this dimension holds the highest response values. This suggests that participants consistently felt a strong Task attraction towards their partner. Part of this will have likely stemmed from their ability to complete the task together. Given that Task attraction represents how easy or worthwhile participants found working with their partner, this result may be partially explained by the nature of the task. The fact that the DiapixUK task is fundamentally collaborative and cannot be completed without the aid of a partner may have had an impact on the results. Since both participants are working towards the same goal and require information from their partner in order to reach that goal, it would seem reasonable to conclude that unless they were deliberately withholding information or attempting to deceive, then there would be little reason to offer anything other than a favourable rating of their partner. One possible way in which to test this would be to use a confederate as the partner who deliberately offers false information or tries to derail the joint objective.

In sum, it appears that there may be a relationship between pairing condition and both Physical and Social attraction, as far as can be determined form this sample. However, these results may be influenced by factors that were not foreseen in the construction of the experiment. Further work, evaluating these additional factors is necessary to rule out their influence.

### 3.4.3 General discussion

The key question to ask here is whether or not the self-selection protocol was effective in assigning participants to a group that either had similar personalities or high levels of interpersonal attraction or both. The results of the BFI demonstrate no clear relationship with the pairing condition. It can therefore be concluded that the self-selection protocol was not able to group participants by personality, in this sample. The results of the IA demonstrated some relationships for two of the IA dimensions (Physical and Social) and pairing condition. However, there were a number of unforeseen factors that might partially explain these findings. As such, it cannot be fully concluded that the self-selection protocol was able to group participants by Physical and Social attraction.

It may be the case that the self-selection protocol is actually grouping by age rather than by any other type of factor. The fact that participants self-selected based on images of each other may have introduced an age bias into the data. If participants do select based on traits in others that they find to be similar to themselves, then it stands to reason that they would self-select others of a similar age. By the self-selected participants choosing others of the same age, the remaining participants tended to be older. This difference between the age of the self-selected and randomly paired groups may have had an influence on the outcome of both the BFI and the IA. As it stands, this assumption of age impacting in the results is only speculative. However, it is a valid concern and could be further investigated by testing a group of younger adults against a group of older adults to see if there is a consistent bias across age groups. Although this is not possible within the scope of this thesis.

The evaluation of the self-selection protocol also suffered from issues such as small sample sizes and non-standard test implementation (in the case of the IA). However, even given the small sample size and non-standard implementation of the IA, if there had been a clear and sizeable distinction between the pairing conditions there may have been some justification to conclude that the self-selection protocol had served its function. This is not what the data demonstrate though. Of the two tests applied to the participants, the IA looks to be the most promising, but even that has its issues. Although the Physical and Social dimensions of the IA do demonstrate sizeable differences between pairing conditions, there are good reasons to believe that these results might be somewhat compromised (see subsection 3.4.2). It may be that the theory behind selection of similar people to oneself based on photographs was being stretched too far. Basing a protocol around this and expecting clear distinctions between groups was perhaps rather ambitious.

Since no conclusive evidence is available from these data to support the use of the self-selection protocol as a method for grouping participants by personality or interpersonal attraction, it will not be taken further in this thesis. More comprehensive screening of participants prior to participation for factors such as sexual orientation or a tighter restriction on the age range might help to resolve some of

the issues here. However, the main aim of implementing the self-selection protocol was to evaluate predictions that similar personalities and greater attraction would maximise accommodative behaviour. Whilst this is now not possible due to the lack of specificity in the self-selection protocol, it does not have a major impact on the experiments carried out here. The hypothesis is that accommodation would still occur as it would in normal speech, although it may not be maximised.

Overall, there cannot be said to be enough evidence to conclude that the self-selection protocol was effective in producing groups of people with similar personality types or with a high degree of interpersonal attraction. However, the interpretations and considerations of the results of the BFI and IA can certainly be used to aid further investigation into the effects of a protocol such as the one proposed to group participants based on factors such as personality and interpersonal attraction.

## 3.5 Phonetic Analyses

Before considering the computational approach for the detection of accommodation, a more traditional phonetic analysis of the data will be presented. Three phonetic features of interest have been selected, these are VOT of stop consonants, F1/F2 of vowels and speech rate. The phonetic features of interest have been selected in order to both mirror the most common acoustic-phonetic measures reviewed in subsections 2.2.3 and 2.2.4 as well as to provide a reasonable coverage of short, medium and long-term speech features.

The aims of this section are to:

- Describe the methods used to extract and preprocess the data for each of the phonetic variables of interest (subsection 3.5.1).
- Describe the statistical methods used to interpret the data (subsection 3.5.2).
- Report the results and provide an interpretation of the statistical analyses for each of the phonetic variables of interest (subsection 3.5.3).
- Provide a general discussion of all of the findings taken together (subsection 3.5.4).

The key questions that are to be addressed concern:

1. Assessing the relationship between the recent realisations of a partner and the current realisation of a speaker.
2. Assessing the relationship between realisations of speakers in relation to the total length of the experiment using a non-linear modelling approach.
3. Assessing the relationship between realisations of speakers in relation to the length of an interaction using a non-linear modelling approach.

### 3.5.1 Data extraction and pre-processing

This subsection provides details of the processes used to extract phonetic variables of interest and of any pre-processing steps taken in order to clean the data. This information is provided separately for each phonetic variable of interest, ie. VOT of stop consonants, the F1 and F2 values of stressed vowels and speech rate. All data was extracted from the LaBB-CAT corpus that had been built from the recordings and transcriptions of the experiment as described in subsection 3.3.3.

#### VOT

VOT here includes both positive and negative VOT. Positive VOT is defined here as ‘the time interval between the burst that marks release of the stop closure and the

onset of quasi-periodicity which reflects laryngeal vibration' (Lisker & Abramson, 1967, pp.1). Negative VOT is the length of time from voicing onset to the creation of the stop closure. Data was extracted from the LaBB-CAT corpus by performing a search for all stressed instances of English voiced and voiceless plosives (/b/, /d/, /g/, /p/, /t/ and /k/) for all speakers across all interactions. Some examples of these are the stop consonants in words such as 'beach','peach','tin' and 'doll'. The search for each of the plosives was performed separately. For each search, LaBB-CAT produced three different file types:

1. CSV File

   A comma separated vales (csv) file with all necessary information concerning participant, orthography, transcript and segment data.

2. Praat TextGrid File

   A Praat TextGrid file was produced for each of the resulting identified plosives. Each of these TextGrid files consisted of the whole phrase that the target plosive was contained in. Each Praat TextGrid file was matched with a counterpart wav audio file.

3. WAV Audio File

   A wav audio file of the recording for the phrase that the target plosive was contained in was produced. Each audio file was matched with a counterpart Praat TextGrid file.

Following extraction of the plosive data, a number of pre-processing steps were taken. The first of these steps concerned an error that arose during LaBB-CAT's automated TextGrid generation where some apostrophes were substituted for commas. These commas had to be identified and replaced with apostrophes. This was completed with a short search and replace BASH script.

The second pre-processing step was to modify the TextGrid files so that they were compatible with the semi-automatic VOT extraction tool, AutoVOT (Sonderegger & Keshet, 2012). AutoVOT requires every TextGrid file to have a tier with boundaries marking the location of the plosive of interest. To complete this task quickly, an R script was written to handle this. Using the data in the corresponding csv file from LaBB-CAT it was possible to edit the TextGrid files so that an additional tier was added with boundaries placed in the same position as the identified segment based on LaBB-CAT's forced alignment. Whilst it may have been possible to use the tier produced for the segments by LaBB-CAT as an indicator of the desired target, having multiple targets in the same phrase would cause issues for AutoVOT.

The third pre-processing step was to downsample the wav audio data. AutoVOT can only handle audio data sampled at 16 kHz but the data stored and outputted by LaBB-CAT was sampled at 44.1 kHz. Resampling of the data was completed using the Linux command line utility SoX (Sykes & Giard, 2015).

Before considering the final pre-processing step, it is worth outling how the way in which AutoVOT is used (as detailed in Stuart-Smith et al., 2015). The process of automatic labelling of VOTs by AutoVOT involves the training of a classifier in order to predict VOT measures for new files. Separate classifiers must be trained for positive and negative VOT. AutoVOT uses a set of hand-labelled VOT measures during the training of the classifier. The input for training is given as speech segments, each containing a VOT region. The VOT regions in the speech segments are identified by the *burst onset* and the *voicing onset*, the difference between these times is the VOT and taken together they are known as an *onset pair*. *Onset pairs* are used for classifier training. Following classifier training, the *onset pairs* of new data are predicted by the algorithm.

**Classifier Training**

The classifier function is a weighted sum of 62 *feature maps*, each of which corresponds to a quantity computed for a given speech segment and hypothesized *onset pair*. The weights for each of the 62 feature maps are set based on the training data of hand labelled VOTs. Training develops maximally different values for 'good' versus 'bad' onset pairs. The process can be thought of as developing a measure of 'goodness' for each hypothesised onset pair.

**Classifier Testing**

Once training is completed, the classifier can be used to predict VOT for any speech segment containing a VOT region. The classifier returns the values corresponding to the onset pair for which the function produced the maximal value.

Since the AutoVOT algorithm requires a classifier to be trained over hand labelled VOTs, this was a necessary pre-processing step. This was done by providing AutoVOT with samples of 100 hand labelled VOTs for both voiced and voiceless plosives (200 samples in total). Hand labelling of the VOTs was performed on randomly selected voiced and voiceless plosives.

Once the above data extraction and preprocessing had been completed, it was possible to submit the resulting Praat TextGrid files and the wav audio files to AutoVOT for VOT extraction. All optional arguments were left as default for AutoVOT except for the *--csv_file* argument which was used to output AutoVOT's classifications to a csv data file. The values for minimum and maximum VOT length were manually set at 5 ms and 250 ms respectively.

Once AutoVOT had finished calculating the VOTs, it provided both Praat TextGrid files for each target plosive along with a csv file for each plosive category (ie. /b/, /d/, /g/, /p/, /t/ and /k/). The csv file contained the wav audio file name, the time at which the VOT begins within the audio file, the VOT itself (measured in seconds) and a confidence measure for each VOT representing the quality of the prediction.

These csv files were merged with data from previous LaBB-CAT searches for other acoustic features using the dplyr package (Wickham & Francois, 2016) in R. After the data were in a format that could be easily handled and analysed in R it was possible to begin cleaning the data.

The data resulting from the AutoVOT processing was entered into R and using existing LaBB-CAT labelling, it was tagged for the position of the target phoneme in the word and for the position of the word in the phrase. These both contained the labels initial, medial and final. The phrase position tagging also contained a 'not phrase' label for phrases that consisted of single words. After the data were tagged, all VOT data was converted from seconds to milliseconds.

Data were then restricted to only word initial, non-final phrase position data values. This was done to remove erroneous VOT values arising from sequences of consecutive plosives and large VOT values arising from plosives occurring at the end of words in phrase final positions. The data was then further restricted to values which had a confidence level of $> 220$ associated with them. This value was selected based on consultation with the authors of AutoVOT (Sonderegger & Keshet, 2012). Finally, conservative outlier removal of $\pm 1.5 \times IQR$ was applied to the data to remove any remaining erroneous data points.

**Vowels**

The target measures for this speech feature were the first two formant values (F1 and F2) of a number of monophthong vowels. In order to extract only the stressed vowels of the Glaswegian vowel inventory (as described in Stuart-Smith (2004)) an additional searchable layer had to be generated within LaBB-CAT. This allowed for automatic tagging of stressed vowels after the transcriptions had been uploaded to the LaBB-CAT corpus. The location of stressed vowels within any given word was produced by a reverse look-up in the UNISYN dictionary (Fitt, 2002). These locations were then transposed onto the phonemic transcriptions (drawn from the previously generated HTK aligned phonemic transcription) in the newly generated 'Stress' layer. The stressed vowels of interest that were extracted are as follows: /i, ɪ, e, ɛ, a, o, ɔ, u, ʌ/. All vowels that were followed by a liquid consonant, /r/ or /l/ were excluded from the search due to the impact that they have on vowel formant values in Glaswegian (Lawson et al., 2013). The search also excluded function words and pronouns that contain a target vowel and are commonly contracted. For example, the search for /i/ excluded "he","she","we" and "the".

In order to ensure accuracy in the vowel searches, each of the vowels of interest were searched for individually across all participants. The resulting data files were then downloaded from LaBB-CAT and concatenated in R to produce a file that contained data concerning all vowels of interest. The files downloaded from LaBB-CAT also contained meta data about the speakers, transcripts and variables.

The data were then resubmitted to LaBB-CAT in order to find the formant values

of all of the vowels of interest. LaBB-CAT called Praat and used the data contained in the uploaded file to work through each of the vowels of interest to return F1, F2 and F3 values at 25%, 50% and 75% of the way through each vowel. Whilst, F3 values were obtained, they were not taken any further in this work. The formant value information was then automatically added to the existing data file. Once the formant values for all vowels had been extracted, the updated data file was returned.

Given that the data were force aligned in LaBB-CAT, and errors can occur (eg. misalignment of vowel boundaries), the data was pruned in order to ensure that the vowel measures were within a sensible range. This pruning consisted of the following steps. Firstly, any errors in the formant extractions as automatically identified by LaBB-CAT were excluded. Then the mean and range as calculated across samples from each vowel at 25%, 50% and 75% of the length of the vowel were taken. Upper and lower bounds for F1 and F2 values were set as $\pm 1.5 \times IQR$. Any vowels with F1 or F2 values outside of this range were excluded. This was classed as the removal of outliers. The data was then corrected for values with disturbing ranges of F1 or F2 values. Any vowels where the range across the three time points exceeded 33% were identified and excluded. This pruning process helped to ensure that any errors introduced into the dataset through less than perfect transcription alignment was minimised.

The data were then imported into R (R Core Team, 2016) for analysis and were Lobanov normalised (Lobanov, 1971). Normalisation was carried out using the `vowels` package (Kendall & Thomas, 2014) in R. Each normalised dataset, containing both F1 and F2 for each target vowel, was produced independently and was drawn from the pruned raw dataset.

### Speech Rate

Speech rate, as the name suggests, concerns the rate at which a speaker produces units of speech (eg. segments, phones, syllables, words). However, the type of unit used to determine speech rate as attracted some debate (Pfitzinger, 1998). Data was extracted from LaBB-CAT by performing a search for all words for all speakers in all transcriptions. The files outputted from LaBB-CAT contained information regarding the syllable count in each word. There were some instances where the forced alignment didn't work and this resulted in some words lacking a start and end time. These were all removed from the data set.

For the analyses in this thesis the same general approach for calculating speech rate as used in Casasanto et al. (2010) is employed. However, where Casasanto et al. (2010) calculated speech rate using words per second, here the number of syllables per second was calculated. The equation for this calculation is provided in equation 3.1. Syllables were chosen over the use of words as they are more robust to lexical differences in speaker styles (ie. more reliable if one speaker tends to use longer words). The number of syllables in a word was calculated by a reverse lookup

to the Unisyn lexicon (Fitt, 2002). This means that syllable counts per word will be based on the values provided in the Unisyn lexicon and was implemented such that it was sensitive to the particular local accent variations of Glaswegian. The speech rate was then calculated for each of the utterances in R (R Core Team, 2016) by dividing the total number of syllables in an utterance by the total length of time of that utterance.

$$SR_{syl} = \frac{N_{(Syllables\ in\ Utterance)}}{t_{(Length\ of\ Utterance)}} \tag{3.1}$$

### 3.5.2 Statistical methodology

Here, the aim was to evaluate if accommodation changes over time for a number of phonetic variables. It has been shown that accommodation is often subtle when extracted from studies that lack tight experimental controls and use acoustic-phonetic measures (Collins, 1998; Evans & Iverson, 2007; Purnell, 2009). Further to this, the tracking of accommodation over time has often been performed post-hoc (eg. *AXB* paradigms). What was trying to be achieved here is to detect accommodation over time through the use of live, interactional speech, as measured with acoustic-phonetic measures. In order to do so with reasonable statistical power, a statistical methodology that combines linear mixed-effect regression models (MEMs; eg. Baayen, 2008) and generalized additive models (GAMs; Hastie & Tibshirani, 1986; Wood, 2011) was employed. The statistical approach for the analysis of the phonetic variables draws upon an approach used by Sonderegger et al. (Accepted) to model change in phonetic variables over time. The use of MEMs is commonplace in phonetic analyses (eg. Adank & Janse, 2010; Bane, Graff, & Sonderegger, 2010; Drager & Hay, 2012; Hay, Pierrehumbert, Walker, & LaShell, 2015) and are useful in expanding the error term of the standard linear regression equation such that more of the error can be accounted for. The use of GAMs are less common but they allow for the modelling of non-linear relationships between variables because the predictive function is learned rather than being established a priori. The predictive function is established by applying smoothing functions to capture the impact of the predictive variables. These functions can be linear or non-linear, depending on the underlying patterns in the data (Hastie & Tibshirani, 1986). The statistical analyses are conducted in the same way for each of the phonetic variables of interest, controlling for appropriate sources of error relative to that phonetic variable. All statistical analyses were conducted in R (R Core Team, 2016).

The aim of the statistical analysis is to provide an answer to the following three questions for each of the phonetic variables of interest:

1. Is there a linear relationship between the recent realisations of the partner and the current realisation of the speaker?

Question 1 is asked to see if the recent realisations of a particular phonetic variable from a speech partner impact a speaker's current production of that phonetic variable. This is akin to asking if, as time progresses within an interaction, local phonetic input drives accommodation. This is performed in the following way. Once all data pre-processing and cleaning has been completed, each phonetic variable (eg. voiced VOT, voiceless VOT, TRAP F1, TRAP F2 etc.) is submitted to an MEM using the lmer function from R's lme4 package (Bates, Mächler, Bolker, & Walker, 2015) with the following structure:

$$X \sim (1|effect_1) + (1|effect_2) + ...(1|effect_n) \tag{3.2}$$

where $X$ is the data for the variable of interest, and $(1|effect_{1...n})$ are random effects that are known to impact on realisation of the phonetic variable under investigation. The residuals are then drawn from this model using the resid function from R's stats package (R Core Team, 2016), which provides values for the variable of interest removed of any known random effects. These residual values are then used to calculate the values for the recent realisations of the partner and the current realisation of the speaker. This is done by iterating across the dataset, selecting the start time of each variable of interest and calculating the mean of the previous 3 variable values for the other speaker (if they exist). This is done for both speakers. The choice of using the previous 3 realisations was to only extract from as local an area as possible. This helps to ensure that realisations from too far back in an interaction were not included. Since the random effects have already been accounted for in the dataset, a linear regression is then performed using the lm function from R's stats package (R Core Team, 2016) to test for any relationship between recent realisations for all speakers and current realisations for all speakers.

2. When the difference between partner and speaker realisations within an interaction is modelled non-linearly, is there a relationship with the presentation position of the stimuli?

   Question 2 uses the presentation position of the particular DiapixUK stimuli as a measure of the overall progress of the experiment. This question can be thought of as asking whether the change in realisation between speakers holds a non-linear relationship within an interaction and if that is related to the overall progression of the experiment. In the traditional interpretation of accommodation, it would be expected that there would be a smaller difference between speakers during interactions that take place later in the experiment. This analysis is performed by fitting a GAM, using the gam function from R's mgcv package (Wood, 2011) to the residual data values obtained from question 1 for each speaker in each interaction as a function of smoothed time (because

the phonetic variable samples do not have regular time intervals).

$$X \sim s(X_t) \tag{3.3}$$

where $X$ are the variable data points of interest, $s$ is the smoothing function and $X_t$ are the time points associated with the data points. A $1000$ point, equally spaced time-series is then produced for this given speaker, in this given interaction, with the same length as the interaction (eg. if the interaction is $500\,\mathrm{s}$ long then the time-series will have 1000 equally spaced points between 0 and $500\,\mathrm{s}$). The difference between the GAMs for each of the speakers in relation to the equally spaced time-series, for each point on the time-series, is then predicted using the $\mathrm{predict}$ function from R's $\mathrm{stats}$ package (R Core Team, 2016). This provides a non-linear measure of change in each of the speakers for a given phonetic variable in relation to a linear $1000$ point time-series. The mean difference between the two resulting $1000$ point predictions is then taken to represent the total difference between speakers over time for that phonetic variable. These can then be plotted against presentation position to ascertain if a relationship is present. Since the non-linear difference between predicted values for the phonetic variable are now integrated into the dataset, they are then submitted to a linear regression, using $\mathrm{lm}$ from $\mathrm{stats}$ (R Core Team, 2016), to statistically verify a relationship.

3. When the difference between partner and speaker realisations within an interaction is modelled non-linearly, is there a relationship with interaction length?

   Question 3 uses the interaction length of the particular DiapixUK stimuli as a measure of difficulty in each interaction. The longer an interaction takes, the more difficult the participants found it. This question can be thought of as asking whether the change in realisation between speakers holds a non-linear relationship within an interaction and if that relates to task difficulty. This is because the interaction was ended either after all differences between images were found or after $15\,\mathrm{min}$, whichever came first. As the experimenter was listening to all of the interactions, it can confidently be asserted that participants did not stray off topic or engage in any activity that was contrary to the completion of the task. It is hypothesised that more difficult tasks, longer interactions, will generate more convergence and that the participants will therefore differ less in their realisations of phonetic variables. The statistical analysis is the same as for question 2 until the plotting stage, where interaction length is used instead of presentation position. The same applies to the linear regression to test for a relationship.

### 3.5.3 Results

Results for all three types of phonetic variables considered are presented here along with discussions for each individual variable. All tables in this subsection were produced using the Stargazer package (Hlavac, 2015) for R.

#### VOT

As outlined above, here three main questions will be addressed:

Q1 Is there a linear relationship between the recent VOT realisations of the partner and the current VOT realisation of the speaker?

Q2 When the difference between partner and speaker VOT realisations within an interaction is modelled non-linearly, is there a relationship with the presentation position of the stimuli?

Q3 When the difference between partner and speaker VOT realisations within an interaction is modelled non-linearly, is there a relationship with interaction length?

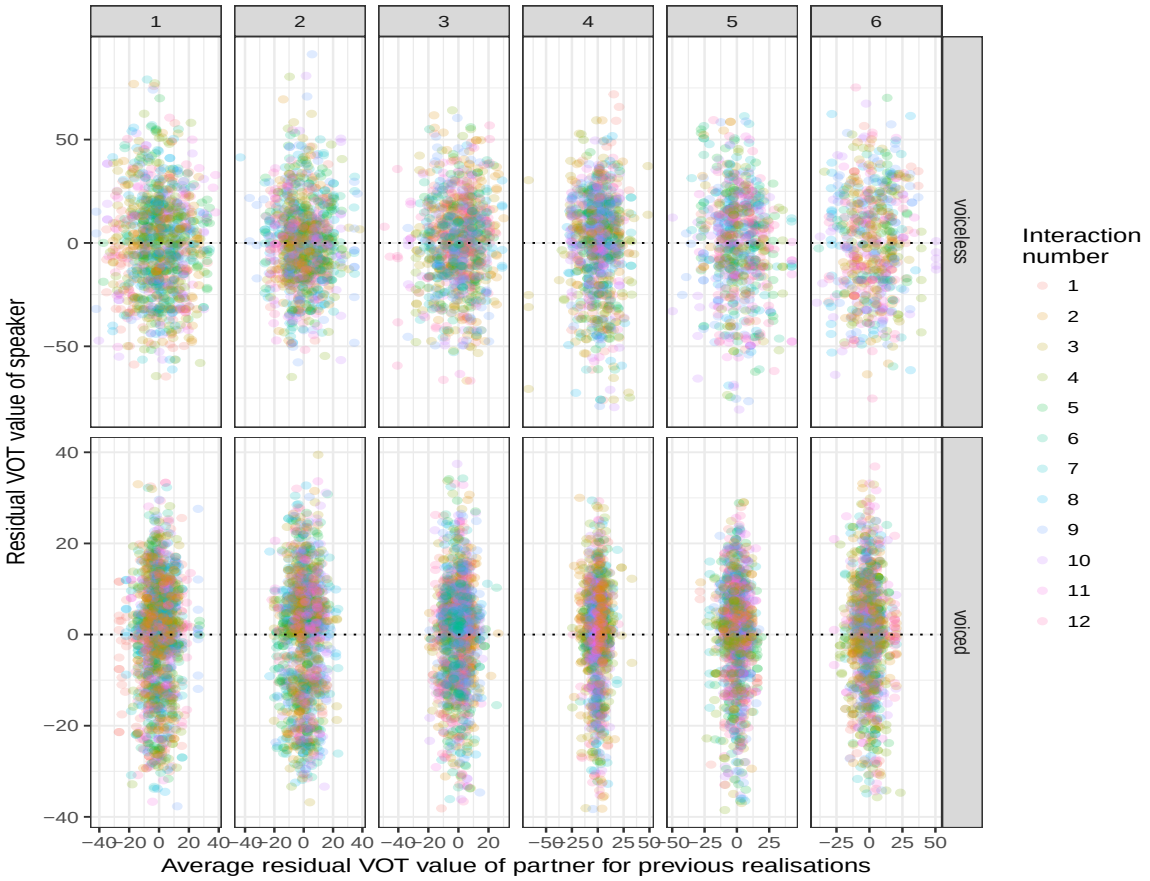



Figure 3.7: Scatter plot of the residual VOT values, where effects attributable to word and speaker have been removed. Columns represent the speaker pairs and rows represent plosive type (voiced and voiceless). Note that the scales for voiced and voiceless values are different. There are no clear trends observable in the data.

Figure 3.7 shows the residual VOT values of a speaker (y-axis) plotted against the average residual VOT values of their partner's recent realisations (x-axis). The data is further separated by speaker pair (columns) and voiced/voiceless VOT (rows, voiceless VOT on top). Colours represent the interaction from which the data were drawn.

If there were a strong and clear relationship between the two variables then the data should demonstrate some pattern (eg. positively sloping diagonal line if positively correlated). What is presented here are clouds of data points with no clear patterns. This would suggest that there is no strong relationship between the variables. However, there may still be a weak relationship that is not detectable by eye.

| | *Dependent variable:* | |
|---|---|---|
| | Residual VOT speaker values | |
| | Voiced | Voiceless |
| Residual VOT partner values | 0.022 | −0.006 |
| | (0.017) | (0.022) |
| Constant | 0.005 | −0.104 |
| | (0.123) | (0.307) |
| Observations | 8,835 | 5,659 |
| $R^2$ | 0.0002 | 0.00001 |
| Adjusted $R^2$ | 0.0001 | −0.0002 |
| Residual Std. Error | 11.530 (df = 8833) | 23.105 (df = 5657) |
| F Statistic | 1.728 (df = 1; 8833) | 0.072 (df = 1; 5657) |
| *Note:* | | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 |

Table 3.3: Output of linear regression model comparing previous realisations of voiced and voiceless VOT from a partner with that of the speaker. VOT values are reported in milliseconds and brackets report standard errors. The constant is the y-axis intercept of the regression line.

Table 3.3 presents the results of the linear regression performed on the residual VOT values of the speakers. In this test, the dependent variable was the current residual VOT value of the speaker and the independent variable was the average residual VOT values of their partner's recent realisations. Because the residuals of the original MEM are being used, the random effects have already been accounted for. In this case, the MEM accounts for random effects of speaker, word and phrase position. DiapixUK task number was also entered as a random effect in the MEM to account for any differences due to DiapixUK stimuli but it accounted for $0.0000$ of the variance so was removed. Table 3.3 shows that the difference in VOT attributable to the previous VOT values of the partner is $0.022\,\text{ms}$ for voiced VOT and $-0.006\,\text{ms}$ for voiceless VOT, we can assume that influence of local partner VOT realisation is

minimal in both cases. This confirms the conclusions drawn from the interpretation of figure 3.7.

When considered in terms of the three questions that are being asked here, this answers question 1:

Q1 There is no linear relationship between the recent VOT realisations of the partner and the current VOT realisation of the speaker.

The remaining questions, questions 2 and 3, can be answered at the same time. Figure 3.8 shows the GAM predicted difference in VOT (ms) between speakers over the course of an interaction (y-axis) plotted against presentation position (x-axis). The data is further separated by voiced/voiceless VOT (rows, voiceless VOT on top). Colours represent speaker pairs.

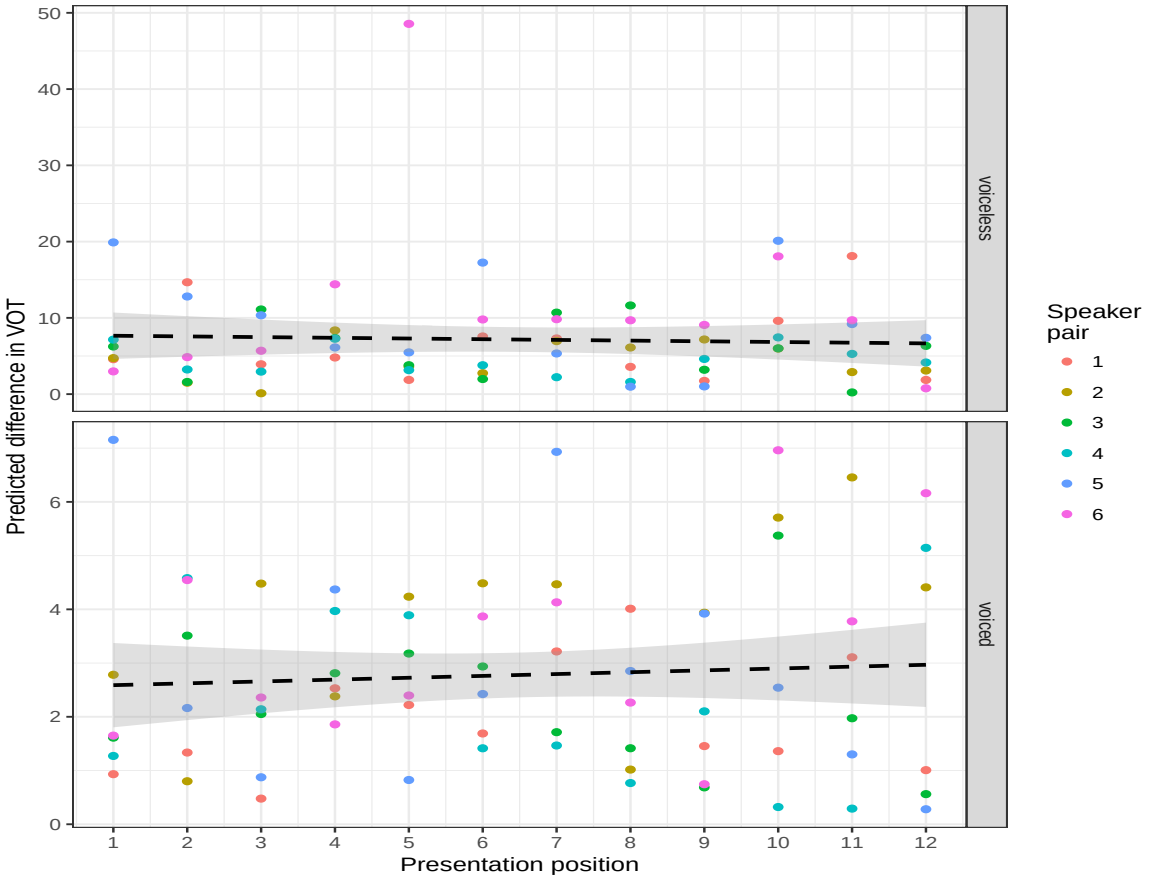



Figure 3.8: Scatter plot of the GAM predicted difference in VOT as a function of presentation position. Data is separated into voiced and voiceless VOT, voiceless VOT on top. Colours represent the speaker pair, the dashed line indicates the linear regression line and the shaded area is the standard error. Note that there is a difference in scale between voiced and voiceless plots on the y-axis. Predicted VOT values are presented in milliseconds.

In figure 3.8 there are smaller predicted differences for voiced VOT than for voiceless VOT, which would be expected given their relative average lengths. Overall, there does not appear to be a general trend for either voiced or voiceless predicted VOTs over the course of the experiment as both lines of best fit are reasonably flat.

Figure 3.9 shows the GAM predicted difference in VOT between speakers over the course of an interaction (y-axis) plotted against interaction length (x-axis). The data is further separated by voiced/voiceless VOT (rows, voiceless VOT on top). Colours represent speaker pairs.

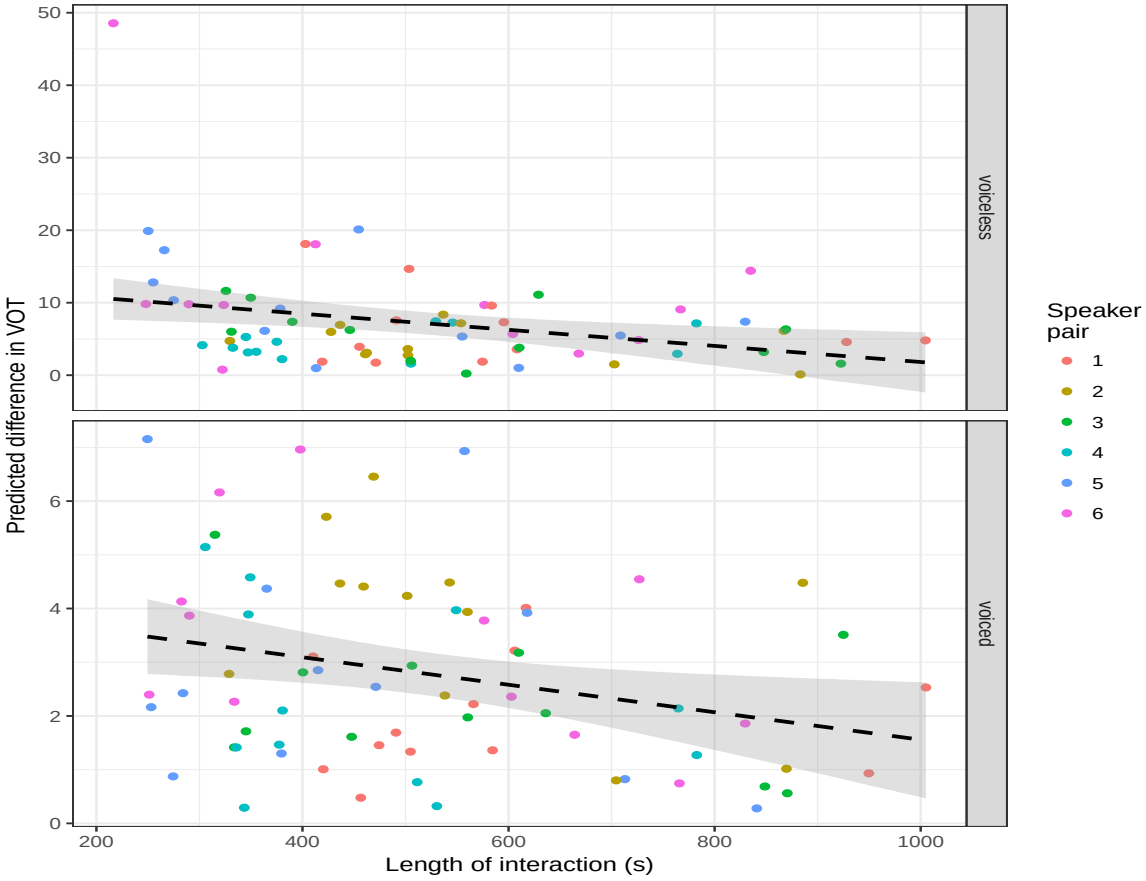



Figure 3.9: Scatter plot of the GAM predicted difference in VOT as a function of interaction length. Data is separated into voiced and voiceless VOT, voiceless VOT on top. Colours represent the speaker pair, the dashed line indicates the linear regression line and the shaded area is the standard error. Note that there is a difference in scale between voiced and voiceless plots on the y-axis. Predicted VOT values are presented in milliseconds.

In figure 3.9 the trends appear to be stronger with less predicted difference in VOT for longer interactions than for shorter interactions. This is true for both voiceless and voiced VOTs, although the trend in voiceless VOT might be being weighted by an outlier. This would suggest that the non-linear relationships between speakers within an interaction vary by the degree of difficulty that a pair has in completing a task. However, even though the trends appear to be stronger than those found in figure 3.8, they still look to be somewhat weak.

The output of the linear regressions run on the GAM predicted VOTs are presented in table 3.4 and provide verification of observed trends.

However, although the significance levels are high, the actual effect sizes are small. There are no significant effects of presentation position on voiced or voiceless VOT. Interaction length has a significant effect on voiced VOT of $-0.003\,\mathrm{s}$ per $1\,\mathrm{s}$

| | *Dependent variable:* | | | |
|---|---|---|---|---|
| | Residual VOT speaker values | | | |
| | Voiced | Voiceless | Voiced | Voiceless |
| Presentation position | 0.034 | −0.091 | | |
| | (0.061) | (0.235) | | |
| Interaction length | | | −0.003** | −0.011*** |
| | | | (0.001) | (0.004) |
| Constant | 2.554*** | 7.749*** | 4.116*** | 12.923*** |
| | (0.446) | (1.729) | (0.581) | (2.209) |
| Observations | 72 | 72 | 72 | 72 |
| $R^2$ | 0.005 | 0.002 | 0.079 | 0.100 |
| Adjusted $R^2$ | −0.010 | −0.012 | 0.066 | 0.087 |
| Residual Std. Error (df = 70) | 1.775 | 6.883 | 1.707 | 6.537 |
| F Statistic (df = 1; 70) | 0.324 | 0.151 | 6.022** | 7.760*** |
| *Note:* | | | | *p<0.1; **p<0.05; ***p<0.01 |

Table 3.4: Output of linear model comparing the predicted overall difference in VOTs for each pair, for each interaction as produced from the GAM against both length of interaction and presentation position. Values are reported in milliseconds and brackets report standard errors. The constant is the y-axis intercept of the regression line.

unit increase of interaction length, $R^2 = 0.079, F(1,70) = 6.022, p = 0.016, r = 0.281$. Interaction length also had a significant effect on voiceless VOT of $-0.011$ s per 1 s unit increase of interaction length, $R^2 = 0.100, F(1,70) = 7.760, p = 0.007, r = 0.316$. This suggests that whilst the effect may be measurable, it is a subtle phenomenon that holds a non-linear relationship within an interaction.

To put this in terms that answer questions 2 and 3, these results suggest that:

Q2 When the difference between partner and speaker VOT realisations within an interaction is modelled non-linearly, there is no relationship with the presentation position of the stimuli.

Q3 When the difference between partner and speaker VOT realisations within an interaction is modelled non-linearly, there is a relationship with the interaction length, but only a small one.

In other words, speakers vary in VOT accommodation in a non-linear way during an interaction but maintain an overall standard VOT across the experiment. The effects reported here are small but significant, this might suggest that an assessment of VOT taken by itself could be sufficient to capture accommodation during a live interaction. However, given the very small effect sizes, it would be difficult to measure accommodation using only VOT without adapting the experimental environment to

maximise accommodation specifically in VOT. This would likely mean removing the element of live interaction which is key to the research goals of this thesis.

### Vowels

Having taken measures for all stressed monophthongs, the assessment of potential accommodation was considered here for three vowels, for which specific additional predictions about possible sociolinguistic variability could be made. The results for the three different vowels are considered and the lexical sets as used in Wells (1982) will be used to assign keywords to each vowel. The vowels considered are STRUT (/ʌ/), THOUGHT (/ɔ/) and TRAP (/æ/). These vowels were chosen for analysis based on phonetic research on vowel change in the Glaswegian accent (Stuart-Smith, 1999; Stuart-Smith, 2004; Macaulay, 1976; Lawson et al., 2013) and fall into three categories: stable vowels, diachronic change vowels and social change vowels. These categories are specific to the local accent and one vowel is presented for each category. The stable vowel category is one where there would not be much expected shift in realisation for a Glaswegian speaker, STRUT is used to represent vowels in this category. The diachronic change vowel category is one where there would be some change expected over time, so there might be different realisations across ages for Glaswegian speakers and/or there may be more variation given that this vowel is known to be undergoing a change in progress, THOUGHT is used to represent vowels in this category. The social change vowel category is where the most variability would be expected since these vowels can be varied within the Glaswegian accent to indicate a number of social factors especially social class, TRAP is used to represent vowels in this category. For all vowels, the first (F1) and second (F2) formant frequencies are considered.

As with the results for VOT, three main questions will be addressed:

Q1 Is there a linear relationship between the recent F1/F2 realisations of the partner and the current F1/F2 realisation of the speaker?

Q2 When the difference between partner and speaker F1/F2 realisations within an interaction is modelled non-linearly, is there a relationship with the presentation position of the stimuli?

Q3 When the difference between partner and speaker F1/F2 realisations within an interaction is modelled non-linearly, is there a relationship with interaction length?

However, unlike the VOT results, question 1 will first be addressed for each vowel separately then questions 2 and 3 will be addressed for all vowels at the same time.

Figure 3.10 presents data relating to the residual Lobanov normalised formant values of the STRUT vowel, representing the stable vowel category. The top plot shows the F2 values and the bottom plot shows the F1 values. Both plots follow the

same form, the x-axis plots the average value of the partner's F1/F2 realisations and the y-axis plots the current F1/F2 value of the speaker. The data is further separated by speaker pair (columns) and colours represent the interaction from which the data were drawn.

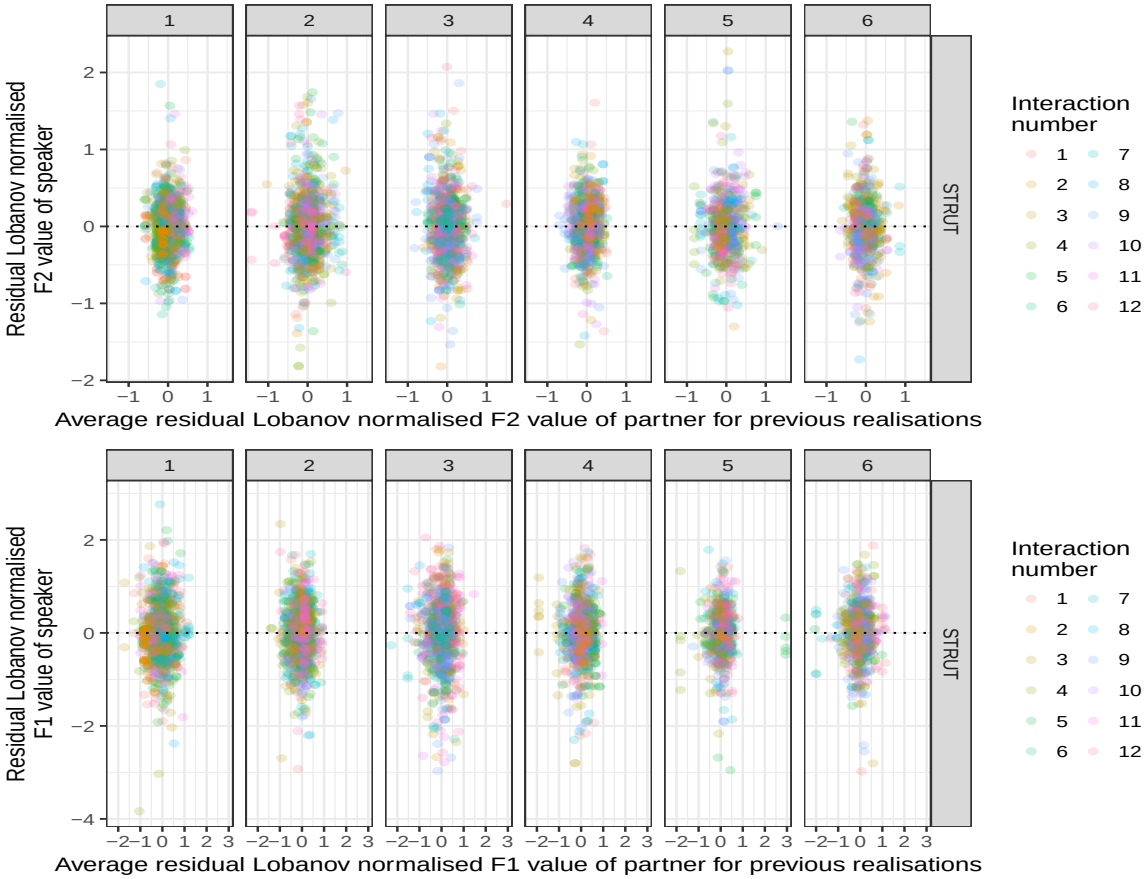



Figure 3.10: Scatter plot of the residual Labanov normalised STRUT F1 and F2 values (F2, top), where effects attributable to word, speaker and DiapixUK task number have been removed. Panels represent the speaker pairs. There are no clear trends observable in the data.

If there were a strong and clear relationship between the previous realisations of the partner and the current realisation of the speaker then the data should demonstrate some pattern (eg. positively sloping diagonal line if positively correlated). What is presented here, for both F1 and F2 are clouds of data points with no clear patterns. This would suggest that there is no strong relationship between the previous STRUT realisations of the partner and the current STRUT realisation of the speaker. A further observation could also be made about the distribution of the data, there are a number of extreme values in the STRUT data that extend the spread. This is especially true for F1 where the x and y axes are somewhat extended to account for the extreme values.

Table 3.5 presents the results of the linear regression performed on the residual Lobanov normalised F1 and F2 values of the speakers. In this test, the dependent variables were the current residual Lobanov normalised F1 or F2 value of the speaker and the independent variables were the average residual Lobanov normalised F1 or

| | *Dependent variable:* | |
|---|---|---|
| | STRUT F1 | STRUT F2 |
| Residual vowel partner value F1 | −0.003 | |
| | (0.021) | |
| Residual vowel partner value F2 | | 0.020 |
| | | (0.022) |
| Constant | −0.002 | −0.0002 |
| | (0.009) | (0.005) |
| Observations | 5,175 | 5,175 |
| $R^2$ | 0.00000 | 0.0002 |
| Adjusted $R^2$ | −0.0002 | −0.00003 |
| Residual Std. Error (df = 5173) | 0.615 | 0.378 |
| F Statistic (df = 1; 5173) | 0.015 | 0.865 |
| *Note:* | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 | |

Table 3.5: Output of linear regression model comparing previous realisations for Lobanov normalised F1 and F2 of STRUT from a partner with the current realisations for Lobanov normalised F1 and F2 of STRUT of the speaker. Brackets report standard errors. The constant is the y-axis intercept of the regression line.

F2 values of their partner's recent realisations. Table 3.5 shows that there are no significant effects of the partners' previous realisations of STRUT on the speaker's current realisation of STRUT. This is true for both F1 and F2. This confirms the conclusions drawn from the interpretation of figure 3.10.

When considered in terms of the three questions that are being asked here, this answers question 1:

Q1 There is no linear relationship between the recent F1/F2 STRUT realisations of the partner and the current F1/F2 STRUT realisation of the speaker.

However, given that STRUT was expected to remain roughly stable for these speakers in terms of variability relating to diachronic change and/or sociolinguistic variation, this is what was expected.

Figure 3.11 presents data relating to the residual Lobanov normalised formant values of the THOUGHT vowel, representing the diachronic change vowel category. The top plot shows the F2 values and the bottom plot shows the F1 values. Both plots follow the same form as figure 3.10.

Again, what is presented here, for both F1 and F2, are clouds of data points with no clear patterns. However, there appears to be a tighter distribution of the THOUGHT data than that seen in the STRUT data. This is especially true for F1, where STRUT had a greater range along the x-axis. Although, the few extreme values

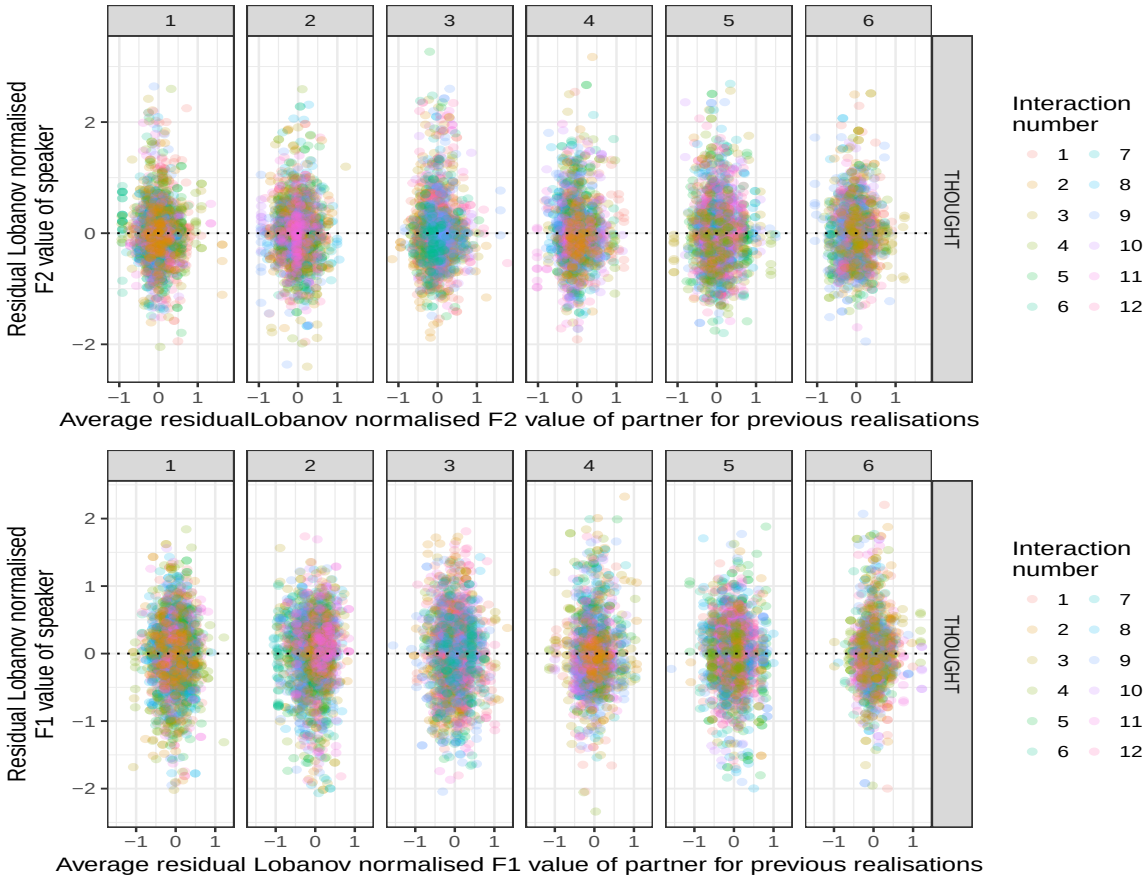



Figure 3.11: Scatter plot of the residual Lobanov normalised THOUGHT F1 and F2 values (F2, top), where effects attributable to word and speaker have been removed. Panels represent the speaker pairs. There are no clear trends observable in the data.

present in the STRUT data might explain the greater distribution. This would suggest that there is no strong relationship between the previous THOUGHT realisations of the partner and the current THOUGHT realisation of the speaker.

Table 3.6 presents the results of the linear regression performed on the residual Lobanov normalised F1 and F2 values of the speakers for THOUGHT. In this test, the dependent and independent variables were the same as those for table 3.5 except for THOUGHT rather than STRUT. The residuals of the original MEM are also the same as those used for STRUT. Table 3.6 shows that there are no significant effects of the partners' previous realisations of THOUGHT on the speaker's current realisation of THOUGHT. This is true for both F1 and F2. This confirms the conclusions drawn from the interpretation of figure 3.11.

When considered in terms of the three questions that are being asked here, this answers question 1:

Q1 There is no linear relationship between the recent F1/F2 THOUGHT realisations of the partner and the current F1/F2 THOUGHT realisation of the speaker.

Figure 3.12 presents data relating to the residual Lobanov normalised formant values of the TRAP vowel, representing the social change vowel category. The top

| | *Dependent variable:* | |
|---|---|---|
| | THOUGHT F1 | THOUGHT F2 |
| Residual vowel partner value F1 | 0.015 | |
| | (0.015) | |
| Residual vowel partner value F2 | | 0.012 |
| | | (0.016) |
| Constant | −0.001 | 0.0004 |
| | (0.005) | (0.005) |
| Observations | 11,468 | 11,468 |
| $R^2$ | 0.0001 | 0.00005 |
| Adjusted $R^2$ | 0.00000 | −0.00004 |
| Residual Std. Error (df = 11466) | 0.512 | 0.552 |
| F Statistic (df = 1; 11466) | 1.016 | 0.546 |
| *Note:* | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 | |

Table 3.6: Output of linear regression model comparing previous realisations for Lobanov normalised F1 and F2 of THOUGHT from a partner with the current realisations of Lobanov normalised F1 and F2 of THOUGHT of the speaker. Brackets report standard errors. The constant is the y-axis intercept of the regression line.

plot shows the F2 values and the bottom plot shows the F1 values. Both plots follow the same form as figure 3.10.

Here too, like the STRUT and THOUGHT vowels, both F1 and F2 are clouds of data points with no clear patterns. However, TRAP does look to have a wider distribution of values than both STRUT and THOUGHT. This is true for both F1 and F2, the range of values appears to be consistently broader for all speakers. This is something that would be expected given the assumed use of TRAP as a vowel for social identification in Glaswegian (Lawson et al., 2013). However, overall there is no evidence to suggest a strong relationship between the previous TRAP realisations of the partner and the current TRAP realisation of the speaker.

Table 3.7 presents the results of the linear regression performed on the residual Lobanov normalised F1 and F2 values of the speakers for TRAP. In this test, the dependent and independent variables were the same as those for table 3.5 except for TRAP rather than STRUT. The residuals of the original MEM are also the same as those used for STRUT. Table 3.7 shows that there are no significant effects of the partners' previous realisations of TRAP on the speaker's current realisation of TRAP. This is true for both F1 and F2. This confirms the conclusions drawn from the interpretation of figure 3.12.

When considered in terms of the three questions that are being asked here, this answers question 1:

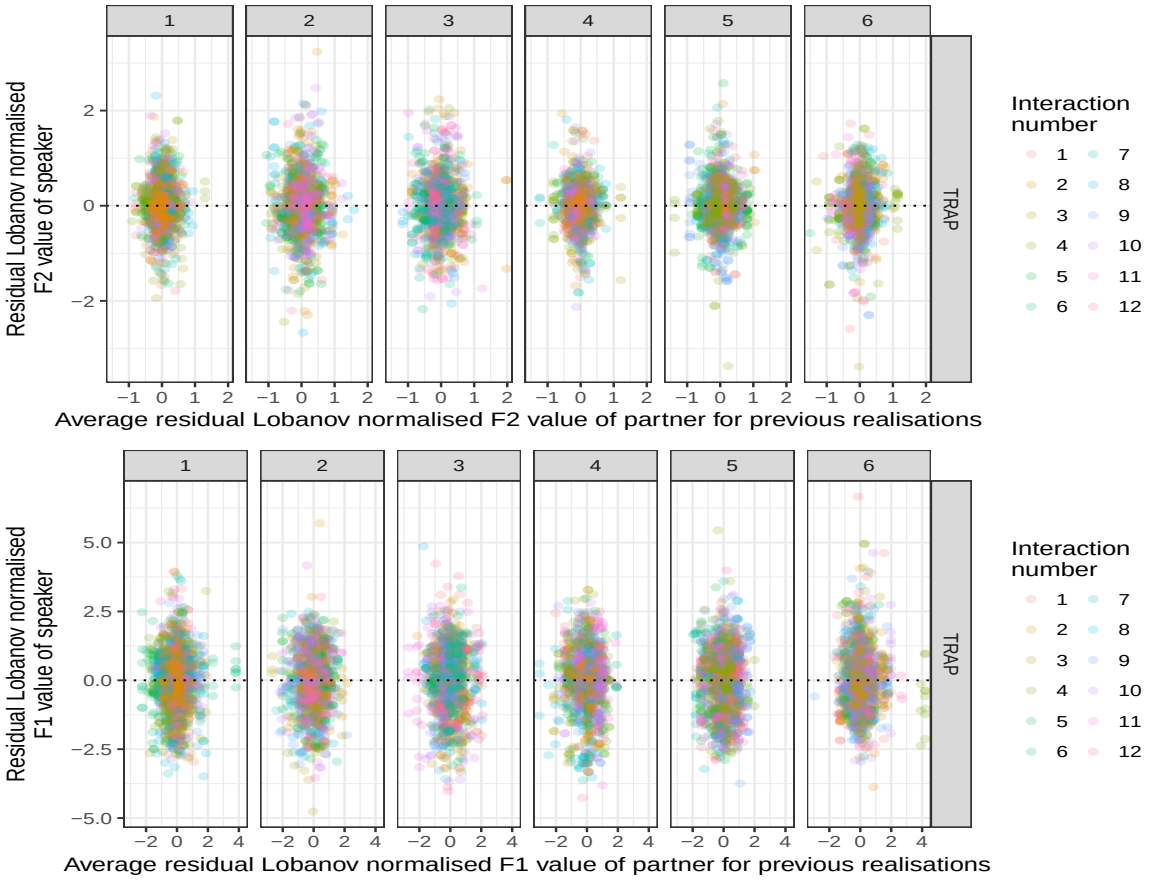



Figure 3.12: Scatter plot of the residual Lobanov normalised TRAP F1 values, where effects attributable to word and speaker have been removed. Panels represent the speaker pairs. There are no clear trends observable in the data.

Q1 There is no linear relationship between the recent F1/F2 TRAP realisations of the partner and the current F1/F2 TRAP realisation of the speaker.

Now the results of the GAMs are presented in order to answer questions 2 and 3. Figure 3.13 plots the results of the F1 and F2 difference GAM predictions against the presentation position, for each of the vowels considered (STRUT, THOUGHT and TRAP). The x-axis plots the presentation position and the y-axis plots the GAM predicted difference in Lobanov normalised F1 and F2. The columns dictate the vowel of interest and the rows dictate the source of the data, F1 or F2 (F2 on top). Colours represent the speaker pair. Dashed lines are the linear regression lines for the data in each panel, shaded areas are the standard error.

Looking at the plots presented in figure 3.13, it can be seen that the GAM predicted values of Lobanov normalised F2 across all vowels look to have a slight negative relationship with presentation position. This suggests that the difference in F2 between speakers decreases as the experiment proceeds. The results for F1, on the other hand, look to be mostly flat although there is perhaps some slight negative relationship for STRUT and perhaps some slight positive relationship for TRAP. In addition, the values for F1 in STRUT and TRAP have some extreme values that may be impacting on the results, although all data points are below 0.5 Cook's distance

| | *Dependent variable:* | |
|---|---|---|
| | TRAP F1 | TRAP F2 |
| Residual vowel partner value F1 | −0.005 | |
| | (0.017) | |
| Residual vowel partner value F2 | | 0.015 |
| | | (0.018) |
| Constant | −0.002 | 0.002 |
| | (0.013) | (0.006) |
| Observations | 8,005 | 8,005 |
| $R^2$ | 0.00001 | 0.0001 |
| Adjusted $R^2$ | −0.0001 | −0.00004 |
| Residual Std. Error (df = 8003) | 1.154 | 0.514 |
| F Statistic (df = 1; 8003) | 0.088 | 0.683 |
| *Note:* | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 | |

Table 3.7: Output of linear regression model comparing previous realisations for Lobanov normalised F1 and F2 of TRAP from a partner with the current realisations of Lobanov normalised F1 and F2 of TRAP of the speaker. Brackets report standard errors. The constant is the y-axis intercept of the regression line.

and are therefore within tolerance for a linear regression. The spread of the data for F2 doesn't appear to follow any pattern other than that previously described. F2 is spread reasonably consistently across the vowels with most pairs also demonstrating a broad usage of the F2 space. The same can be said for F1 across the vowels, with a few notable deviations, as mentioned. These plots suggest that there could be a weak relationship between the predicted difference in Lobanov normalised F2 and presentation position.

The results of the linear regressions performed on the GAM predicted difference between Lobanov normalised F1 and F2 values for all vowels against presentation position are presented in table 3.8. The results of the linear regressions show no significant relationships between presentation position and any of the dependent variables. However, the observations about the trends of F2 are validated by the larger $R^2$ values in the F2 columns, meaning that more of the variance for these dependent variables is explained by presentation position. However, these values still remain small and are not significant.

Taken as a whole, these findings answer question 2 for all of the vowels investigated:

Q2 The difference between partner and speaker F1/F2 realisations for STRUT, THOUGHT and TRAP within an interaction when modelled non-linearly, does not hold a relationship with the presentation position of the stimuli.

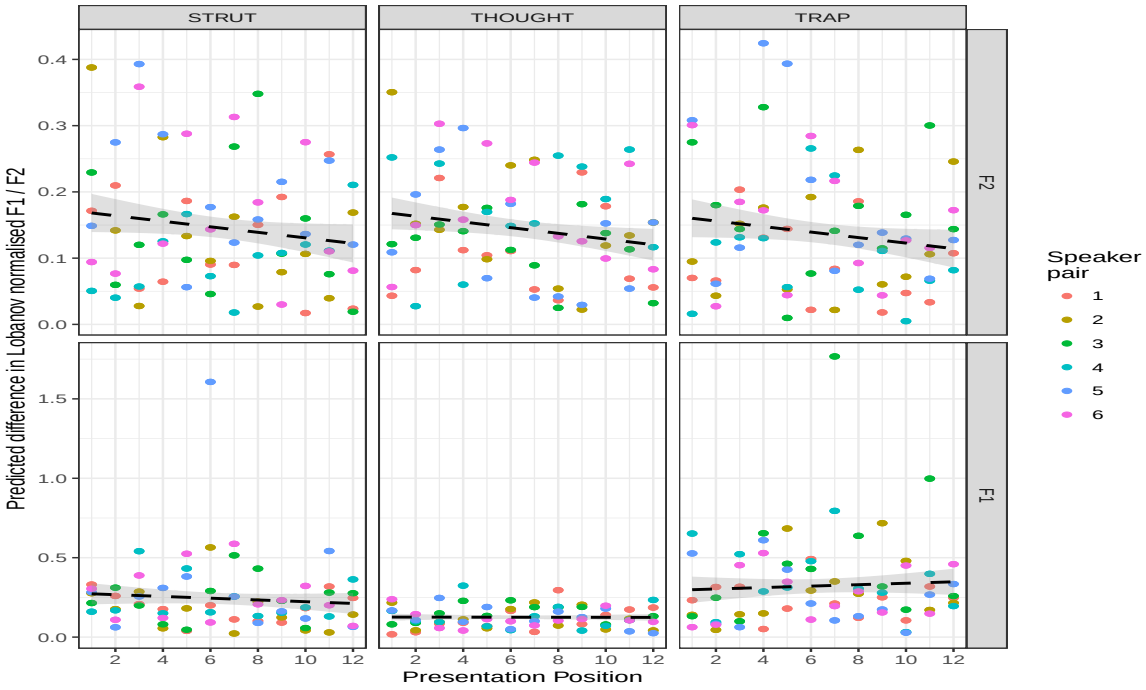



Figure 3.13: Scatter plot of the GAM predicted difference in F1 and F2 values for all vowels investigated as a function of presentation position. Columns indicate vowel type (STRUT, THOUGHT, TRAP) and rows indicate formant type (F1 or F2, F2 on top). Colours represent the speaker pair, the dashed line indicates the linear regression line and the shaded area is the standard error. Note that there is a difference in scale between F1 and F2.

Figure 3.14 follows the same format as figure 3.13 and plots the same data except against the length of the interactions rather than the presentation position.

| | *Dependent variable:* | | | | | |
|---|---|---|---|---|---|---|
| | STRUT: F1 | STRUT: F2 | THOUGHT: F1 | THOUGHT: F2 | TRAP: F1 | TRAP: F2 |
| Presentation position | −0.005 | −0.004 | 0.00004 | −0.004 | 0.005 | −0.004 |
| | (0.007) | (0.003) | (0.002) | (0.003) | (0.009) | (0.003) |
| Constant | 0.275*** | 0.171*** | 0.125*** | 0.172*** | 0.293*** | 0.164*** |
| | (0.054) | (0.023) | (0.017) | (0.020) | (0.067) | (0.023) |
| Observations | 72 | 72 | 72 | 72 | 72 | 72 |
| $R^2$ | 0.007 | 0.022 | 0.00000 | 0.037 | 0.004 | 0.024 |
| Adjusted $R^2$ | −0.007 | 0.008 | −0.014 | 0.023 | −0.011 | 0.011 |
| Residual Std. Error (df = 70) | 0.217 | 0.093 | 0.069 | 0.078 | 0.267 | 0.093 |
| F Statistic (df = 1; 70) | 0.489 | 1.583 | 0.0003 | 2.700 | 0.247 | 1.757 |
| *Note:* | | | | | | *p<0.1; **p<0.05; ***p<0.01 |

Table 3.8: Output of linear regressions performed on the GAM predicted difference between Lobanov normalised F1 and F2 values for all vowels against presentation position. Brackets report standard errors. The constant is the y-axis intercept of the regression line.

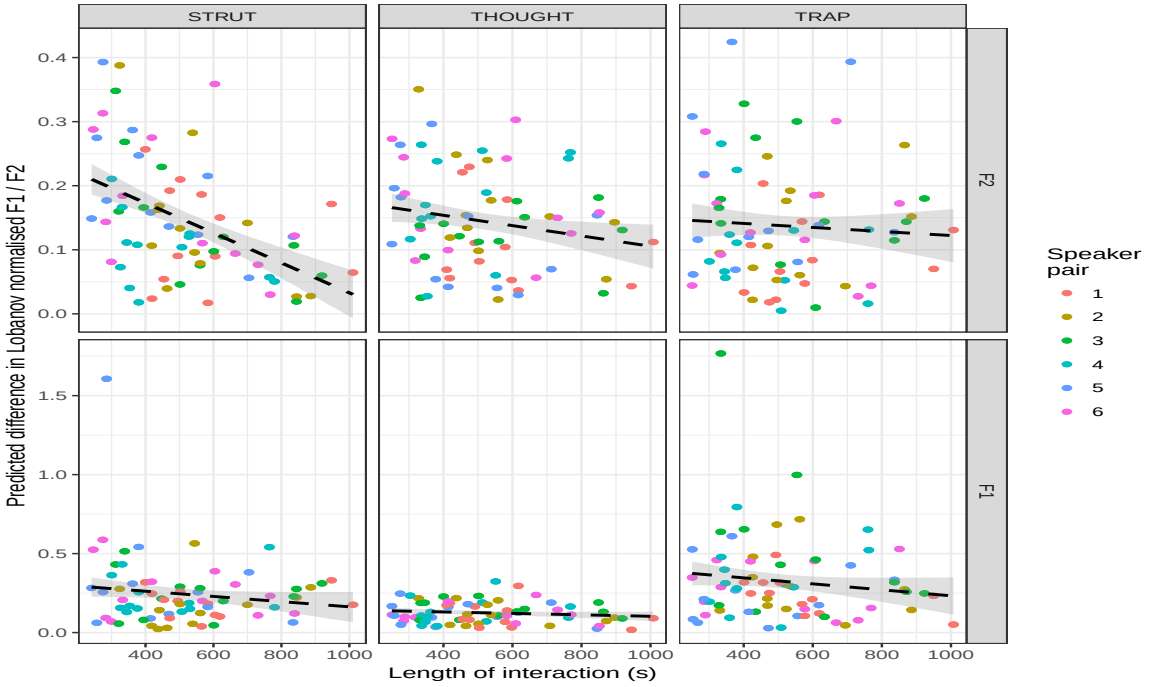



Figure 3.14: Scatter plot of the GAM predicted difference in F1 and F2 values for all vowels investigated as a function of interaction length. Columns indicate vowel type (STRUT, THOUGHT, TRAP) and rows indicate formant type (F1 or F2, F2 on top). Colours represent the speaker pair, the dashed line indicates the linear regression line and the shaded area is the standard error. Note that there is a difference in scale between F1 and F2

The data in figure 3.14 show the same general trends as that found in figure 3.13 for the linear regressions against presentation position. However, the trends appear to be more pronounced for most vowels and their formant values, with the exception of F2 for THOUGHT and TRAP. The largest relationship clearly looks to be for F2 in STRUT, suggesting that the difference between speakers for that variable decreases in interactions that lasted longer. The same kind of trend can be seen in all other

vowels for both F1 and F2 although to lesser degrees with some, such as F1 for THOUGHT and F2 for TRAP appearing quite markedly less so. These results would suggest that there may be some relationship between these dependent variables and interaction length, although for most any relationship is likely to be weak.

| | *Dependent variable:* | | | | | |
|---|---|---|---|---|---|---|
| | STRUT: F1 | STRUT: F2 | THOUGHT: F1 | THOUGHT: F2 | TRAP: F1 | TRAP: F2 |
| Interaction length | −0.0002 | −0.0002*** | −0.00005 | −0.0001 | −0.0002 | −0.00003 |
| | (0.0001) | (0.0001) | (0.00004) | (0.00005) | (0.0002) | (0.0001) |
| Constant | 0.331*** | 0.266*** | 0.149*** | 0.185*** | 0.421*** | 0.153*** |
| | (0.073) | (0.028) | (0.023) | (0.027) | (0.090) | (0.032) |
| Observations | 72 | 72 | 72 | 72 | 72 | 72 |
| $R^2$ | 0.024 | 0.238 | 0.017 | 0.038 | 0.019 | 0.004 |
| Adjusted $R^2$ | 0.010 | 0.227 | 0.003 | 0.024 | 0.005 | −0.010 |
| Residual Std. Error (df = 70) | 0.215 | 0.082 | 0.069 | 0.078 | 0.265 | 0.094 |
| F Statistic (df = 1; 70) | 1.712 | 21.819*** | 1.190 | 2.740 | 1.352 | 0.287 |
| *Note:* | | | | | | *p<0.1; **p<0.05; ***p<0.01 |

Table 3.9: Output of linear regression performed on the GAM predicted difference between Lobanov normalised F1 and F2 values for all vowels against interaction length. Brackets report standard errors. The constant is the y-axis intercept of the regression line.

Table 3.9 presents the results of the linear regressions performed on the GAM predicted difference between the Lobanov normalised F1 and F2 values for all vowels against interaction length. These results show one significant relationship between the F2 of STRUT and interaction length, $R^2 = 0.238, F(1,70) = 21.819, p = 0.00001, r = 0.487$.This is in line with what was observed in figure 3.14 for the F2 of STRUT. It is an interesting finding because it is not what was expected from the predictions based on sound change and socially sensitive vowels in Glaswegian. No other significant relationships were found in this data. This might tentatively suggest that the STRUT vowel might be on the verge of changing. However, given the very small effect size for STRUT, this cannot be said for certain. The general trend across the vowels is that there is no relationship with interaction length.

These data provide an answer to question 3:

Q3 The differences between partner and speaker F1/F2 realisations for STRUT, THOUGHT and TRAP within an interaction when modelled non-linearly, do not generally hold a relationship with interaction length. However, there is a small, tentative relationship between the F2 of STRUT and interaction length.

### Speech Rate

The results for speech rate, as extracted based on syllable counts (see subsection 3.5.1) are considered. Unlike the results for VOT and the vowel formants, there is only one measure extracted for speech rate. As such, the results for each of the key questions can be answered sequentially. To recap, the questions that are being asked here are:

Q1 Is there a linear relationship between the recent speech rate of the partner and the current speech rate of the speaker?

Q2 When the difference between partner and speaker speech rate within an interaction is modelled non-linearly, is there a relationship with the presentation position of the stimuli?

Q3 When the difference between partner and speaker speech rate within an interaction is modelled non-linearly, is there a relationship with interaction length?

Figure 3.15 presents the data for the residual speech rate values of the participants. The plot shows the average residual speech rate of the partner's recent utterances on the x-axis and the speaker's current residual speech rate on the y-axis. As with previous speech measures, the residuals are the result of a MEM accounting for speaker, word and DiapixUK task number as random effects. The plot is further subdivided into each of the participant pairs, each panel represents one participant pair. Colours represent the interaction from which the data were drawn.

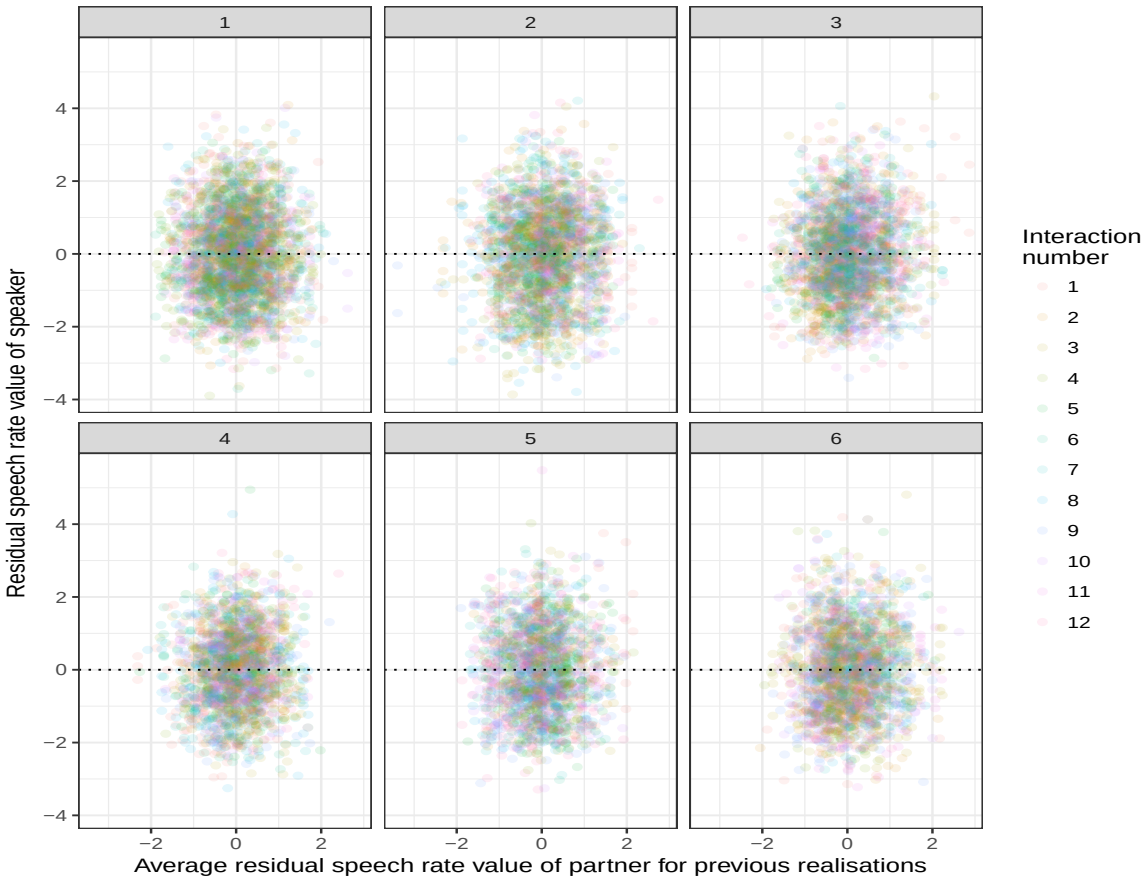



Figure 3.15: Scatter plot of residual speech rate values, where effects attributable to word, speaker and DiapixUK task number have been removed. Panels represent the speaker pairs. There are no clear trends observable in the data.

As with the previous measures of VOT and vowel F1/F2 values, these residual speech rate values show no clear trends or relationships between the recent speech rate of the speech partner and current speech rate of the speaker. To test if this

observation holds, a linear regression was performed on the data, the results of this can be found in table 3.10.

| | *Dependent variable:* |
|---|---|
| | Residual speaker speech rate values |
| Residual partner speech rate values | $0.058^{***}$ |
| | (0.014) |
| Constant | $-0.0001$ |
| | (0.010) |
| Observations | 14,287 |
| $R^2$ | 0.001 |
| Adjusted $R^2$ | 0.001 |
| Residual Std. Error | 1.185 (df = 14285) |
| F Statistic | $17.710^{***}$ (df = 1; 14285) |
| *Note:* | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 |

Table 3.10: Output of linear regression model comparing previous speech rate from a partner with that of the speaker. Brackets indicate standard errors. Speech rate values are provided in syllables per second.

The results of the linear model show that there is a significant relationship between the recent average speech rate of the partner and the current speech rate of the speaker, $R^2 = 0.001, F(1, 14285), p = 0.0009, r = 0.035$. Whilst this finding could indicate that as the local speech rate of the partner increases, so too does the speech rate of the speaker, this seems unlikely given the small $R^2$ value. Looking at the value and significance of the constant provides some further evidence and context for this finding. Generally speaking, it is somewhat meaningless to interpret the constant associated with a linear regression. The constant indicates the point at which the regression line intersects the y-axis and for most analyses this point of intercept would be meaningless. Consider the example of data collected on age and height, imagine these data plotted on a graph where the x-axis represents age and the y-axis represents height. Running a linear regression on data collected in that sample would likely produce a negative constant since the relationship with height and age is positive, as one's age increases, so does one's height (at least initially). At age zero, it is not possible to measure height so the regression line is not likely to pass through the origin. In addition, the relationship between age and height is not one-to-one, humans do not grow at a constant rate of one inch per year, for example. Thus, it is likely that the constant will have a negative value. It is not possible for a human to have a negative height, therefore the constant is meaningless but it is still required to compute a regression. Now consider the data presented for speech rate. Because both axes on the graph have meaningful negative and positive values, the point at which the regression line intercepts the y-axis does have mean-

ing. The values represent the residuals generated by the local speech rate of the partner and residuals of the the current speech rate of the speaker, both corrected for random factors of word, speaker and DiapixUK task number. These values have meaningful interpretations both above and below the zero line. A plot of all speech rate data used in this regression, along with the linear regression line is presented in figure 3.16 in order to aid visualisation. Recall that the residuals are the result of a MEM and thus the negative values for speech rate represent the difference between predicted and observed values, hence the negative values for speech rate.

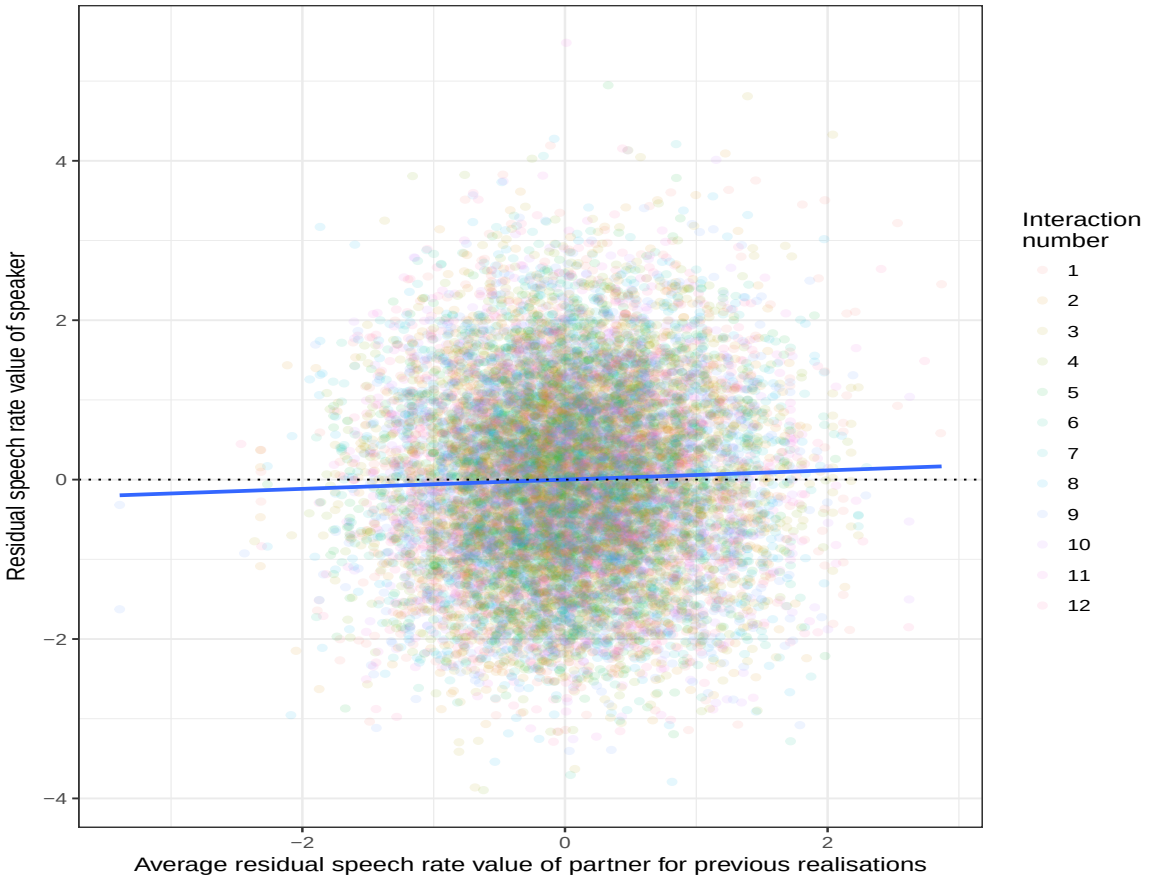



Figure 3.16: Scatter plot of residual speech rate values, where effects attributable to word, speaker and DiapixUK task number have been removed. The blue line is the regression line.

Given that the constant represents the point at which the regression line intersects the y-axis, the p-value associated with it can be interpreted as an indication of the validity of the null hypothesis that the intercept differs significantly from zero. In this case, the constant is non-significant. This means that the null hypothesis that the intercept with the y-axis does not differ from zero cannot be rejected. As such, it is possible that the actual trend in the data may not differ from zero, meaning that the relationship between the residuals of the two speakers may not hold more generally. Effectively, the result in the linear regression for this data can be seen as a strong but unreliable effect.

In answering the first question for the speech rate data, the answer would be a

somewhat nuanced one:

Q1 There may be a linear relationship between the recent speech rate of the partner and the current speech rate of the speaker but for this dataset the effect remains unreliable.

Looking for a non-linear trend in the speech rate data takes the same form as previous evaluations of non-linear trends in this section. The predicted differences in speech rate are produced by submission of the data to a GAM, predicting the difference in speech rate between speakers over the course of an interaction. This is plotted against the presentation position of the DiapixUK tasks in figure 3.17. The y-axis shows the predicted difference in speech rate and the x-axis provides the presentation position of the DiapixUK task. Colours represent the speaker pair and the dashed line represents the linear regression with the grey shaded area representing the standard error.

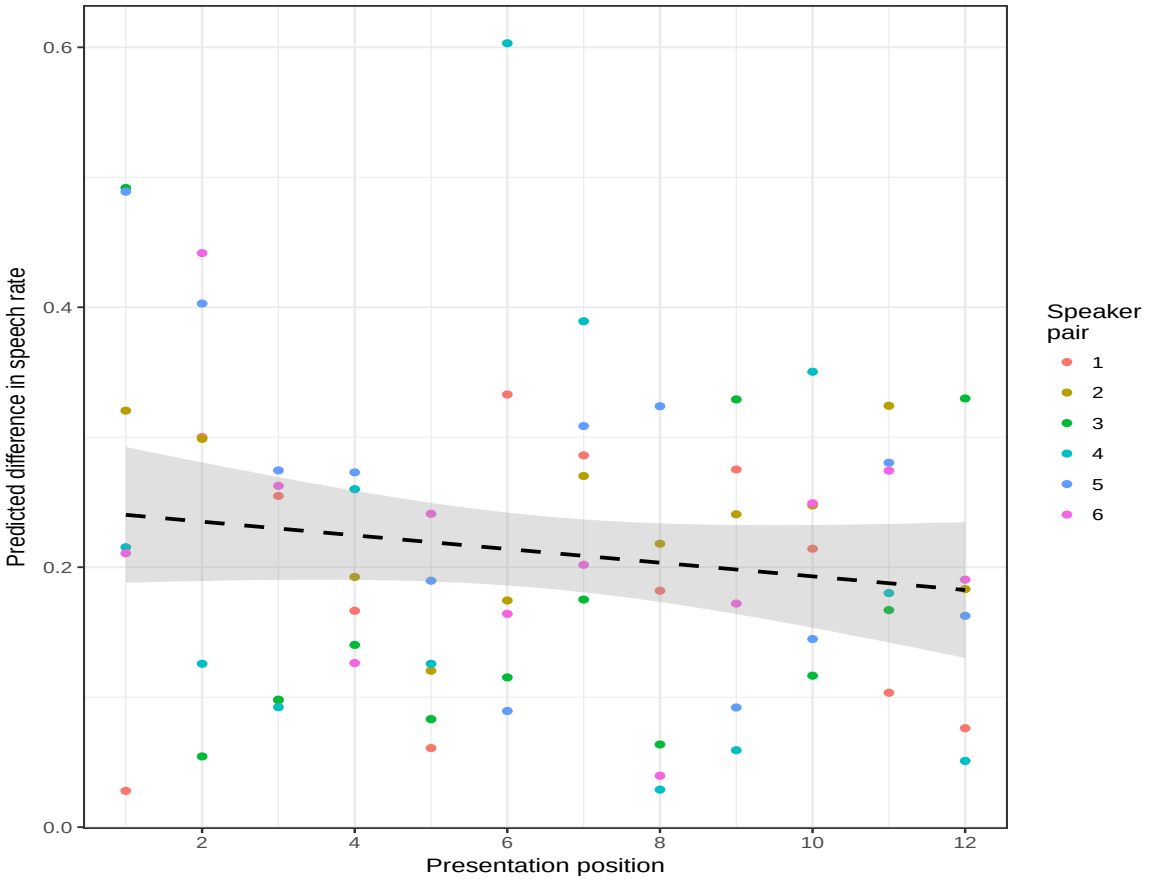



Figure 3.17: Scatter plot of the GAM predicted difference in speech rate values (syllables per second) as a function of presentation position. Colours represent the speaker pair, the dashed line indicates the linear regression line and the shaded area is the standard error.

The data in figure 3.17 demonstrate a negative trend, where the difference in speech rate between the speakers decreases as the experiment progresses. To test this, a linear regression was run on the results of the GAM against the presentation position. The results of this test are presented in table 3.11.

| | *Dependent variable:* |
|---|---|
| | Speech Rate |
| Presentation position | −0.005 |
| | (0.004) |
| Constant | 0.246*** |
| | (0.030) |
| Observations | 72 |
| $R^2$ | 0.024 |
| Adjusted $R^2$ | 0.010 |
| Residual Std. Error | 0.118 (df = 70) |
| F Statistic | 1.701 (df = 1; 70) |
| *Note:* | *p<0.1; **p<0.05; ***p<0.01 |

Table 3.11: Output of linear regression performed on the GAM predicted difference between speaker speech rates against presentation position. Brackets provide standard errors.

The results of the linear regression demonstrate no significant effect of presentation position on the difference between the speech rate of the speakers during an interaction. Since the GAM has integrated time during an interaction into the values used in this analysis, this finding can be said to be running contrary to the findings presented for the effects of a partner's local speech rate on a speaker's speech rate. Further to this, GAMs tend to be good at accounting for variability in the data leaving only the largest and most robust effects intact. Given that the finding for local speech rate demonstrated a weak effect and may not be reliable, the lack of a result here is not wholly unexpected.

These results allow for an answer to be provided for the second question regarding speech rate:

Q2 The difference between partner and speaker speech rate within an interaction, when modelled non-linearly does not have a statistically significant relationship with the presentation position of the stimuli.

The final question for speech rate, and for the phonetic analyses, concerns the relationship between non-linearly modelled speech rate differences and interaction lengths. The same GAM predicted data used in the assessment of the relationship between differences in speech rate and presentation position was used to evaluate a relationship between differences in speech rate and interaction length. This data is presented in figure 3.18.

A similar trend to that seen in figure 3.17 can be seen in figure 3.18. There is a slight negative relationship between the length of interaction and the GAM predicted differences in speech rate. This would suggest that as interactions got longer,

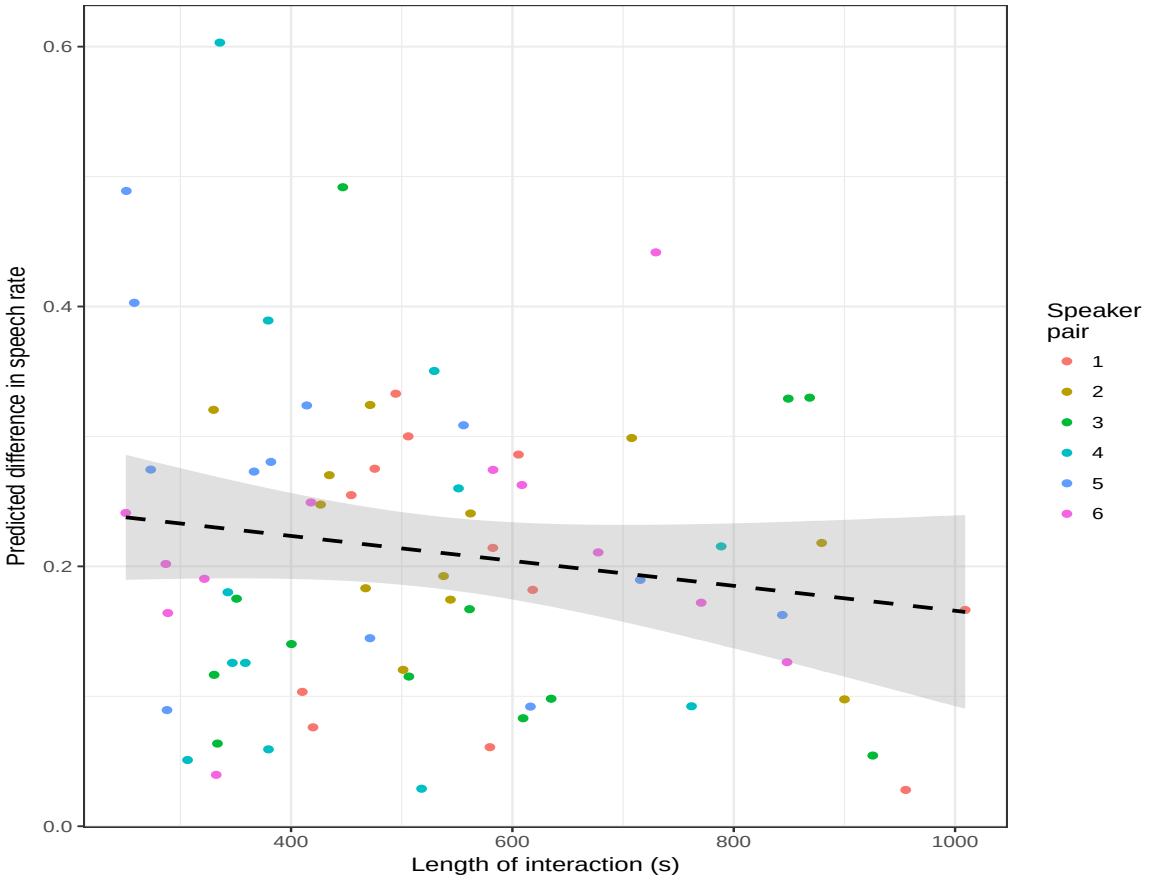



Figure 3.18: Scatter plot of the GAM predicted difference in speech rate values (syllables per second) as a function of interaction length. Colours represent the speaker pair, the dashed line indicates the linear regression line and the shaded area is the standard error.

meaning that the participants were finding the task more difficult, the difference between the speech rates of the speakers decreases. In other words, it suggests that speakers converge when they find a task more difficult. To test if this effect was meaningful, a linear regression was performed, comparing the GAM predicted differences against interaction length. These results can be found in table 3.12

These results show that the trend seen in figure 3.18 is not significant. The results of the statistical analyses show that although there may be a subtle trend of convergence in longer interactions, it is not a significant one.

With these results, the third question pertaining to speech rate can be answered:

Q3 The difference between partner and speaker speech rate within an interaction, when modelled non-linearly, does not show a statistically significant relationship with interaction length.

| | *Dependent variable:* |
|---|---|
| | Speech Rate |
| Interaction length | −0.0001 |
| | (0.0001) |
| Constant | 0.262*** |
| | (0.040) |
| Observations | 72 |
| $R^2$ | 0.025 |
| Adjusted $R^2$ | 0.011 |
| Residual Std. Error | 0.118 (df = 70) |
| F Statistic | 1.788 (df = 1; 70) |
| *Note:* | *p<0.1; **p<0.05; ***p<0.01 |

Table 3.12: Output of linear regression performed on the GAM predicted difference between speaker speech rates against interaction length. Brackets indicate standard errors.

### 3.5.4 Discussion

Interpretations of the results presented in subsection 3.5.3 are now presented. These are first considered in turn, in the same order that the results were presented in (VOT, vowels, speech rate), before considering the results of the phonetic analyses as a whole. Recall that for each of the phonetic variables that were investigated, there were three key questions being asked:

Q1 Is there a linear relationship between the recent realisations of the partner and the current realisation of the speaker?

Q2 When the difference between partner and speaker realisations within an interaction are modelled non-linearly, is there a relationship with the presentation position of the stimuli?

Q3 When the difference between partner and speaker realisations within an interaction are modelled non-linearly, is there a relationship with interaction length?

#### VOT

The findings for VOT demonstrate only some evidence for accommodation being detected, this was true for both voiced and voiceless VOT. The results demonstrated no significant results for the effect of the partner's previous VOT on the speaker's current VOT and there was no significant effect of presentation position on VOT. There was however, a significant effect of interaction length on VOT. The findings

provide support for the notion that accommodation is a subtle phenomenon that has a non-linear relationship with both time and for certain features of speaker context.

If a trend had been seen in the initial linear model based on the residuals of the MEM, then two things could have been concluded, (1) that the local VOTs of the partner share a relationship with the current VOTs of the speaker and (2) that this relationship demonstrates a clear linear trend. As it stands, the results do not show this. So what must be concluded instead is either that there is no relationship between the recent VOT realisations of the partner and the speaker or that the tools used to detect the effects are not sensitive enough. Having said that, an alternative interpretation could be that the method used to sample the previous VOT values of the partner might not have been appropriate. It might be the case that the number of VOTs that was used to calculate the mean value for recent partner VOT could have been drawn from words that were somewhat distanced in time. If this was the case then it could have led to a good degree of variance in the mean VOT values for the partner, thus leading to a greater spread of values in the data. Having said this, given the number of observations for both voiced and voiceless VOT, $8,835$ and $5,659$ tokens respectively, this variance in mean partner VOT would have needed to be highly consistent, which seems unlikely.

Another potential explanation for the non-significant finding might be that accommodation is phoneme dependent. What is meant by this is that the analysis that was performed simply evaluated all voiced and all voiceless plosives, it did not distinguish by the phoneme of each plosive (ie. /b/, /d/, /g/, /p/, /t/ and /k/). If the participants are accommodating to mental representations of phoneme categories rather than the global features of voiced or voiceless plosives, then any effects might have been normalised out by grouping based on voicing type. In order to address this, a further analysis separating out the data by phoneme may prove insightful. However, if the broad effects cannot be seen at the group level then it suggests that the phenomenon of interest is considerably more subtle than originally thought. It may prove too fine grained to be captured with standard techniques when assessing a continuous stretch of speech. Having said that, accommodation in VOT has been demonstrated in a restricted setting by Shockley et al. (2004) (see subsection 2.2.2) where the length of VOTs were shown to be imitated. This suggests that whilst VOTs may be imitated generally, they may not be distinguishable from the ongoing variation that is present within a speaker's own productions.

The second finding in the results for VOT also showed no significant effects. For these findings, the independent variable was the presentation position of the stimulus and the dependent variable was the GAM predicted difference between the VOTs of speakers within an interaction, separated by voicing type. For these results, time as represented across an interaction had been integrated into the data values and normalised by use of a GAM to predict the values against a $1,000$ point time series. The data had also been modelled non-linearly to more accurately capture

the ongoing speaker trends for each interaction. The expected trend would have been towards convergence as the experiment progressed, meaning that a negative coefficient for the regression would have been expected. Since there were no significant results, this finding indicates that more time spent with someone does not necessarily lead to greater convergence for certain features.

The findings in relation to question three provide the only significant results for VOT. They demonstrate a small but significant negative regression coefficient for both voiced and voiceless VOT. This finding suggests that longer interactions tended to produce VOT realisations from the speakers that were closer to each others realisations than when engaged in shorter interactions. Recall that for these evaluations, the time dimension within an interaction has been normalised. When considered in terms of the function that a longer interaction length has on the task, it could be interpreted that participants are demonstrating more convergence in interactions that they are finding the hardest. This would seem reasonable given that the only reason that any given interaction would take longer is if the participants had not found all twelve of the differences between the target stimuli.

In the results section, these two analyses were presented together. This was because they present two different windows on the same data. Whereas the analysis with presentation position looks at the overall time that the participants spent together, the analysis using interaction length aims to capture more of what participants were doing during the tasks. Comparing these two analyses, it would seem that what participants were doing during an interaction appears to have more of an effect on the predicted difference in VOT than overall time spent together.

Taken as a whole, the results for VOT provide tentative support for this feature of accommodation being linked to the context in which a person is engaged.

#### Vowels

For the analysis of vowel accommodation, three vowels were selected based on the nature of variability of these vowels observed by previous sociolinguistic apparent and real-time studies of Glaswegians (Macaulay, 1976; Stuart-Smith, 2004; Lawson et al., 2013). STRUT was selected to represent a stable vowel category where no shift was expected, THOUGHT was selected to represent a diachronically changing vowel where some shift might be expected and TRAP was selected to represent a vowel that could be used to indicate social factors, this vowel type was expected to shift the most. Results of the analyses demonstrated little evidence for any adaptation for any of the vowels that were expected to shift in either the F1 or F2 dimensions. The only vowel that demonstrated any significant adaptation was the vowel that was predicted to be most stable, STRUT. As such, the answers to the first two key questions were that there was no measurable relationship whilst for the third key question the answer was that generally there were no relationships except for the F2 of STRUT.

As with the findings for VOT, the first two analyses present no significant results. This lack of significance is consistent across all of the vowels analysed for the first two analyses. What this demonstrates is that any relationship between recent F1 and F2 productions for each vowel by the partner is not likely to be linear or particularly strong. Further to this, any relationship between differences in F1/F2 production for vowels and overall time is also unlikely to be strong. Sensible precautions were taken to eliminate formant values during transitions that were known to be either susceptible to outside influence from the surrounding phonetic context but perhaps this might play a role in accommodation during a live interaction. It might be the case that the vowels a speaker hears are affected by outside factors from the surrounding phonetic context and they are embedded in the representation of the vowel to which the speaker is attempting to accommodate towards. Effectively, it may be the case that speakers do not accommodate to canonical prototype vowels but that they might bring the context in which the vowel is uttered into their accommodative target. Having said that, evidence from studies such as Purnell (2009) and Bailly, Lelong, et al. (2010) have shown reasonably clear effects of vowels being accommodated towards during an interactional setting. It would have been interesting to have performed a pre and post task on the participants so that an assessment could have been made of the differences found in vowel realisations when sampled from a continuous engagement and when sampled from a more restricted engagement. An evaluation such as this would have allowed for a clearer comparison with other studies.

Further to considerations surrounding the nature of the accommodative objects that speakers use to adapt their behaviour towards, it is also worth considering the way in which the vowels were evaluated. It may be the case that breaking the vowels down into their first two formants and then comparing only like with like (ie. F1 of partner with F1 of speaker) lost an element important to accommodation. For instance, it could be the case that some element of F2 in TRAP produced by the partner had an influence on the F1 production in the speaker. Equally, it could be possible that higher formants play a role in accommodation. An ability to capture the vowel as a human hears or interprets it would perhaps allow for a keener insight into the patterns of vowel realisation that contribute to accommodation. This could have been achieved through a Bark (Smith & Abel, 1995) or Mel (Tokuda, Kobayashi, Masuko, & Imai, 1994) transformation of the data prior to normalisation. Additionally, a further analysis could be employed to look at the relationships between different formant values within the same vowel. However, this still would not provide a whole understanding of the interaction between the production of vowels and the features which are used to integrate accommodative movements into production.

As mentioned above, the only vowel to present a significant result was STRUT, in the F2 dimension (ie. the front-back dimension). This result was only found

in relation to the interaction length. This vowel's F2 was found to demonstrate a significant tendency towards convergence when interactions were longer. It is interesting that the only vowel to show an effect was the one vowel that is thought to be the most stable in the Glaswegian accent. Perhaps it was the case that since the other vowels evaluated were able to vary more freely, they did vary but only to a small degree in both directions and the effects cancelled each other out. This would have then only been compounded by the fact that the difference values used in the analysis were absolute. Within a system such as accommodation where adaptations can be made towards or away from another speaker, the use of absolute differences may prove to be too coarse. The use of absolute differences collapses maintenance and complementarity into the same category since the total difference would remain the same across an interaction. However, these two categories of accommodation actually represent different behavioural trends. If vowels are being used adaptively throughout an interaction, a measure that separates the different accommodative behaviours would be needed. Having said that, this collapse of accommodation categories does not explain why the F2 of STRUT was found to be significant.

It might be the case that the shifting observed in the STRUT vowel could be linked to instability and change in the THOUGHT vowel, which is rising. This would be consistent with a classic pull chain in the vowel space. Vowels move up the back dimension, and central vowels are then pulled back. As a result of this pull chain the STRUT vowel would become backer.

The results for the vowels demonstrate that even with longer domain features that allow for more movement, accommodation still remains subtle and nuanced. Whether this is the result of speakers using subtle variations within a vowel category to display accommodation or whether it is due to an inability in the methods used to capture accommodation in vowels remains to be determined. In either case, there is little evidence that standard approaches to the measurement and interpretation of vowels are able to detect accommodation in a continuous interaction.

#### Speech rate

Speech rate is somewhat of a different measure to the VOT and vowel measures. This is because rather than making specific observations about small sections of speech, it is providing a generalisation across a comparatively large section of speech data. Speech rate is calculated by dividing the total time of an utterance, in seconds, by the number of syllables in that utterance. The inclusion of this measure should offer a window onto accommodative behaviour at a broader scale than that of the VOT and vowel measures. Indeed, it is the only phonetic variable to show an accommodative effect in the initial models. Although there might be reason for a cautious interpretation of these results (as mentioned in subsection 3.5.3), the fact that a significant relationship was found between the recent speech rate of the partner and current speech rate of the speaker suggests that speech rate might be capturing some

broad accommodative trends. These results may perhaps be reflected in the visual trends for both presentation position and interaction length, but were not significant in the GAM based modelling.

For both of the analyses that made use of the GAM predicted differences in speech rate, the independent variable was a time dependent measure. What is meant here is that both presentation position and length of interaction provide measures that represent the time during the experiment in some way. For presentation position, it is the overall time of the experiment. For interaction length, it is the length of time taken by the participants to complete a particular task. It may be the case that because speech rate is also a measure that is dependent on time, at a greater scale than VOT, and because it samples more of the data (every utterance), the element of time that would be predictive is not present in the GAM based analyses. This would explain why a general comparison of speech rate against the previous speech rate of the partner provided a significant result whilst the GAM analysis did not. Applying an analysis to speech rate within interactions and across the experiment may provide some indication as to the validity of this claim but comparisons between tasks and experiments would be difficult due to time not being normalised across all conditions.

The results for speech rate are not completely clear. In terms of the key questions asked of speech rate, there are some weak suggestions for a general tendency towards convergence over local stretches of engagement. Having said this, the effect size is small and there is additional evidence to suggest that the results may not be reliable. In other words, the answer to question one is that there may be an effect but it cannot be conclusively determined from this analysis. However the answers to questions two and three are somewhat more complicated. There may be some interaction between the different time scales involved in the GAM analyses that cause any visible trends found in the original linear regression to be muted. Ultimately, the results presented for speech rate must be taken as a potential indicator of some accommodative effects.

### General discussion

Individually, each of the phonetic variables that have been investigated here show different effects in response to local realisations of the partner, the presentation position of task stimuli and to interaction length. The significant results for analyses were:

1. Interaction length having a significant effect on the GAM predicted differences for both voiced and voiceless VOT.
2. Interaction length having a significant effect on the F2 of the STRUT vowel.
3. The previous speech rate of the partner having a weak effect on the current speech rate of the speaker.

However, all of the phonetic variables that were investigated did demonstrate at least some evidence of accommodative behaviour but the sizes of these responses were consistently small with little support for accommodation at a broad scale. These small effects may be due to a number of factors. It may be the case that by sampling the speech data and extracting specific segments such as those related to VOT or vowels, the whole accommodative effect is not being captured. Sampling the data over more consistent spreads of speech may prove to be more insightful than taking slices of speech data at irregular intervals since they would provide better coverage of the speech data as a whole. Indeed, the method of predicting the phonetic variable against a 1000 point time series using the GAMs as predictors was an attempt to account for the irregularity in measurement spread over the course of in interaction. This may have had some success in its implementation given the significant results for VOT and the F2 of STRUT. On the other hand, the weak coefficients associated with the results suggest either that even when accounting for the irregular spread of phonetic measures the accommodative effect is still small or that the use of GAMs and predicted time series in this way lacks the sensitivity to properly observe the effect. Whilst these findings do provide some findings relating to the detection of accommodation in a continuous interaction, the fact that the findings are a little sporadic means they also offer some support for an evaluation of different approaches.

A second possible reason for the lack of significance in the results concerns the evaluation of single phonetic variables in isolation. Studies have demonstrated that speakers can use different elements of their phonetic repertoire to accommodate in relation to their speech partner (see section 2.2). Whilst there may be some specific instances where the use of a particular phonetic feature indicates a specific social goal, a more global measure of accommodation should be able to capitalise on all available information pertaining to accommodation. Accommodation in the speech signal is likely to involve not only multiple phonetic features but also interactions between those features. Observing the movements of a single phonetic feature may be akin to observing the movement of a single starling when trying to interpret a murmuration of starlings. Some information about the whole form will most certainly be captured but the full effect is not truly observable without a broader scope. In order to understand global trends, a different perspective on the phenomenon of interest (whether it be murmurations or accommodation) must be taken. The results of the data analysis performed here supports this view. The lack of detection of strong accommodative effects provides some motivation for the use of a more holistic measurement approach to detecting accommodation.

With regards to the results of the linear regressions performed on the residual values of the MEMs, speech rate was the only phonetic feature that produced a significant response. Although reasoning for applying caution when interpreting this result was provided in subsection 3.5.3, if it is assumed that the result does repre-

sent a real effect, this raises some interesting questions. Speech rate was the only phonetic feature that accounted for all of the available speech data. What is meant by this is that all of the utterances that were produced by the speakers were used to calculate the speech rate. Whilst the speech rate that is then calculated might be a rather coarse measure, it is still a measure that is likely to have more consistent intervals than that of VOT or the vowels. In addition, the length of time over which the partner's local speech rate is calculated is likely to be a better representation of the recent context of speech since there is less potential for the samples to be extracted from parts of speech that are concerned with different conversational context. Both of these things taken together may have been enough to allow for the initial analysis of speech rate to return as significant where VOT and vowel analyses did not. Further to this, the fact that speech rate accounts for far larger sections of speech than VOT or vowel measures may have played a role in the non-significant results for the liner regressions of the GAM predicted, time normalised values.

When the values for speech rate are predicted from the GAM for each of the speakers over the $1000$ point time series, it may be the case that the curves representing the model are more likely to be flat because of smaller differences between consecutive speech rate measures. This would then manifest as smaller differences in predicted speech rate between the speakers. Looking at the results presented in figures 3.17 and 3.18 it can be seen that the differences in speech rate are indeed low. The maximum difference between speakers can be seen to be $0.6$ syllables per second (although this value is in itself somewhat extreme), the mean speech rate on the other hand is $4.42$ syllables per second. This would place the greatest difference between speaker speech rate at roughly half of a standard deviation from the mean, this is also true when standard deviations are considered within participant pairs. Whilst the degree of difference in speech rate that is necessary for participants to detect and adapt to accommodative behaviours may not be known, it would seem that changes as small as these over a stretch of time as large as an utterance or a sequence of utterances would not be sufficient for accommodation unless the effect was cumulative in some way. Although this is speculation, given that larger differences were found in shorter term features such as VOT, it seems unlikely that changes in a long domain feature such as speech rate would be so small. Explanations for this could take a number of forms. It might be the case that the DiapixUK task lends itself either to highly consistent speech rate levels or to levels that vary but that ultimately cancel each other out (ie. speeding up and slowing down). Alternatively, it might suggest that accommodation in speech rate does not take place in this data (which doesn't seem likely given previous research (Casasanto et al., 2010) and the results of the initial linear regression). However, a more likely answer would be that the use of time normalisation during the GAM based analyses eliminated some of the speech rate effect. An alternative approach would be not to time normalise for speech rate, but then results between phonetic variables would

be less comparable.

One consistent finding across all of the phonetic variables is that interaction length in the GAM based analyses was a better predictor of the difference between the phonetic variables of speakers than presentation position. Across all of the analyses performed on the GAM predicted data, the $R^2$ value tends to be higher for interaction length than that for presentation position. This suggests that the context of the interaction explains more of the variance in the data than the overall amount of time spent engaging with a speech partner. Although most of these results are non-significant, they may be an indication of a subtle trend that is not detectable with the statistical methods employed here. It provides some tentative suggestions that a more holistic investigation of the data might expect to find greater convergence during longer interactions.

In general, the findings of this experiment provide only weak support for the detection of accommodation in a continuous interaction when focussing on phonetic segments using more standard statistical approaches. Whilst the methods used in this analysis have aimed to utilise available statistical tools in a manner appropriate to the task at hand there will naturally be alternative approaches to the interpretation of this data. Indeed, there may be more appropriate tools for analysis of this kind of data that have not been explored in this thesis. Further exploration of statistical tools for extracting the appropriate information from continuous speech data is encouraged. However, this thesis chooses to turn to machine learning in order to better understand the continuous nature of accommodation. The following section, section 3.6, concerns the application of HMM based, machine learning approaches to this data.

## 3.6 Computational Analysis

It is the aim of this section to demonstrate an alternative approach to the detection and interpretation of accommodation during a continuous interaction to that explored in section 3.5. The findings of the phonetic analysis employed in section 3.5 were, for the most part, inconclusive. They did however provide a tentative suggestion that accommodation might be related to the context of the interaction between speakers by way of a relationship with interaction length. It was also noted that the phonetic analyses employed were not sufficiently able to distinguish between the various types of accommodation that were detailed in subsection 2.1.1 (convergence, divergence, complementarity and maintenance). Further to this, the role that time played in the analyses proved to be a stumbling block when interpreting results. Taking this forward, this section aims to evaluate if a HMM based approach might be able to overcome some of these shortcomings and provide a potential solution to detecting and evaluating accommodation in a continuous interaction.

Recall that HMMs are able to characterise the general form of a continuous signal by modelling the joint probability distributions between the unknown or *hidden* states and the observable data. If an HMM is constructed for a particular speaker, it can then be used to determine the likelihood that a particular word or speech segment was produced by that speaker. However, in order to construct an HMM for a speaker, the data stream needs to be turned into a low-dimensional sequence of numbers, known as vectorisation. Here vectorisation is completed through transformation into Mel frequency cepstral coefficients (MFCCs). Part of the reasoning behind performing this transformation is to avoid an over fitting of the HMM. If the raw waveform had been used as input for the HMM then all comparisons made against the model would be unrelated unless they had been sampled from the same interaction that the HMM was trained on. This then allows for an HMM to be trained before submission of speech samples to the HMM for likelihood extraction. The process used in this thesis is, in essence, a form of speech recognition where speakers are tested against themselves and their interlocutor across a series of interactions to determine which speaker they more closely represent for a given speech element at a given time point. This process is described in detail in this section.

The remainder of this section is structured as follows. Subsection 3.6.1 provides the specifics of the analysis performed on the data, subsection 3.6.2 presents the findings of the HMM based analysis, subsection 3.6.3 discusses the results of the HMM based approach and provides an indication of its success.

### 3.6.1 Methodology

The analysis was performed on the same data that was used to conduct the the phonetic analyses in section 3.5. The basic unit used as input to the HMM based approach presented here is the word. This is because the HMMs will be estimating the

underlying states that make up a word which are roughly equatable to phonemes or general spectral information, depending on the HMM type. Looking at larger segments of speech would mean that states would be representing something possibly akin to words or syllables and processing smaller segments of speech would be somewhat meaningless since the states would then represent sub-phonemic units.

The process itself consists of a vectorisation of the data stream, training HMMs for each of the speakers based on their vectorised speech samples, computation of the likelihood of each word in an interaction as being produced by either speaker *A* or speaker *B* through comparison to the HMMs for the speakers and then finally, the correlation of these likelihoods with time in order to classify the general trend of accommodation for the interaction. This process can be roughly broken down into four steps: (1) converting the acoustic signal to MFCC vectors, (2) training the speaker models, (3) computing the likelihood ratios for each speech element in an interaction and (4) correlating the likelihood ratios with time. Each of these steps are now described.

Step 1: Convert acoustic signal to MFCC

This step provides a form that can account for acoustic properties across instances of the same word. Signal segments with similar acoustic properties are represented by similar vectors.

The Mel-Frequency Cepstrum Coefficient (MFCC) is a useful representation of the short-term power spectrum in an audio signal because it represents some transformation akin to the human auditory system and has been used regularly in sound processing and artificial speech recognition (eg. Davis & Mermelstein, 1980; Moreno, 1996; Mistry & Kulkarni, 2013). It has proven to be a robust measure for the linguistic and communicative content in human speech when used in conjunction with artificial speech recognition systems. For the purposes of this thesis, MFCCs for the data were calculated within the Hidden Markov Model Toolkit (HTK) (Young et al., 2006) using the HCopy function.

Figure 3.19 offers a schematic view of the generalised process used by HTK to generate MFCCs. More specifically, the MFCC is derived from discrete Fourier transform based log spectra and is arrived at by performing the following operations:

1. The signal is arranged into short frames determined by the configuration parameters written by the user. These frames should be short enough to allow for the assumption of statistical regularity in the signal whilst remaining long enough to be able to measure a reliable spectral estimate. Window durations of 20 - 40ms with a 10ms frame period are usually acceptable.
2. A power spectrum is extracted for each of the windows, identifying the frequencies that are present (much like the process performed by the human

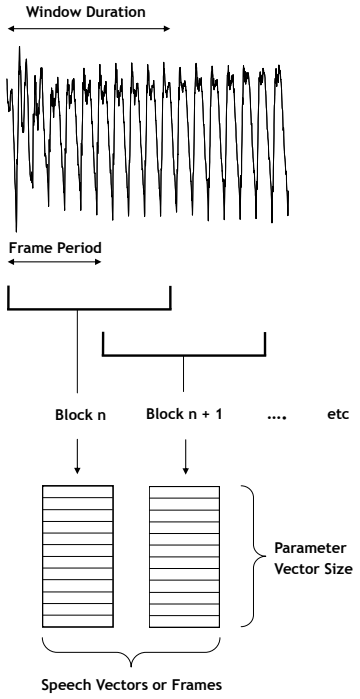



Figure 3.19: Schematic of the generalised process used by HTK to convert a waveform to speech vectors or frames. Adapted from Young et al. (2006) pp.62

cochlea).

3. The Mel filterbank is then applied to the power spectra. This is performed to emulate the human cochlea's difficulties in differentiating between two closely spaced frequencies, especially at higher frequencies. The Mel filterbank bins the power and sums the energy in each filter bin to indicate how much energy exists in the frequency regions. The Mel filterbank consists of triangular frequency bins which are narrow at lower frequencies to give high resolution and wider at higher frequencies where measuring variation is less important due to the nature of the cochlea's processing that is being emulated.

4. The logarithm of all filterbank energies is taken. This step is driven by the fact that humans do not hear loudness on a linear scale (Howard & Angus, 2001; Kreiman & Sidtis, 2011).

5. The discrete cosine transform (DCT) of the log filterbank energies is computed. Because the filterbank energies overlap with one another, there is a certain degree of correlation between them. Computing the DCT decorrelates the filterbank energies which aids with the HMM classifier.

6. The discrete cosine transform coefficients 1-12 are kept. This is done to capture the most relevant information in the signal with out also including fast changes in the filterbank energies which again aids the HMM classifier. Traditionally, the very first DCT coefficient is the sum of all the log-energies. The HCopy function in HTK assigns this to the $0^{th}$ value and is not included as default since it would lack much information about the spectral content

of the signal. Higher DCT coefficients tend to represent fast changes in the filterbank energies. These coefficients are now your MFCCs.

The general process can be considered as a conversion from a 2-byte integer (waveform) to a multi-component vector (as represented by the MFCCs) as HTK considers both the waveform and the MFCCs as sample sequences.

HTK has a number of configuration parameters for conversion of signals into vector form. The parameters concerning the work presented here are TARGETKIND, TARGETRATE, WINDOWSIZE, SOURCEKIND and SOURCEFORMAT. The configuration parameters used are provided below along with a description of their function. Any additional parameters that are not listed remained at the default values as defined by HTK or were taken from the meta data contained in the source file (all durations specified in 100nsec units):

TARGETKIND = MFCC

This parameter determines the output format that the user has requested. This can be set to a number of values including LPC (linear prediction coefficient) and USER for a user defined vector form.

TARGETRATE = 100000.0

This parameter sets the frame rate to be used when producing the target vector output form.

WINDOWSIZE = 250000.0

This parameter sets the window size to be used when producing the target vector output form.

SOURCEKIND = WAVEFORM

This parameter indicates the format of the incoming signal to be converted. This can be any HTK recognised format including values such as LPC and USER.

SOURCEFORMAT = WAV

This parameter indicates the file format of the incoming signal to be converted.

Step 2: Train speaker models

HMMs are produced for both speakers in an interaction, the speaker models are trained over all available data contained in interactions between speakers. Models are trained for both speaker $A$ and speaker $B$, i.e. to set the value of the parameters in $\Lambda_A$ and $\Lambda_B$ so that the probabilities $p(X_A, S_A|\Lambda_A)$ and $p(X_B, S_B|\Lambda_B)$ are maximized, where $X_A$ is the sequence of all observation vectors extracted from all words uttered by $A$ (same for $B$), $S_A$ is the sequence of states for speaker $A$ in the given HMM (same for $B$) and $\Lambda_A$ are the parameters' set for speaker $A$ in

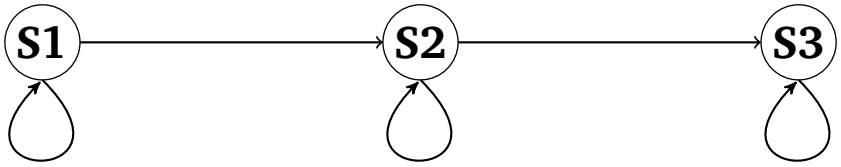



Figure 3.20: Topography schematic for the left-right and word-dependent models, S = state. This topography allows for states to progress sequentially or to remain in their current state. It is not possible to return to a previous state upon leaving it (ie. S1 cannot be returned to once in S2 or S3).

the given HMM (same for $B$). The training is performed through a mathematical model (the Baum-Welch algorithm) implemented in HTK. In this thesis, three different types of HMM are evaluated. These consist of a single state HMM, a left-right HMM and a word dependent HMM. Each of the models are described here in terms of speaker $A$ for ease of explanation. However, all notations and descriptions containing $A$ also apply for $B$. In other words, everything that was done for speaker $A$ was also done for speaker $B$. In the single state HMM the HMM corresponding to speaker $A$ has one state ($D = 1$, where $D$ describes the number of parameters for the state or states) and $p(X,S|\Lambda_A)$ is the probability of speaker $A$ having uttered the words from which the sequence of observations $X$ has been extracted with the underlying state sequence $S$ given $\Lambda_A$, the parameter set of the HMM corresponding to speaker $A$. In the left-right HMM the HMM corresponding to speaker $A$ has three states ($D = 3$, where $D$ describes the number of parameters for the state or states) where the states are only able to progress in a left-to-right topography as presented in figure 3.20. This is so because speech is constrained by time, meaning that phonemes must be produced in a given order to form a word. It cannot be the case that the word 'cat' is produced with a /t/ before a /k/. In this model the HMM is same for all words. As with the single state HMM, $p(X,S|\Lambda_A)$ is the probability of speaker $A$ having uttered the words from which the sequence of observations $X$ has been extracted with the underlying state sequence $S$ given $\Lambda_A$, the parameter set of the HMM corresponding to speaker $A$. Finally, for the word-dependent HMM, the HMM corresponding to speaker $A$ has three states ($D = 3$, where $D$ describes the number of parameters for the state or states) and the same state topography as the left-right model (see figure 3.20). For this HMM a separate set of models is produced for each word uttered by the participants. Again, $p(X,S|\Lambda_A)$ is the probability of speaker $A$ having uttered the words from which the sequence of observations $X$ has been extracted with the state sequence $S$ given $\Lambda_A$, the parameter set of the HMM corresponding to speaker $A$.

Step 3: Compute likelihood ratio

By this point the models for each speaker have been trained and the next step in the process is to submit speech sample to each of the HMMs (one for each speaker) in order to determine whether the word itself was more likely to have been uttered

by either speaker $A$ or speaker $B$. After models have been trained, it is possible to estimate the probability that a given word $w$ has been uttered by a given speaker: if $X_w$ is the sequence of observation vectors extracted from the speech signal segment corresponding to word $w$, $S_A$ is the state sequence associated with speaker $A$ (same for $B$ ) and $\Lambda_A$ is the parameter set for speaker $A$ (same for $B$), then $p(X_w, S_A|\Lambda_A)$ is the probability of that word having been uttered by $A$ and $p(X_w, S_B|\Lambda_B)$ is the same probability for $B$. The right hand side of the following expression:

$$\theta = \frac{p(X_w, S_A|\Lambda_A)}{p(X_w, S_B|\Lambda_B)} \tag{3.4}$$

is called *likelihood ratio* $\theta$. When $\theta > 1$, it is more likely that the word has been uttered by $A$ than by $B$ and vice versa when $\theta < 1$.

Step 4: Correlate time with changes in the speech spectrum

Here the likelihood ratios that were generated for each word in the previous step are linked with the start time at which that word was produced. This allows for an interpretation of accommodation across the course of an interaction. An Interaction can be thought of as a sequence of words uttered either by $A$ or by $B$. If $w_i^{(A)}$ is the $i^{th}$ word uttered by $A$, then the following likelihood ratio can be considered a measure of how speaker $A$ becomes more similar to speaker $B$:

$$\theta_i = \frac{p(X_{w_i^{(A)}}, S_A|\Lambda_A)}{p(X_{w_i^{(A)}}, S_B|\Lambda_B)} \tag{3.5}$$

The ratio $\theta_i$ can be measured for each word uttered by $A$ resulting in a sequence of pairs $(\theta_i, t_i)$, where $t_i$ is the time when word $w_i$ starts. If the correlation between the $\theta_i$'s and the $t_i$'s is negative to a statistically significant extent, then $A$ tends to converge to $B$ and vice versa if the correlation is positive to a statistically significant extent. If the correlation is not statistically significant, then there is no evidence for change. Switching $A$ and $B$ in the expression of the likelihood ratio demonstrates how $B$ shifts with respect to $A$.

Using Figure 3.21 as a visual representation of the whole process, we can consider the terms $corr[t_i^{(A)}, d_i^{(A)}]$ and $corr[t_j^{(B)}, d_j^{(B)}]$ to be the output of the measure. Here, $t_i^{(A)}$ represents the start time of a given word uttered by speaker $A$ and $d_i^{(A)}$ represents the likelihood ratio for that word having been uttered by speaker $A$ (the same holds for speaker $B$). Taking $corr[t_i^{(A)}, d_i^{(A)}]$ as an example, there are three possible outcomes for this value:

1. $corr[t_i^{(A)}, d_i^{(A)}]$ is positive and statistically significant, speaker A tends to become more similar to speaker B over time.

2. $corr[t_i^{(A)}, d_i^{(A)}]$ is negative and statistically significant, speaker A tends to become less similar to speaker B over time.

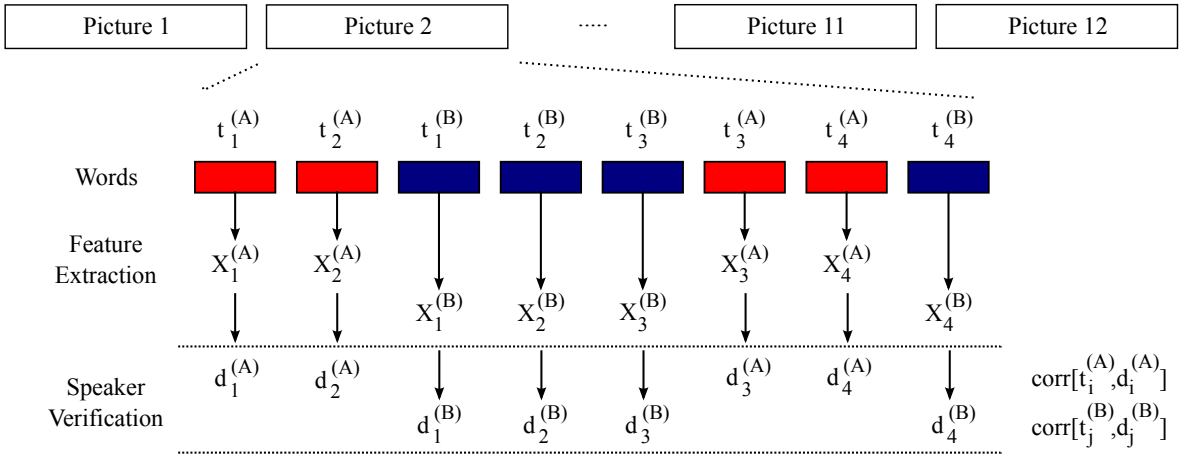



Figure 3.21: *Schematic of the approach. The words uttered during the conversation interval related to a specific picture were transcribed manually, automatically segmented and split into two groups, namely words uttered by A (red rectangles) and words uttered by B (blue rectangles). Each word is converted into a sequence of observation vectors (here, 12-dimensional MFCC vectors). For a given sequence of observation vectors, the distance measurement $d_i^{(A)}$ or $d_i^{(B)}$ are obtained using mixtures of Gaussians. The Spearman coefficient is used to measure the correlation between the distance measures and the time at which words have been uttered. Hence allowing for the assessment of the potential for accommodation to have taken place or not.*

3. $corr[t_i^{(A)}, d_i^{(A)}]$ is not statistically significant, accommodation (if any) is invisible to the model.

These values can be reduced down to the general form $L(A,B)$:

$$L(A,B) = \rho[l_i(A,B), t_i^{(A)}] \tag{3.6}$$

where $\rho(.,.)$ corresponds to the *Spearman Correlation Coefficient*, used as it is more robust than standard correlation coefficient to outliers (this is equivalent to $corr[t_i^{(A)}, d_i^{(A)}]$). $L(B,A)$ is defined in the same way as for $L(A,B)$ but with the $A$ and $B$ terms reversed. The correlation being performed here is between the likelihood ratios generated for each word uttered by a speaker and the start time of that given word, the correlation is not being made with time itself.

We can then consider these three possible outcomes within the speaker dyad. When we pair each of these possible outcomes up with the outcome of the correlation relating to the other speaker, there are four possible cases:

1. **Convergence (CO)**

   Where either speaker A or B demonstrate a positive and statistically significant correlation whilst the other either demonstrates no change or a positive and statistically significant correlation, we can say that convergence has occurred.

2. **Divergence (DI)**

   Where either speaker A or B demonstrate a negative and statistically significant

correlation whilst the other either demonstrates no change or a negative and statistically significant correlation, we can say that divergence has occurred.

3. **Complementarity (CM)**

   Where both speakers demonstrate statistically significant correlations in opposite directions, we can say that complementarity has occurred.

4. **Maintenance (MN)**

   Where both speakers demonstrate no statistically significant correlations, we can say that they demonstrate maintenance.

Table 3.13 provides a visual schematic of this differentiation.

| | **B+** | **B=** | **B–** |
|---|---|---|---|
| **A+** | CO | CO | CM |
| **A=** | CO | MN | DI |
| **A–** | CM | DI | DI |

Table 3.13: Dyad correlation combination table. $A+$: $A$ demonstrates convergence, $A=$: $A$ demonstrates maintenance, $A-$: $A$ demonstrates divergence, the same notation is used for $B$. CO: convergence, CM: complementarity, MN: maintenance and DI:divergence.

Using this system of differentiation of the outcomes from the measure, each conversation can be allocated to one of these four categories.

### 3.6.2 Results

Before reporting the results for this experiment it is worth recounting what each of the HMM types represent. Recall from section 2.4.1 that each of the HMM types represent the following:

GMM: general distribution of acoustic evidence in the feature space.

left-right: temporal patterning of the observed acoustic evidence.

word-dependent: change in acoustic evidence over time.

Where a GMM is an HMM with a single state, the left-right HMM has three states allowed to proceed in a left-right topography (see figure 3.20) where the state sequence remains the same for all words and the word-dependent HMM has three states allowed to proceed in a left-right topography where the state sequence is specific for each word. Holding this in mind when considering the results will help to clarify some of the interpretations.

In terms of the results themselves, the goals of the computational analysis can be considered to be two fold:

- To assess whether the HMM-based approach is able to detect adaptation patterns in the speech signal.
- To determine if results from the HMM based approach are compatible with the trends suggested by the phonetic analysis.

With regards to the second aim for this results section, what this effectively means is that the relationship between the classification of accommodation types and both presentation position and interaction length will be assessed. The detection of accommodation in relation to local speech realisations of the partner is not possible with this HMM based approach because adaptation patterns are determined based on the dyad rather than on individual responses.

In order to evaluate if the HMM-based approach was able to detect real patterns in the data and that the results were not the brought about by chance, a binomial test was performed. The results of this test are provided in Table 3.14.

| Model Type | No. Significant Cases | $p$ |
|---|---|---|
| GMM | 57 | $< 0.001$ |
| Left-Right | 50 | $< 0.001$ |
| Word-Dependent | 42 | $< 0.001$ |

Table 3.14: Results of binomial test for behavioural experiment.

The data used for the test were counts for the number of times at least one of the two correlations $L(A,B)$ and $L(B,A)$ was found to be statistically significant. For a confidence level of $0.05$ (after Bonferroni correction) the number of statistically significant correlations was found to be $57$, $50$ and $42$ for the GMMs, left-right models and word-dependent models, respectively. For all models, the binomial tests returned $p < 0.01$. This indicates that the results obtained are unlikely to have been generated by chance. The result of this test suggests that the trends being observed by the models are related to the physical trace of communicative phenomenon in the speech signal and are not the result of chance. It can be concluded that the HMM-based approach, which view the speech signal as holistic and continuous, is able to detect real trends in the data.

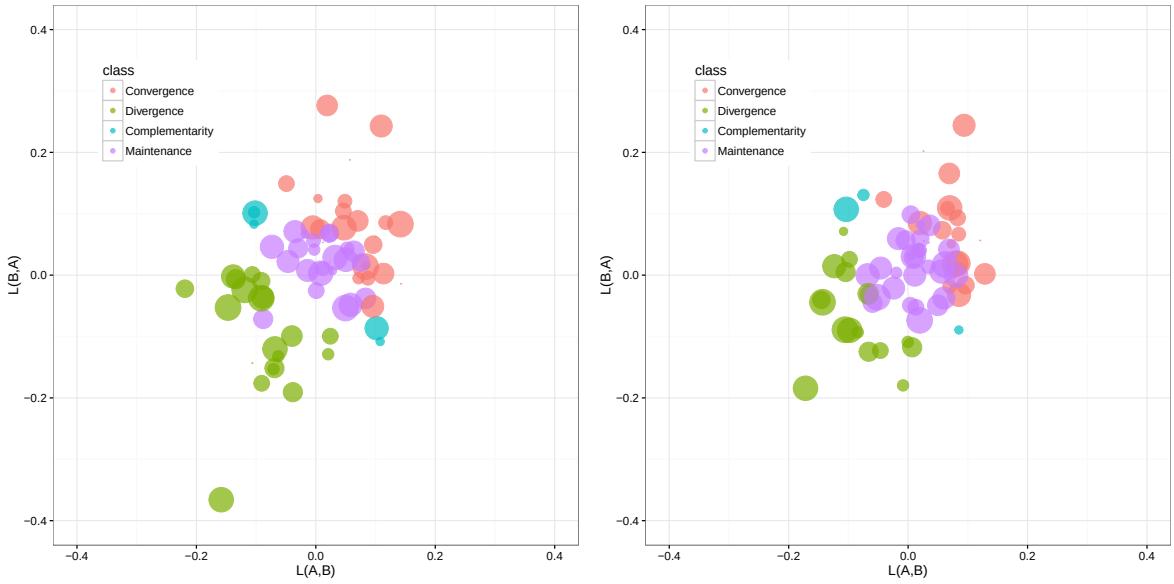
The second aim of the computational analysis is to determine if results of the HMM analyses are compatible with the results of the phonetic results. To test this, the classification results of the HMM analyses will be compared with both the presentation position and interaction length. First, the results of a comparison with presentation position will be presented. The classification results of the three HMM types (GMM, left-right and word-dependent) were compared to the position of the task in the experiment. Table 3.15 reports the counts of the interactions classified as convergence, divergence, maintenance or complementarity for each of the three HMM types. It also presents the average presentation position associated with each of the accommodation and HMM types.

| | Model Type | Convergence | Divergence | Maintenance | Complementarity |
|---|---|---|---|---|---|
| Count | GMM | 22 | 20 | 25 | 5 |
| | Left-Right | 21 | 17 | 31 | 3 |
| | Word-Dependent | 22 | 12 | 37 | 1 |
| Avg. Position | GMM | $6.00 \pm 0.77$ | $7.15 \pm 0.76$ | $6.60 \pm 0.67$ | $5.60 \pm 2.01$ |
| | Left-Right | $5.86 \pm 0.79$ | $6.65 \pm 0.87$ | $6.97 \pm 0.58$ | $5.33 \pm 2.85$ |
| | Word-Dependent | $6.68 \pm 0.76$ | $7.00 \pm 1.31$ | $6.32 \pm 0.51$ | $3.00 \pm NA$ |

Table 3.15: This table reports the counts and average position ($\pm$ the standard error) of interactions classified as either Convergence, Divergence, Maintenance or Complementarity.

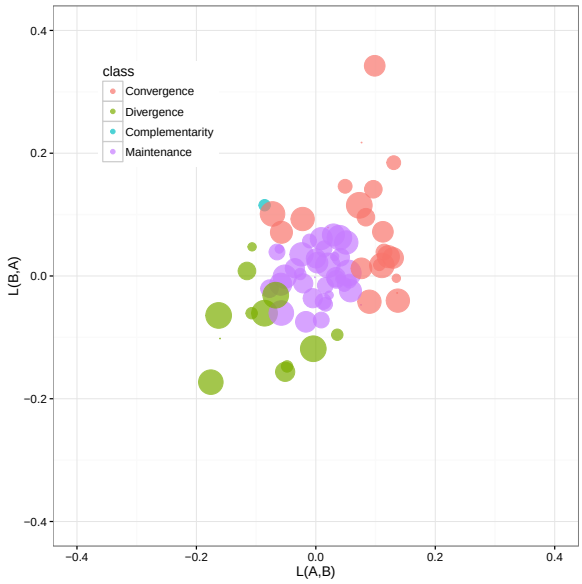
Looking at table 3.15, it is clear that for all HMM model types (GMM,left-right and word-dependent), the most common classification is of maintenance, followed by convergence as the next most common and then divergence. Complementarity is somewhat more rare and occurs only a handful of times across the experiment. When the average presentation positions of the tasks are compared across accommodation patterns, there doesn't appear to be any major deviations from the middle of the experiment. This statement holds for all accommodation patterns and across all HMM types except for the complementarity accommodation pattern where there are too few cases to draw definite conclusions from. Figure 3.22 offers a general, visual overview of the data and presents the results of the three HMM types as a bubble plot where the size of the bubbles are proportional to the position of the task it represents within the experiment. Larger bubbles represent interactions that took place later in the experiment and smaller bubbles represent interactions that took place earlier in the experiment.

It should be noted that the space in which the results are plotted in figure 3.22 is a representational space of the Spearman correlation coefficients associated with $L(A,B)$ and $L(B,A)$. The space is a representation of the movements of one participant in relation to the other, within a pair. Convergence would be expected to be associated with positive correlation values for at least one participant and negative correlation values for at least one participant would be expected for divergence. Indeed, this is what is seen in the plots. If the HMM-based approach detected more convergence in the later trials in the experiment then the plots would show generally larger bubbles for the convergence cases than for the divergence conditions. In general, across the three model types, this does not seem to be the case. There looks to be a mostly similar distribution of bubble sizes (and therefore presentation order positions) for both the convergence and divergence cases. There also seems to be a similar spread of bubble sizes for the maintenance condition. The number of complementarity cases are too few to make reasonable interpretations. The data presented here appear to support the intuitions that there is no relationship between accommodation pattern and presentation position. Figure 3.23 presents the data as a bar graph with standard errors.




(a) Gaussian mixture model (GMM)

(b) Left-right HMM




(c) Word-dependent HMM

Figure 3.22: Results for all three of the HMMs compared to the position of the task in the experiment. The size of the bubbles is proportional to the position of the task in the experiment. A larger bubble indicates that the task associated with that bubble was presented later in the experiment.

The bar graphs in figure 3.23 for the average presentation positions for each of the accommodation patterns show a trend for convergence, divergence and maintenance to all pattern around the mid point of the experiment (the dotted line). This suggests that for every instance of convergence, divergence or maintenance that occurs towards the end of the experiment, there is another instance that took place towards the beginning of the experiment. Clearly, there is some variation between the models but they all provide roughly equivalent average values for presentation position. Looking at the values presented in table 3.15 for average position, it can be seen that across the models, there is less than a 1.00 difference between the

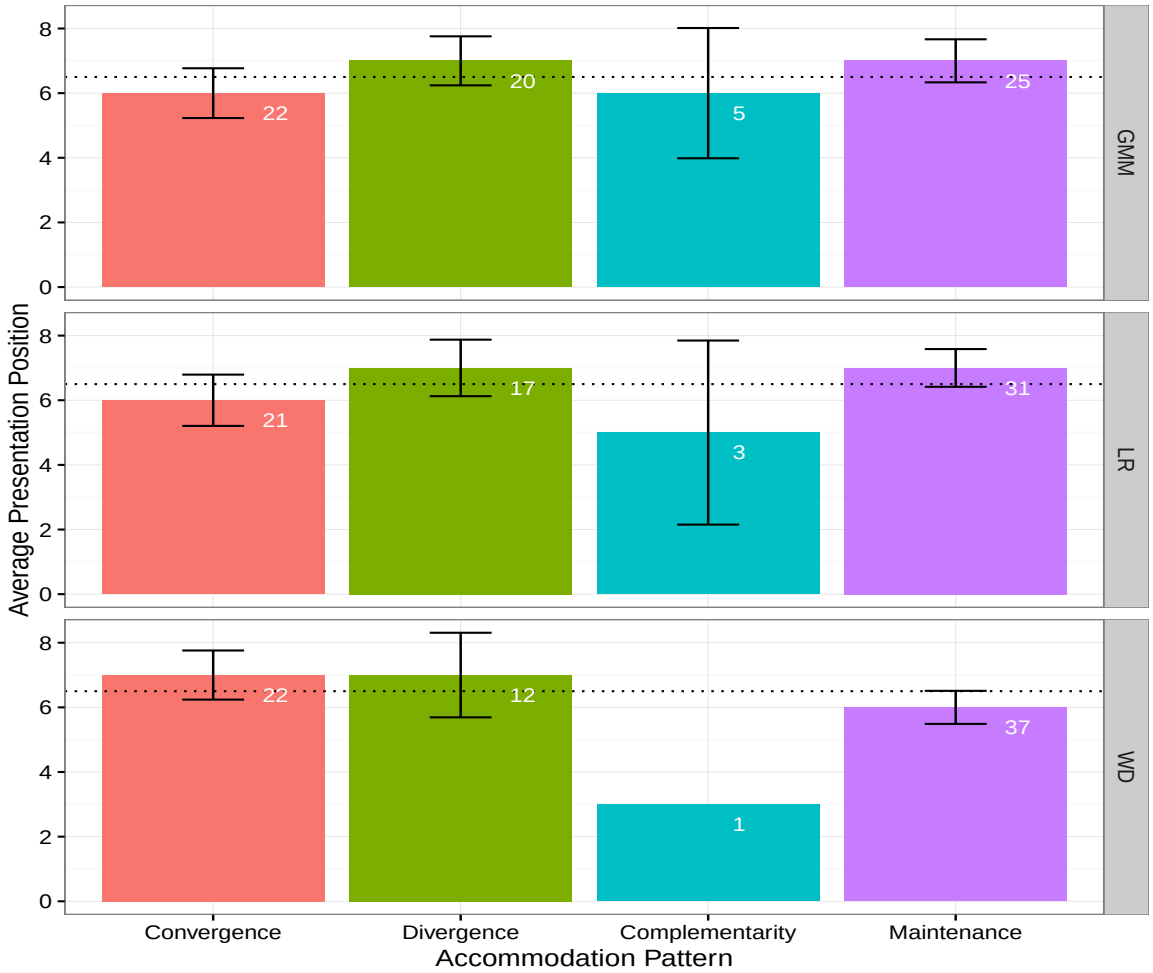



Figure 3.23: Average presentation position of tasks for each of the four possible adaptation patterns. Error bars represent the standard error of the mean. Numbers in each bar indicates the number of observations for that adaptation pattern. Dotted black line indicates the middle point of the overall experiment (ie. halfway through interaction 6). GMM = Gaussian mixture model (single state HMM), LR = left-right HMM, WD = word-dependent HMM.

values across all models in each accommodation pattern type (excluding complementarity). This is coupled with standard error values that all either encompass the middle trial of the experiment or pattern very close to it. Taken together, this all suggests that there is no real difference between the accommodation patterns for any of the HMM types. Table 3.16 presents the results of a series of t-tests testing for differences between the accommodation patterns.

The results presented in table 3.16 confirm intuitions drawn from previous data representations that there are no real differences between the accommodation patterns. This is true across all HMM types and mirrors the results for presentation position demonstrated in the phonetic analyses.

The next set of results presented here relate to the relationship between the HMMs accommodation classification and the length of time taken to complete a task. The classification results of the three HMM types (GMM, left-right and word-dependent) were compared to the interaction lengths of the tasks in the experiment.

| Model Type | Comparison | $p$ |
|---|---|---|
| GMM | convergence:divergence | 0.295 |
| | convergence:maintenance | 0.556 |
| | divergence:maintenance | 0.588 |
| Left-Right | convergence:divergence | 0.508 |
| | convergence:maintenance | 0.255 |
| | divergence:maintenance | 0.755 |
| Word-Dependent | convergence:divergence | 0.823 |
| | convergence:maintenance | 0.688 |
| | divergence:maintenance | 0.563 |

Table 3.16: This table presents the t-test results for comparisons between accommodation category classification by the different HMM types and the presentation position of the DiapixUK images.

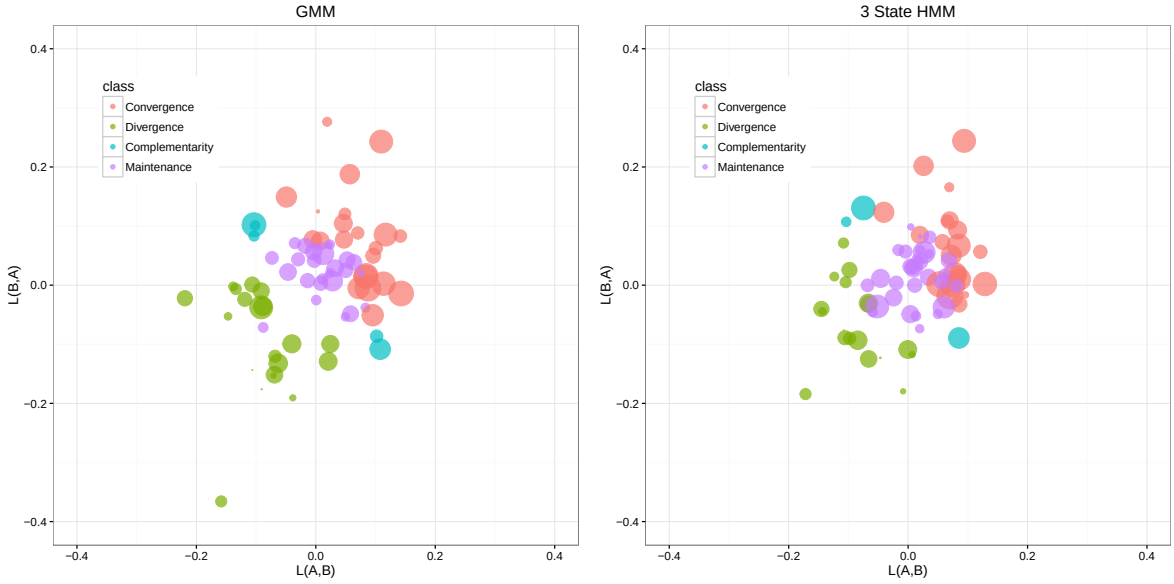
Table 3.17 reports the counts for the interactions classified as convergence, divergence, maintenance or complementarity for each of the three HMM types. It also presents the average presentation position associated with each of the accommodation and HMM types.

| | Model Type | Convergence | Divergence | Maintenance | Complementarity |
|---|---|---|---|---|---|
| Count | GMM | 22 | 20 | 25 | 5 |
| | Left-Right | 21 | 17 | 31 | 3 |
| | Word-Dependent | 22 | 12 | 37 | 1 |
| Avg. Duration (s) | GMM | $656 \pm 47$ | $468 \pm 36$ | $455 \pm 25$ | $511 \pm 112$ |
| | Left-Right | $669 \pm 46$ | $420 \pm 32$ | $476 \pm 26$ | $659 \pm 164$ |
| | Word-Dependent | $594 \pm 42$ | $460 \pm 52$ | $499 \pm 30$ | $900 \pm NA$ |

Table 3.17: This table reports the counts and average durations ($\pm$ the standard error) of interactions classified as either Convergence, Divergence, Independence or Complementarity. Values are rounded to the nearest second.

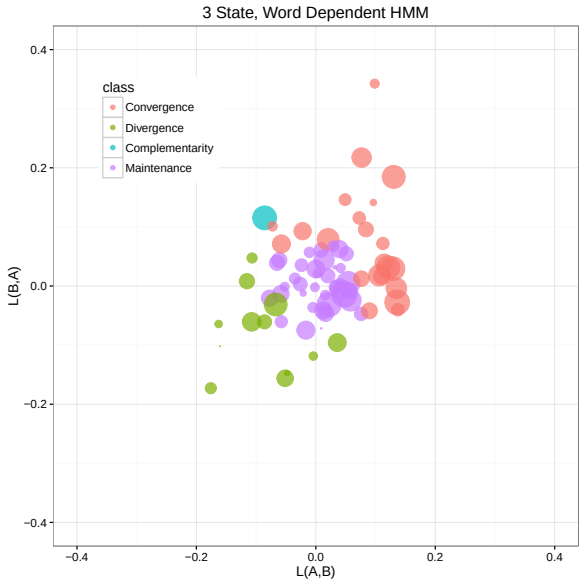
Looking at the results presented in table 3.17 it can be seen that maintenance is again the most common classification across all HMM types. Since the number of complementarity classifications is again low, it is difficult to draw any clear interpretations and they will not be interpreted further. Comparing the classifications of the remaining accommodation patterns, it can be seen that longer interactions tend to be associated with a classification of convergence, this is consistent across all HMM types. Further to this, there is a general trend for the shortest interactions to be classified as divergence. However, this only holds for the left-right and word-dependent HMMs, for the GMM maintenance is associated with the shortest interaction. Figure 3.24 offers a general, visual overview of the data and presents the results of the three HMM types as a bubble plot. The size of the bubbles are proportional to the length of the task it represents within the experiment.

The same type of interpretation that applied to figure 3.22 also applies here. The




(a) Gaussian mixture model (GMM)

(b) Left-right HMM




(c) word-dependent HMM

Figure 3.24: Mixtures of Gaussians results for length of interaction comparisons. The size of the bubbles in 3.24a, 3.24b and 3.24c is proportional to the time taken to complete the task. A larger bubble indicates a greater time taken.

size of the bubbles is proportional to the length of the interaction that it represents. Larger bubbles represent longer interactions and smaller bubbles represent shorter interactions. If the HMM-based approach detected more convergence in the longer trials in the experiment then it would be expected that the plots would show generally larger bubbles for the convergence cases than for the divergence conditions. In general, across the three model types, this seems to be true. Whilst there are most definitely a number of smaller bubbles associated with convergence, the majority of the larger bubbles are also associated with convergence. The divergence classification appear to have consistently smaller bubbles than convergence. The observation broadly holds across all of the HMM types although perhaps less so for

the word-dependent HMM. There also seems to be a similar spread of bubble sizes for the maintenance condition. The number of complementarity cases are too few to make reasonable interpretations. Figure 3.25 presents the data as a bar graph with standard errors.

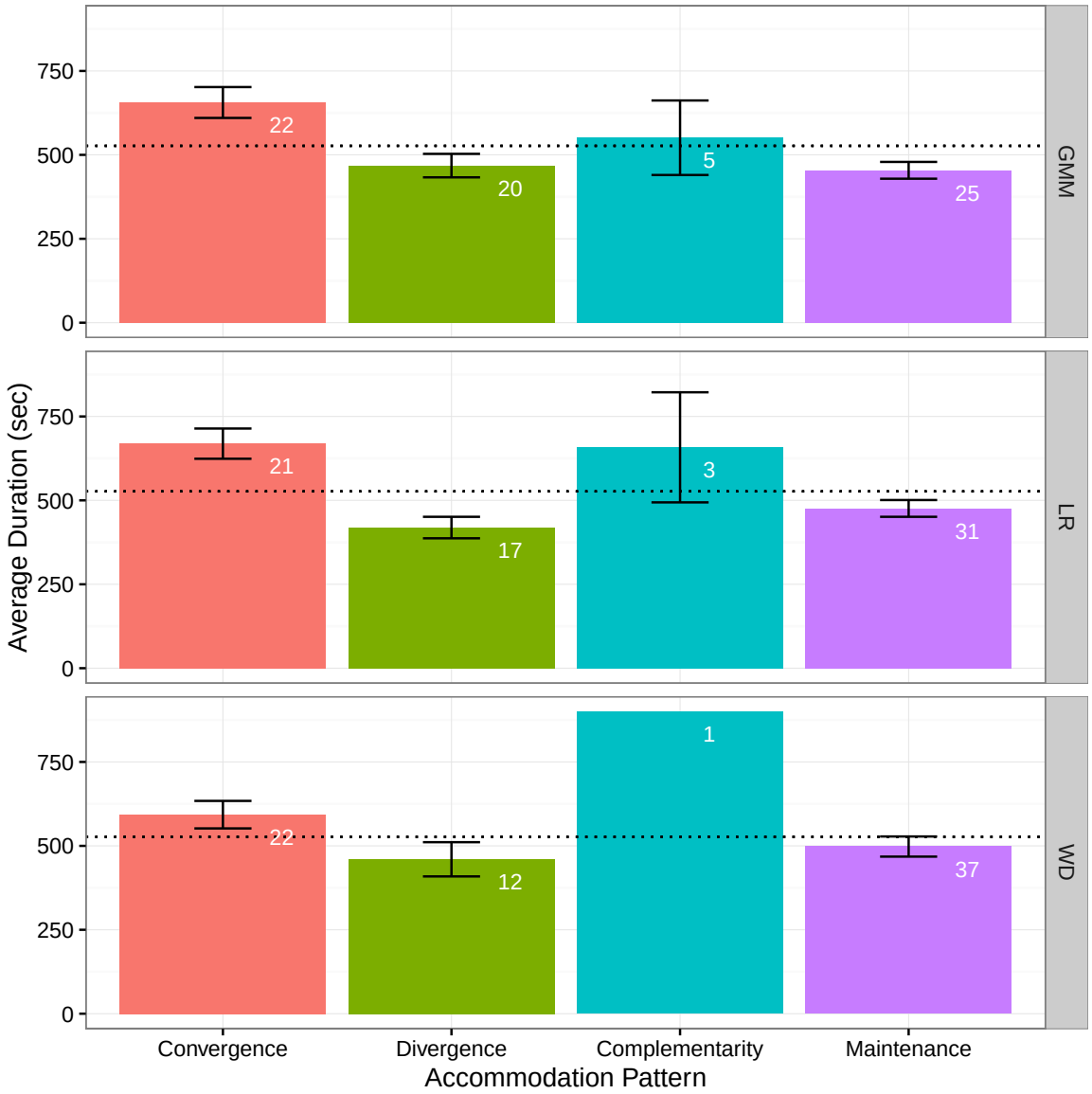



Figure 3.25: Average duration of tasks for each of the four possible adaptation patterns. Error bars represent the standard error of the mean. Numbers in each bar indicates the number of observations for that adaptation pattern. Dotted black line indicates the mean duration of all interactions. GMM = Gaussian mixture model (single state HMM), LR = left-right HMM, WD = word-dependent HMM.

The bar graphs for the average interaction length for each of the accommodation patterns show a trend for convergence to be classified in the interactions that took the longest to complete. Divergence and maintenance appear to pattern quite similarly with divergence possibly being classified in interactions that were slightly shorter than those for maintenance, on average. Across all HMM types, interactions classified as convergence are longer than the average length of an interaction (the dotted line) and interactions classified as divergence are shorter than the average

length of an interaction. Maintenance also tends to occur in shorter interactions but appears to track a little closer to the average interaction length than divergence. This suggests that there is likely to be a difference between the lengths of interactions that are classified as convergence and the lengths of those that are classified as divergence or maintenance. This looks to be the case across all HMM types although perhaps to a smaller degree for the word-dependent HMMs. Table 3.18 presents the results of a series of t-tests testing for differences between the accommodation patterns.

| Model Type | Comparison | $p$ |
|---|---|---|
| GMM | convergence:divergence | $0.003^{**}$ |
| | convergence:maintenance | $0.000^{***}$ |
| | divergence:maintenance | 0.749 |
| Left-Right | convergence:divergence | $0.000^{***}$ |
| | convergence:maintenance | $0.000^{***}$ |
| | divergence:maintenance | 0.193 |
| Word-Dependent | convergence:divergence | 0.060 |
| | convergence:maintenance | 0.070 |
| | divergence:maintenance | 0.531 |

Table 3.18: This table presents the t-test results for comparisons between accommodation category classification by the different HMM types and the duration of the interactions.

The results presented in table 3.18 confirm intuitions drawn from previous phonetic analyses, as presented in section 3.5.3, that the differences between interaction length for convergence classifications and divergence and maintenance classification is real. This finding does vary across the HMM types with the left-right HMM providing the strongest effects. The GMM model reports a greater difference between the length of convergence interactions and the length of maintenance interactions than the equivalent comparison between convergence and divergence. Having said that, both still report significance at below the $1\%$ level. There are no significant results for the word-dependent HMMs but both comparisons involving convergence are approaching significance. The length of interactions classed as convergence compared against divergence is reported at $p = 0.06$ and convergence compared against maintenance is reported at $p = 0.07$. These results suggest that the use of convergent behaviour across the auditory speech signal as a whole tends to occur in longer interactions. As such, they are in keeping with the small number of significant findings, and weak trends, found in the phonetic analyses.

### 3.6.3 Discussion

The results presented in section 3.6.2 provide support for both the use of an HMM based approach in the detection of accommodation in a continuous interaction and provide support for the tentative conclusions regarding accommodation and interaction length presented for the phonetic analyses in section 3.5.3. The HMM based approach can be said to be detecting the presence of accommodation which eludes phonetic analyses of single features modelled as functions of time, as determined through the use of a binomial test. Having said that, no relationships were found between accommodation classification and presentation position. However, classification of convergent interactions was found to be higher during interactions that lasted for longer.

Again, prior to engaging with the main bulk of the considerations for this subsection, a recap of what each of the HMM types used in here represent. Recall that each of the HMM types represent the following:

GMM: general distribution of acoustic evidence in the feature space.

left-right: temporal patterning of the observed acoustic evidence.

word-dependent: change in acoustic evidence over time.

The finding that the HMMs that are employed here are able to detect trends in the speech signal rather than just random noise is promising and essential to ensuring the validity of an approach such as this. The relationship between the words uttered by the speakers is modelled as an adaptation towards or away from a speech partner. This is the essence of accommodation, adjusting one's speech forms to become either closer to or further away from that of an interlocutor. The fact that significant results reflected the tentative trends in the phonetic analyses certainly provides further support that it is indeed detecting accommodation. In order to truly test this HMM based approach to accommodation detection, a true baseline of accommodation would be useful. However, this is a difficult (if not near impossible) baseline to acquire. Accommodation is known to vary in relation to a number of social factors (see subsection 2.1.2) and controlling for all of these factors to produce a clean baseline would not only be incredibly difficult but may also prove to be removing a core aspect of accommodation itself. However, the application of this approach to larger datasets may prove useful in determining the broad validity of its application.

Across both the analysis for presentation position and interaction length, the counts for accommodation classification remains the same. This allows for some interpretations to be made regarding the movement from GMM through left-right HMM to word-dependent HMM. As the HMMs are progressed through in this manner, they can be said to be getting more and more specific about the sources of

speech information that they account for. For instance, the GMM doesn't account for time, whereas the left-right and word-dependent HMMs do. Whilst the number of cases of convergence remains roughly stable, it can be seen that divergence cases decrease as cases of maintenance increase. What this suggests is that convergence is a robust and more easily identifiable phenomenon than divergence. Where divergence can be widely seen at the level of general speech characteristics, it becomes harder for the models to classify as increasing restrictions are applied to the classification criteria. This leads to an increase in the classification of maintenance cases since this represents the accommodation pattern were no correlation between participants is detected. Overall, this suggests that convergence tends to take place across all levels being evaluated by the HMMs whilst divergence takes place more easily at the level of acoustic evidence distribution.

The non-significant findings between presentation position and the classification of accommodation patterns suggests that the overall time that participants spent together did not have a major impact on the patterns of accommodation employed. This finding mirrors that found in the phonetic analyses where no effects were found for presentation position. This may be related to the length of time that participants were paired for during the experiment. It may not have been enough time for large accommodative effects to have emerged to the extent that they could be determined from a holistic viewpoint within a continuous interaction. Having said that, it may be the case that a perceptual evaluation of the early and late utterances of the speakers may reveal some overall trends. This might be possible because, although the HMMs used here have been trained over the available data, they are not trained either as specifically as the human perceptual system is and are not able to draw on additional top-down information like the human perceptual system can. Whilst the machine learning approach presented here certainly offers an alternative to both phonetic and perceptual studies, there are limitations to its capabilities. It is also worth noting that although machine learning approaches might be able to classify speech data, it is not to say that the method through which this is accomplished is the same mechanism that the human perceptual system uses. It could perhaps be seen as a crude approximation but direct relational links between the process used by HMMs and cognitive function should not be made.

The series of significant findings for a link between the classification of accommodation patterns and interaction length follows the findings for VOT and the F2 of the STRUT vowel as well as some of the tentative trends suggested by other phonetic measures. The fact that each of the HMMs can be thought of as representing different information contained in the speech signal allows for multiple interpretations of what the results of each HMM might indicate. In the GMMs, convergence is found to differ significantly from both divergence and maintenance (complementarity is not considered due to too few reported cases), both of which were found to be below the $1\%$ level. No difference was found between divergence and maintenance.

These findings suggest that during tasks that take longer, and that the participants therefore find more difficult to complete, there is a tendency for convergence to take place. Given that the GMMs only account for the general distribution of acoustic evidence in the feature space, it can be assumed that associated results indicate broad changes across general speaker characteristics. Since GMMs do not account for time, holding only a single state for a speaker, the criteria used for determining the likelihood of any given word having been uttered by either speaker $A$ or speaker $B$ must represent the general similarities rather than any word or context specific features. This finding could be taken to suggest that in situations where further clarification is needed to complete a joint task, speakers will adapt the general characteristics of their speech to improve communication.

The left-right HMMs represent something different to the GMMs although they roughly report the same findings. As with the GMMs, the left-right HMMs demonstrate a differences in the length of interactions classified as convergence and divergence as well as differences between lengths of interaction for convergence and maintenance. Having said this, the left-right HMMs appear to demonstrate that the inclusion of time into the models leads to a stronger classification of divergence for interactions that have shorter interaction lengths than compared to the GMMs. This finding provides further support for the suggestion that challenging situations lead to an increase in convergent behaviour.

The findings for the word-dependent HMMs present results that tend in the same direction as the results for the GMMs and left-right HMMs but do not demonstrate significant findings. For both of the convergence comparisons, the $p-values$ of the tests are approaching significance and both are below the $10\%$ significance level. The lack of significance at the $5\%$ level may have its roots in the number of states used for this type of HMM. For the word-dependent HMMs, the states representing each word can be thought of as roughly equating to the underlying phonemes in the target word. In the work presented in this thesis, different numbers of states were not trialled, only $3$ state HMMs were used. What this means is that for longer words that have more phonemes, a number of phonemes may have been collapsed into the same state and for shorter words, a single phoneme may have been categorised over multiple states. It may be the case that this inability to distinguish between longer and shorter phoneme strings may have led to this reduction in significance. Having said that, as mentioned above, the results are still approaching significance so the overall impact of the number of states may not be a huge factor but a more adaptive approach would still likely aid in accommodation detection.

One benefit of the approach presented here is that it is able to classify accommodation into its constituent categories of convergence, divergence, complementarity and maintenance. Further to this, it is able to do so at three main levels, the distribution of acoustic evidence in the feature space, temporal patterns in the distribution of acoustic evidence and word dependent temporal patterns of the acoustic

evidence. It provides a tool that allows for a holistic assessment of accommodation that is drawn from the acoustic evidence found in the speech signal rather than being based on the perceptual judgements of others.

There are a number of technical considerations that might provide improvements in performance and classification accuracy that might be worth exploring. Many of these considerations concern the implementation of the HMMs within HTK. For instance, it would be worth trialling different vectorisation parameters to determine whether there is a specific spectral form that best tracks with accommodation. In addition to this, the frame and window size for this experiment were left at $10$ and $25$ ms, respectively. Whilst these values have been shown to be reliable in traditional speech recognition systems, this implementation of speech recognition techniques is non-standard. An exploration of different settings may yield improvements. Additionally, the number of states used in the left-right and word-dependent models was kept at $3$. This means that within each word, transition is only possible between three states. This may be under representative for some larger words that might be used by participants. Exploring a variety of different state values could again improve classification although a form of adaptive algorithm for determining the most appropriate number fo states for any given word would perhaps offer the best results. Further improvements to the approach presented here might explore the inclusion of word content. This could lead to improved performance in detection rates for the approach by further clarifying the context in which specific words and acoustic features are used. In the specific context of this experiment, the inclusion of higher order information may provide further indications about the interplay between adaptation and task performance. However, as a starting point, HMMs as implemented here provide a solid base from which to proceed.

Broadly, the application of the HMM based approach can be considered to be able to detect and classify accommodation during a continuous interaction based on holistic measures of the speech signal. The findings provide support for the use of machine learning approaches in the detection of subtle phenomena in the speech signal. This may provide a potential new avenue for research in accommodation, allowing for acoustic assessments of accommodation to be made for speakers engaged in continuous dialogue.

## 3.7 General Discussion

The general findings for this chapter and the behavioural experiment are that HMM based approaches to detecting accommodation in a continuous interaction outperform traditional acoustic phonetic approaches, at least for the features evaluated here (VOT, F1/F2 of STRUT, THOUGHT and TRAP, speech rate). Whilst the HMM based approach may still need refining, even as a coarse tool, it still demonstrates a clear advantage over more traditional measurement approaches. This section offers a review of the experiment conducted in chapter 3, it offers suggestions for the possible implications of the findings, discusses the drawbacks and possible improvements that could be made to the experiment and summarises the key points from this chapter that should be taken forward.

Taken together, the findings of the phonetic and computational analyses support the notion that accommodation is a multi-feature, time-dependent phenomenon. For the acoustic-phonetic analyses prior to the non-linear integration of time using GAMs, accommodation was only detected in speech rate. Even once the time integration had been performed, the detected accommodation was minimal for any given acoustic-phonetic feature. Part of the reason for this may be due to the way in which the speech signal is sampled when collecting acoustic-phonetic features. There are only a finite number of any given acoustic-phonetic features in a speech signal and once they are extracted, the rest of the data are ignored. For example, the total time used in the analysis of all three vowels amounts to less than $1\%(0.817\% : STRUT = 0.133\%, THOUGHT = 0.438\%, TRAP = 0.246\%)$ of the total length of utterances in the dataset. Being able to make an assessment of the degree of change between speakers across an experiment based on less that $1\%$ of available data would appear to be wishful thinking. However, the experiments reviewed in subsections 2.2.3 and 2.2.4 have all demonstrated effects based on acoustic-phonetic sampling of the data. Although the approaches used do not make use of continuous speech data and often apply methods that evaluate accommodation with a measure from outside the interaction (ie. $AXB$), their findings, when contrasted against the ones presented here, do demonstrate how robust a phenomenon accommodation can be.

The one measure that does make use of all available data is speech rate. Indeed, in this case an effect was found prior to integration of time using the GAMs. However, from the total number of utterances in the sample, one would expect a trend to emerge given the task that the participants were undertaking (see speech rate discussion in subsection 3.5.3). Further to this, the fact that no trends were found once time was integrated within each interaction suggests that while speech rate may vary in general, time dependency is an important factor and is able to explain a good deal of the variance. So, whilst more fine-grained acoustic-phonetic measures may provide an insufficient sample of available data, speech rate may prove to be

too broad a measure to detect the subtle variations associated with accommodation.

The computational method presented in this chapter, may provide a way of integrating improved sampling of speech material and the capture of broader speech features. As discussed in subsection 2.4.1, machine learning approaches have been successfully applied to many speech recognition problems. They have the ability to recognise words from a string of continuous speech (Deng & Li, 2013) as well as identifying specific speakers (Lan et al., 2013). The findings presented here demonstrate that when a coarse version of the available machine learning approaches is applied to the problem of accommodation detection, it is able to detect effects that are overlooked by more traditional approaches. Because the HMM based approach makes use of all of the identifiable speech data, it can be said to be maximising the available data. By making use of as much of the data as possible and considering it in a time-dependent manner, the continuous, ongoing aspects of accommodation are able to be captured. Further to this, because it utilises MFCCs during its vectorisation process, it can be said to be capturing a more holistic view of the speech signal. This allows for more of the interaction between ongoing acoustic activity to be captured since the speech signal is being sampled across an interaction rather than being categorised based on the content of a sample window. It is also better able to deal with the impact of time on the speech signal since this is integrated into the model from the outset for the Left-Right and Word-Dependent models. However, this is not to say that the approach is not without its drawbacks.

The HMM based approach cannot provide an indication of the exact speech features that are being used to produce accommodative behaviour. Because MFCCs, as used in this thesis, do not map neatly onto aspects of speech such as VOT or fundamental frequencies, it is not possible to retroactively determine the features that are most present in convergence, for example. Thus, if the research question of interest concerns the specific nature of use of a particular acoustic-phonetic feature, then it would be best to apply traditional approaches. What the approach offers is a detection of similarity in speech forms, not an explanation of what that similarity is constructed of. Having said that, the HMM based approach was able to more consistently and more accurately detect accommodation than more traditional phonetic analyses. Further to this, it was also able to disambiguate the different forms of accommodation (convergence, divergence, complementarity and maintenance) because it evaluates the two interactants at the smae time. What this represents is a key contribution to the field, opening the door for the use of HMM and other machine learning based approaches for the investigation of accommodation.

Something that might help to validate these findings and that could potentially provide a form of baseline for accommodation would be to perform a perceptual study using these data. The human perceptual system is sensitive to changes in the speech signal that are pertinent to social interactions (see section 2.1) and does not perceive speech in terms of discrete acoustic-phonetic units (eg. VOT). As such,

it could provide the measure against which the HMM based approach should be tested. However, the standard way of applying such tests is to use the *AXB* or something akin to the *AXB*. The problem with this is that the continuous nature of accommodation is then lost since the data must be categorised into 'before' and 'after' groups (ie. before interaction vs. after interaction). This relies on carry over effects into a post-task (which this experiment doesn't have) or on comparing early utterances with late ones. So, the ability to represent the effect of local speech events on their surrounding context is lost. Further to this, it may be the case that some accommodative behaviour is unconsciously produced and unconsciously adapted to through the linked production perception routes (see subsection 2.3.1). If this is something that takes place, then it is unlikely that human listeners will be able to detect this, even if they are from outside of the interaction. Machine learning approaches on the other hand, base their classifications on the available data and do not base these classifications on human perception. Whilst the application of this data to a perceptual study would most certainly yield some interesting results, it is difficult to say how they could be directly compared to the HMM based results presented here.

Further to considerations about the role that different methods play in the detection of accommodation, it appears that both methods point to a similar finding. This is that it is likely that the content of an interaction tends to hold a relationship with the type of accommodation employed. This relationship, in the context of the experiment presented here, appears to be that if a task is more difficult, then speakers are more likely to converge. This is evidenced through the fact that although small, the only significant effects found for VOT and F1/F2 of vowels was in the comparison made with interaction length. Further to this, the amount of variance explained in the data was consistently higher for interaction length than for presentation position. Given that the data for each analysis are the same, this suggests that interaction length is a better predictor of the difference in speech between speakers carrying out this collaborative task than presentation position. This finding was mirrored in the HMM results for the GMMs and the Left-Right models. What this demonstrates is that the overall amount of time spent with someone, over the course of an experiment lasting a few hours, has less of an impact on accommodation than the immediate context in which someone is engaged. This makes sense from both a social and information processing point of view. The experiment was designed so that longer interactions necessarily indicated tasks that participants naturally found more difficult. Recall that the participants could not see each other, so no non-verbal communication was possible. Thus, the only mechanism that they could employ in order to complete a given scene and proceed to the next (or complete the experiment) was to modulate their voices. If it is believed that the participants both wished to complete the tasks as quickly as possible and that to do so required efficient information transfer then it is reasonable to assume that they would aim to

reduce social distance and improve processing through convergent behaviour. The reduction in social distance through convergence could be being used to indicate that they are both working towards a shared goal whilst the potential improved processing helps to reach that goal through alignment of internal representations. As such these findings provide support for this premise which will be tested in chapter 4.

In addition to contrasting a traditional and a machine learning approach to accommodation detection, this experiment also attempted to group participants by personality and/or interpersonal attraction. Since this is discussed in section 3.4, only a brief recap will be provided here. It was thought that personality and degree of liking might have a bearing on accommodation. For example, it could have been the case that those participants who liked each other to a greater degree would have a tendency to converge more. There were no effects found in the BFI but Physical and Social attraction for the IA were found to differ between pairing conditions. However, it was concluded that the protocol used to group participants was susceptible to external factors and was not able to perform its task. The groups that the protocol placed participants into differed in the average age of both the overall group and the pairs within. Whilst the age of participants in general is unlikely to impact the rates of accommodation, age may impact the types of personality and levels of interpersonal attraction reported using the BFI and IA, respectively. Further to this, both forms of test suffered from a small sample size and the IA was also, necessarily, implemented in a non-standard manner. These two things are enough to cast doubt upon the outcome of the self-selection protocol and may provide explanations for some of the effects found in the IA. Because of these concerns, the findings related to the self-selection protocol were not taken further in this thesis. It is suggested that a more comprehensive screening of the participants or a tighter restriction on age might have helped in eliminating these issues. However, not linking the BFI and IA findings to accommodation will not have a major impact on the main goals of the thesis.

The main findings and conclusions that can be drawn from the work reported in this chapter are:

- Accommodation can be detected in a continuous interaction through the use of single phonetic measures but that effects tend to be weak.
- An HMM based approach is able to provide a holistic account of accommodation during a continuous interaction.
- Accommodation occurs across a number of speech features, as evidenced by weak effects in the acoustic-phonetic analyses and the holistic results of the HMM based approach.
- Accommodation might be more sensitive to local contextual factors than global factors relating to an interaction with another speaker.

# Chapter 4

# Neural Experiment

Chapter 3 tackled a number of issues linked to the detection of speech accommodation in a continuous engagement between two people. The key assertions and findings in the chapter were that:

- Accommodation occurs across a number of acoustic features, assessing them individually does not account for any interactions between the features. If accommodation is to be studied in its entirety, it will need to be studied with both holistic and more traditional (segmental/suprasegmental) methodologies.
- Since accommodation occurs continuously (there are continuous adjustments to various acoustic features during an interaction), the more traditional approach of comparing 'before and after' measurements to that of the interlocutor may be missing some of the accommodative behaviour. Considering accommodation on a continuous basis may allow for this gap in the assessment of this phenomenon to be filled.
- HMM-based approaches present a way to evaluate accommodation as a continuous holistic process. This was validated through the detection of relevant looking accommodation processes that aligned with the findings from a traditional segmental approach to accommodation detection.
- Whilst time is undeniably a factor in accommodation, the impact of other behavioural factors may have been overlooked due to a lack of sensitivity in investigation methods. There may be some suggestion that a shift towards investigating accommodation in relation to behavioural triggers will uncover more insight about the driving factors behind accommodation.

This chapter builds on and expands the assertions made in chapter 3 by addressing three main research questions:

1. Is an HMM based approach able to detect shifting trends in brain activity patterns relative to an interlocutor?

2. Is there a relationship between speech accommodation patterns and brain activity patterns between speakers?

3. Can the findings of the HMM based approach to accommodation detection presented in chapter 3 be replicated?

The main goal of this chapter is to investigate whether this approach can be extended to the evaluation of other bio-signals, namely brain activity. Because an HMM-based approach is non domain specific and can be applied to any signal that varies in relation to another dimension, it can also be used to measure the degree to which brain activity patterns from conversational interactants tend towards or away from one another. Further to this though, the chapter also performs a replication of the experiment for the HMM based approach to speech accommodation detection. This is done in order to further verify the applicability of an HMM-based approach for the detection and classification of accommodation in a continuous interaction.

However, before going on to consider the experiment itself, it is worth recapping the motivation for the experiment, presented in section 2.3, and the potential challenges that an experiment such as this might face, detailed in subsection 2.4.2. This is the aim of section 4.1. The remainder of the chapter is organised as follows. Section 4.2 reports the methodology employed in this experiment. This includes participant recruitment, the materials and lab arrangement for the experiment including details of the EEG hyperscanning set-up, the experimental procedure used to elicit speech and neural responses, the approach taken for speech transcription and data management, a recap of the method employed for HMM analysis of speech data and an explanation of the method employed to analyse the EEG data. Following this, the results of both the speech and EEG analyses are presented in section 4.3. This is broken up into two subsections, one dedicated to the speech analysis and the other to the EEG analysis. The EEG data are reported in relation to the speaker's utterances, the partner's utterances and in relation to all speech produced during the experiment. The final section, section 4.4, provides an interpretation of both the speech and EEG data and concludes with a general summary of the findings.

## 4.1 Introduction

The majority of the reasoning for this experiment can be found in section 2.3 although a brief recap will be offered here. The drive for conducting this experiment stems from the relationship between research into accommodation and the cognitive mechanisms that are assumed to underpin it (see: Pardo et al., 2016a). Of the key cognitive theories that are used to explain accommodation, the thread that runs through all of them is that they all suppose a non-trivial link between speech perception and speech production (Liberman & Mattingly, 1985; Fowler, 1986; Goldinger, 1998; Pickering & Garrod, 2013). Given that this feature is consistent across all of

the models, it is reasonable to assume that a perception-production link is feasible. If perception and production are linked, it is then possible that perception can feed into production. This is the essence of accommodation and provides a system for the influence of incoming speech on outgoing productions. However, the relationship between perception and production clearly isn't a simple one. There are a number of factors that can influence the perception-production link including a whole host of top-down processes (Davis & Johnsrude, 2007; Sohoglu, Peelle, Carlyon, & Davis, 2012). This means that production is not necessarily immediately influenced by perception due to additional filters applied to the perceptual signal prior to impacting on production. One potential bio-psychological mechanism that has been suggested to optimise processing and therefore could provide a way to optimise these filters, is the process of neural entrainment.

Neural entrainment is the coupling of brain activity to environmental stimuli (Henry & Obleser, 2012). This entrainment allows for the brain to predict the likely upcoming information that it is about to receive. If this prediction is correct the current settings that the brain is using to interpret the incoming signal are kept constant, otherwise a more appropriate train of activity is enacted through realignment of the oscillatory phase - the phase of locally firing neuronal populations (Peelle & Davis, 2012). What this suggests is that there might be a link between measurable neural activity and the speech signal of both the speaker and their partner. If the brain does entrain to the speech signal by way of phase adaptation then it could be expected that as convergence of speech signals occurs, so too does convergence of brain activity. Likewise for divergence, as speech signals become more dissimilar so too would the brain activity across the two speakers. If oscillatory entrainment is used by the brain as a method for increasing processing efficiency and given that this entrainment is being used to actively predict the incoming speech signal, a convergence of brain activity between speakers could be an indication that the speakers are actively predicting each other's speech. This would be in keeping with certain theories of communication such as the work proposed by Pickering and Garrod (2013) which posits that speakers actively seek to align their mental representations of a conversation. Given that the task used in this thesis, the DiapixUK task, is a spot-the-difference task where participants have to align their representations of the stimuli and find the differences, it is particularly well suited for an investigation of this type.

Whilst there may be theoretical backing to assuming a link between brain activity and accommodation, the practical implementation of an experiment to explore these theories must also be carefully considered. Section 2.4 explores some of the challenges that an experiment looking to investigate a phenomenon such as this might face. It also offered some suggestions for overcoming these challenges. These challenges fell into three categories:

- How to assess a continuous EEG signal for effects related to accommodation.

- How to eliminate muscular artefacts associated with speech from the EEG signal.
- How to perform an EEG hyperscanning experiment whilst recording speech.

Where section 2.4 offered a broad interpretation of the potential solutions each of these challenges, section 4.2 offers the specifics of the approaches employed in this thesis. However, a brief recap is again offered prior to presentation of those specifics.

The first of the three challenges to the practical implementation of an experiment such as the one proposed is to be able to ensure that reliable data is drawn from the continuous EEG signal. The most widely used method for interpreting EEG signals is to evaluate ERPs (eg. Campanella et al., 2002; Bieniek et al., 2012; Swaab et al., 2012; Henrich, Alter, Wiese, & Domahs, 2014; Davis et al., 2015). These are produced by exposing a participant to multiple trials containing the same stimulus and the averaging across those trials to uncover the specific neural activity that is produced by exposure to the given stimuli. This is necessary because the electrical signal detected by the electrodes on the surface of the scalp is weak and can be contaminated by a number of other electrical signals (eg. muscular, digital). The averaging process helps to filter out EEG data that is not a response to the presented stimulus. In an experiment that is evaluating the continuous interaction between two participants in a live interaction, the presentation of clear and repeated stimuli is not possible. As such, ERP based approaches are not suited to an experiment of this kind. The proposed solution to this problem is the same as is applied to the evaluation of accommodation as a continuous phenomenon. As discussed previously in this thesis, HMMs are able to characterise the general form of a continuous signal. In this case, the continuous signal is the ongoing EEG signal. An HMM can be used to produce a model of a participant's general EEG characteristics and this can be used in the same way as the models for speech data to evaluate the shift towards or away from the brain activity of their partner. Rather than averaging across trials, a probabilistic model of brain activity is produced for each participant and then extracts of brain activity taken from across the interactions are compared to both participant models and correlated with time to determine the total degree of accommodation demonstrated in the EEG signal in relation to participant speech. This is expanded upon in section 4.2.6.

The remaining two challenges are technical hurdles that a number of researchers have been engaged with solving (eg. Delorme & Makeig, 2004; Dumas et al., 2010; Delorme et al., 2011; Ganushchak et al., 2011; Bigdely-Shamlo et al., 2015). As a result of research into improvements for EEG data collection and interpretation, the challenges of eliminating muscular artefacts from speech and performing an EEG hyperscanning experiment whilst recording speech can be resolved with existing tools. Here it is suggested that muscular artefacts from speech can be effectively removed

through the application of EEGLAB (Delorme & Makeig, 2004) preprocessing tools for the identification and removal of suspicious signal variances. Further details of the exact application of preprocessing tools can be found in subsection 4.2.6. A parsimonious solution for the physical set-up of EEG hyperscanning experiment was presented by Dumas et al. (2010), this design is emulated in this thesis. The basic premise is that the EEG signals of two participants can be fed to the same amplifier, allowing for low latency time synchronisation of EEG signals. The output of the amplifier is then read as if it were a single participant and signals are separated back into the separate participants during processing. The addition of a signal carrying the speech of the participants meant that the inclusion of a system for synchronising multiple streams of information was required. This was solved through the use of an application called Lab Streaming Layer (LSL) which was developed to integrate multiple streams of data including EEG, video and audio (Delorme et al., 2011).

The remainder of this chapter is structured as follows. Section 4.2 presents the specifics of the issues discussed above in greater detail along with details of the methodology used in the experiment presented in this chapter. Section 4.3 offers the results of the experiment separated into the results of the speech data and the results of the EEG data. Since the aims of this chapter are concerned with the application of HMMs, the speech data do not have a phonetic analysis and are instead evaluated only using the HMM based approach used in chapter 3. Finally, section 4.4 provides a discussion of findings for both speech and EEG data.

## 4.2 Methodology

### 4.2.1 Participants

12 female participants (different to those recruited for the experiment in chapter 3) were recruited to take part in the study. They were assigned to one of 6 female-female (same sex) dyads, the participants' ages ranged from 20 to 65 (mean 36.33 years). They were all recruited from the city of Glasgow conurbation. The participants in this study were not acquainted with one another and were strangers to one another before participating in this study. All speakers have English as their native language and they were all screened to ensure normal hearing and normal or corrected to normal eyesight. Participants were compensated with a monetary payment of £15 per hour in return for their participation. All participants were assigned a 'participant code' which was used in place of their names in order to maintain their anonymity. Data regarding the location of the participants' local home district was also recorded. All participants were born and raised in the greater Glasgow area. No participants had lived outside of Glasgow for periods of more than three months. Participant data is summarised in Table 4.1.

| Pair Number | Age | Home Town | Location | Participant Code |
|---|---|---|---|---|
| Pair 1 | 65 | Cumbernauld | North-East | FGA12 |
| | 47 | Scotstoun | West | GDL01 |
| Pair 2 | 27 | Uddingston | East | AGY13 |
| | 37 | Cumbernauld | North-East | LTY01 |
| Pair 3 | 31 | Croftfoot | South | SWN03 |
| | 25 | Newton Mearns | South | KWE29 |
| Pair 4 | 27 | Rutherglen | South | RTN03 |
| | 34 | Cumbernauld | North-East | EJN28 |
| Pair 5 | 29 | Bearsden | North-West | KCN16 |
| | 33 | Drumchapel | North-West | JNE18 |
| Pair 6 | 53 | Cardonald | South-West | JEN08 |
| | 20 | Pollokshaws | South | TKA09 |

Table 4.1: Demographic and pairing information for participants in the neural experiment.

### 4.2.2 Experimental Set-up

The experiment performed here is an expansion on the experiment described in chapter 3 except with a few key differences. Firstly, the self-selection protocol proved to be too time consuming and led to high attrition rates so this was not included in this experiment. Secondly, because the self-selection protocol had been dropped, the BFI questionnaire no longer took place before participants were paired. It was instead completed by each of the participants on the day of the experiment whilst separated from their experimental partner. The final difference is that both participants in an experimental pair were fitted with EEG caps for this experiment. Other than these key differences, the experiment ran in the same way as described for the behavioural experiment in section 3.3.

Recordings were made in a sound attenuated, Faraday shielded booth. The booth contained a table with two Dell UltraSharp 1908FP, 19 inch flat screen LCD monitors, one placed at either end, angled away from one another. A divider was placed next to the table, between the two monitors such that when sat on either side of the divider, it was not possible to see the other screen. Microphone stands with AKG SE 300B pre-amps equipped with an AKG CK91 condenser capsule serving as a mono microphone, were placed at either end of the table and were angled towards the respective participant. These microphones are designed to suppress off-axis sound and served to minimise the amount of speech captured from the conversational partner. Each microphone was connected to a separate channel on a mixing desk and the respective channels were assigned to either the left or the right channel of a stereo channel. This provides a mono recording channel for each of the participants whilst

maintaining low-latency time stamping. This was then connected to the stimulus presentation PC through a pre-amp.

At the other end of the divider, the end furthest away from the table, there was a stand for the BrainVision PowerPack brand batteries and BrainVision amplifiers. Each amplifier had a dedicated battery pack and was connected to the BrainVision USB2 adapter interface through a fibre optic cable. Each participant had a dedicated amplifier/battery set-up and each feed was delivered through a dedicated fibre optic cable to the computer interface where they were combined into a 64-channel montage. The set-up essentially followed the standard protocol for performing a single subject EEG experiment with a 64-channel EEG kit. The difference was that the two 32-channel amplifiers that would have been used in conjunction to provide a 64-channel montage for one person are in fact being used to collect data from two separate participants. The data are then combined in the same fashion as a 64-channel single subject experiment. This helps to ensure accurate time-stamping of the incoming EEG data from both participants.

A schematic for the complete arrangement and set-up of the experiment can be found in Figure 4.1.

#### Computer Network Set-Up

Due to the build of the lab that was used to collect data, it was necessary to set up a small local network of computers. The network was linked by a model GS105v5 NETGEAR® ProSAFE® 5-port Gigabit Ethernet Switch, using LabStreamingLayer (LSL) (available from: https://github.com/sccn/labstreaminglayer) to collect incoming data streams from the stimulus PC sources. This network consisted of two computers:

1. Stimulus Presentation PC - Windows 7

   This PC was used to present the DiapixUK images to participants, using an in-house presentation script. It was also used to collect the audio data. The LSL app AudioCaptureWin was used to pull the data from the audio input source into a stream used by the sychronization PC running the LSL LabRecorder synchronisation program. This PC was also used to collect the EEG data from participants. It had the BrainVision Recorder software installed and the associated access keys needed in order to run the program. This software was only used to ensure that the electrode impedances were within acceptable levels. The actual recording of the EEG data was performed through the LSL app BrainAmpSeries. The EEG data was recorded as a 64-channel recording which would later be decomposed back into two 32-channel data sets.

2. Synchronisation PC - Windows 8

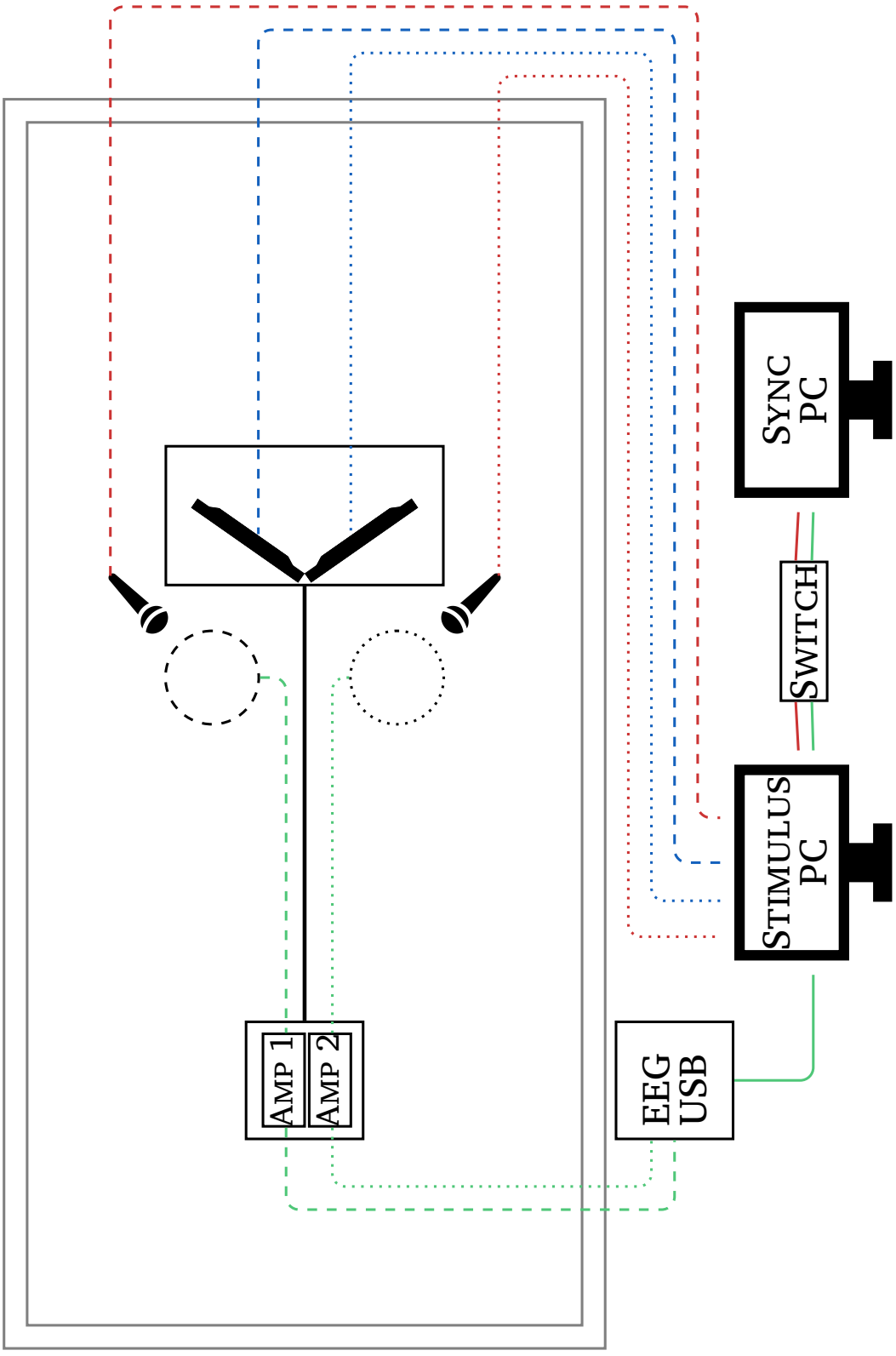



Figure 4.1: Physical set-up of experiment. Dashed lines represent all signals and cables associated with speaker A. Dotted lines represent all signals and cables associated with speaker B. Green lines represent EEG signals, red lines represent audio signals and blue lines represent visual stimuli signals.

This PC was used to pull together and time-stamp the different streams of data from the stimulus presentation PC. This was done using the LSL Labrecoder app. This PC was also used to continuously monitor incoming EEG signals to identify any major deviations in the signal so that adjustments could be made to the EEG electrodes to correct for this. This was completed using the vis_stream function from the LSL MATLAB® toolbox.

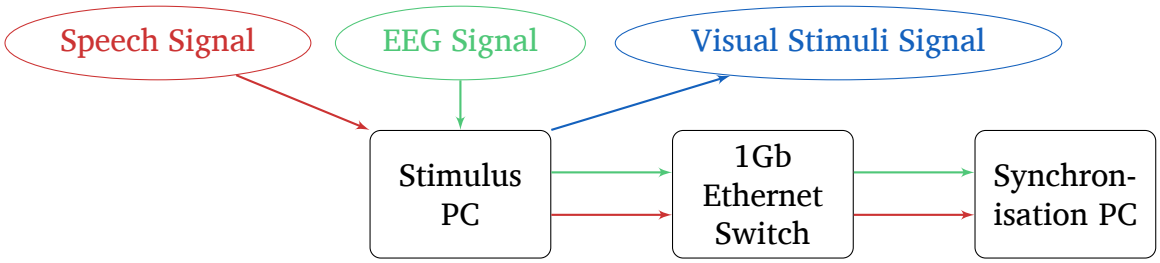



Figure 4.2: Conceptual schematic of the network set-up. Red lines indicate the route taken by audio speech recordings. Green lines indicate the route taken by EEG signal recordings. Blue lines indicate the route taken by the visual stimuli signals.

Figure 4.2 provides a conceptual representation of the network set-up. It presents the network arrangement from the point of view of the signal source, allowing easy identification of where each signal comes from and where it goes to.

### EEG Materials

Resource restrictions meant that only two 32-channel cap sizes were available. Both EEG caps were from the BrainCapMR Plus range offered by BrainVision. The two sizes used were the 58cm and 56cm circumferences. Because only these two sizes were available, when selecting a suitable cap size for each of the participants priority was given to the participant that had the closest match to either of the cap sizes.

The electrolyte of choice for this experiment was Abralyt 2000 HiCl paste. This was the recommended electrolyte for use with this particular brand of EEG recording cap. Having said this, SignaGel was also considered as a potential electrolyte but performed very poorly in tests.

The applicators used to apply the electrolyte consisted of a fine blunt needle and a plastic horned syringe. The blunt needle was used to scratch the surface of the scalp beneath each electrode and to separate any large sections of hair beneath an electrode. However, the needles were too fine to apply the electrolyte. To apply the electrolyte, plastic horned syringes with a wider diameter than the blunt needles were used.

To secure the EEG caps, participants had a choice of either a chin strap or a chest strap. Initially there were concerns that the chin strap may impede the ability of the participants to speak or that it might modify their speech characteristics. Because of this concern, every participant was offered a choice of strap to secure

their cap. Every participant elected to use the chin strap, no participants reported any significant discomfort or noticeable perturbations to their speech.

### 4.2.3 Experimental procedure

The set-up of the EEG recording system used two 32 channel amplifiers run through into one BrainVision USB interface box. This was done in the same way suggested by the manufacturer for a 64 channel EEG recording system using the same hardware (as represented by the green lines in Figure 4.1). When this is read into the BrainVision Recorder software, it is in the same way that a 64 channel system would be read. A custom montage file was written so that the electrodes from the first participant in the pair received the standard labelling of the electrodes whilst the second participant in the pair had their electrode names suffixed with a 'b' (eg. Fp1b, Ozb, Tp10b etc.). This allowed for easy differentiation of the which electrodes belonged to which participant.

Participants were invited into the lab and were asked to fill out the necessary data consent forms. They had the experiment explained to them, their hometown was recorded and they were offered a comfort break before proceeding. At this point the participants were separated and asked to complete the BFI questionnaires. After the BFI questionnaires had been completed, the EEG caps were fitted to the participants. The circumference of the participants heads were recorded using the nasion and inion locations on the head as reference points. These circumference values were used to select the best fitting EEG cap. Participants were asked to wipe their faces with alcohol prep pads, paying special attention to the forehead and under the eyes. The EEG caps were then fitted to the participants, ensuring that the Cz electrode was aligned with the vertex and that the Fp1 and Fp2 electrodes lay either side of the vertex. Abralyt 200HCl paste was then applied to the electrodes in the caps. Syringes were used to separate the hair in the centre of an electrode to expose the scalp, then paste was applied. This was repeated for all electrodes.

Once all electrodes had been filled with paste, the participants were taken into the recording booth and plugged into one of the two EEG amplifiers. Both the amplifiers and the fibre optic cables leaving the amplifiers had previously been indexed for easy identification. The BrainVision recorder software was then used to check the impedances of all of the electrodes for both participants. Adjustments were made to each electrode until all impedances were below $20\,\mathrm{k}\Omega$. The EOG and ECG electrodes were then prepared and applied to the participants. The EOG electrode was placed below the left eye after ensuring the area had been prepared with an alcohol wipe. Abralyt 200HCl paste was applied to aid with conductivity and the electrode was secured with microporous surgical tape. The ECG electrode was prepared in the same way as the EOG electrode except that it was placed on the centre left of participants' chests.

Once satisfied with the application of the electrodes, the EEG montage was checked to ensure good signals across all channels. Participants were checked to ensure their comfort and to issue specific instruction to minimise body movement. They were also asked to minimise eye blinks as much as possible given the nature and length of the experiment. Once satisfied with the comfort of the participants, the custom MATLAB® presentation script was started. It would initialise a LSL stream for manual timestamp markers of stimuli presentation onset then provide an opportunity to start up the necessary LSL applications. Nothing was yet presented to the participants. The LSL LabRecorder program was opened on the sync PC then LSL AudioCaptureWin and BrainAmp applications were opened on the stimulus presentation PC. The AudioCaptureWin application was linked to LabRecorder first, then the BrainAmp application was linked. Before the trials began, a visualisation of the incoming EEG data was opened in MATLAB® on the sync PC using the $\mathrm{vis\_stream}$ function from the LSL MATLAB® toolbox (Delorme et al., 2011). This was done to monitor the incoming signal throughout the experiment. After receiving confirmation from LabRecorder that all necessary applications were streaming data the MATLAB® presentation script was allowed to proceed.

The trials for the experiment were then started. The participants were initially presented with a set of instructions detailing the task, they were then presented with a practice trial to familiarise themselves with the task. Following the practice task the participants were asked to confirm that they understood the procedure and if they confirmed, they were allowed to proceed to the experiment proper.

For the experiment itself, pairs of DiapixUK images were presented to the participants on the monitors at a resolution of $1280 \times 1024$ pixels and a refresh rate of $50\,\mathrm{Hz}$. The images were presented as a single image spanning across the two monitors. The monitors shared the same workspace so the onset time for the images was the same for both participants. Each participant was presented with one half of a 'spot-the-difference' image, their partner was presented with the other half. A reminder of the types of image presented to the participants can be found in Figure 4.3. Each of the two images shown in Figure 4.3 would have been presented to one of the two monitors. For example, the image on the left in Figure 4.3 might have been presented to speaker *A* and the image on the right to speaker *B*.

The participants were tasked with finding the twelve differences between the images, using verbal communication only. They could not see each other's images. They were asked to begin in the top left hand corner of the image and work around in a clockwise direction. The trial ended either after the participants had found all twelve differences or after fifteen minutes had passed, whichever occurred first. Between each trial the participants were offered a break and if the EEG signal had degraded, adjustments were made to the EEG caps. The images for each trial were randomised for every pair of participants so as to minimise order effects. Upon completion of the DiapixUK task, participants were once again separated and were

asked to complete the McCroskey Interpersonal attraction questionnaire. Once this had been completed, the experiment was finished. Participants were then debriefed and compensated for their time.


Figure 4.3: An Example of the DiapixUK Stimuli. The Above Image Pair is from the 'Beach' Scene Category.

### 4.2.4 Transcription and Data Management

Transcription and data management were undertaken in the same way that was outlined in Section 3.3.3. The only slight change to the method used in Experiment 1 was that a greater number of transcribers were employed to reduce the total time required to complete the transcription. This meant that roughly half of the transcriptions were completed by the author. Random samples from each of the transcriptions were taken and reviewed to ensure consistency across all transcriptions. All data was uploaded to LaBB-CAT and force-aligned. Once uploaded and force-aligned, the data was once again checked for errors.

### 4.2.5 Speech Analysis

The analysis of the speech produced by the participants was handled in the same way as outlined in subsection 3.6.1. Because of this, only a brief reminder of the core aspects of the analysis will be presented here.

The methodology essentially considers the speech of any given speaker as a string of words uttered over the length of a given task element of DiapixUK. The speech signal associated with each word is extracted from the recording and is converted into a feature vector. For the purposes of this work, the feature vector selected were the first 12 Mel Frequency Cepstral Coefficients (MFCCs) of the speech sample. Conversion from a wave file to MFCC was performed by the Hidden Markov Model Toolkit (HTK). After the extraction of the feature vectors, speaker models were trained over all of the available data, ensuring that the parameters $\Lambda_A$ and $\Lambda_B$ in the probabilities $p(X_A, S_A|\Lambda_A)$ and $p(X_B, S_B|\Lambda_B)$ are maximised. $X_A$ represents the sequence of all feature vectors extracted from all words uttered by speaker $A$ (the

same for speaker $B$ in relation to $X_B$) and $S_A$ represents the state sequence associated with speaker $A$ (the same for speaker $B$ in relation to $S_B$). This is implemented using the Baum-Welch algorithm in HTK. Three types of HMM are produced, the first is a Gaussian Mixture Model (GMM) where the HMM has only one state. GMMs are time-independent, meaning that the value of the probability does not depend on the order of the observed feature vectors, and word-independent, meaning that the model is the same for all words. This type of model allows for the assessment of the general distribution of acoustic evidence for any given speaker. The second type of HMM implemented is a *left-right* model with a pre-defined number of states, here set at three. These models are time-dependent, meaning that the value of the probability does depend on the order of the observed feature vectors, but word-independent, the model is the same for all words. This type of model offers an insight into the temporal patterning of the observed acoustic evidence provided by speakers. The third type of model implemented here is a concatenation of left-right HMMs with each model corresponding to one of the phonemes that compose a given word. These models will be referred to as word-dependent models. The word-dependent models are time-dependent, the value of the probability depends on the order of the observed feature vectors, and word-dependent, meaning that there is a different model for every word. This third type of model allows for an evaluation of the change in acoustic evidence over time in relation to specific words. Once these models have been produced, the following log-likelihood ratio can be estimated for any given word $w_i^{(A)}$:

$$l_i(A,B) = \log \frac{p(X_i^{(A)}|\Lambda_B)}{p(X_i^{(A)}|\Lambda_A)} \tag{4.1}$$

where $w_i^{(A)}$ is the $i^{th}$ word uttered by speaker $A$ and $p(X|\Lambda_A)$ and $p(X|\Lambda_B)$ are the probability distributions of the HMMs trained over the data of speaker $A$ and $B$, respectively. When $l_i(A,B) > 0$, it can be said that the model for speaker $B$ better explains the speech utterances of speaker $A$ than the model of speaker $A$. Conversely, when $l_i(A,B) < 0$, it can be said that the model for speaker $A$ better explains the speech utterances of speaker $A$ than the model of speaker $B$. Because $l_i(A,B)$ can be estimated for every word uttered by A, it is possible to measure the following correlation:

$$L(A,B) = \rho[l_i(A,B), t_i^{(A)}] \tag{4.2}$$

where $\rho[.,.]$ corresponds to the *Spearman Correlation Coefficient* and $t_i^{(A)}$ corresponds to the $i^{th}$ time unit that corresponds to the position of the word $w_i$ uttered by speaker $A$. Of course the same can be estimated for speaker $B$ by redefining the

equation as $L(B,A) = \rho[l_j(B,A), t_j^{(B)}]$. In this way it is possible to estimate the degree of change towards or away from their interlocutor over time.

### 4.2.6 EEG Analysis

There are very few studies measuring EEG during a conversation between two people and this is linked to both the noise that speaking introduces to the EEG signal (Brooker & Donald, 1980; Vos et al., 2010) and to the limitations of traditional ERP analyses (eg. requiring multiple trials, see subsection 2.4.2). In order to tackle both of these issues, novel pre-processing of the EEG signal was required in addition to the application of an HMM-based approach for EEG analysis.

#### EEG Preprocessing

EEG data were preprocessed using Matlab® 2015a and the open-source toolbox EEGLAB 13.5.4b (Delorme & Makeig, 2004). Pre-processing of EEG data allows for the removal of signal noise that is known to relate to non-neurophysiological activity. Such sources of noise include, line noise from electrical equipment and muscular activity from eye/head movements. EEG data were down-sampled to 512 Hz using the $\mathrm{pop\_resample}$ function from the EEGLAB toolbox and the ECG channel was removed. This was done to both help reduce the size of the data (this helps with processing times, reducing the time to complete some processing steps from many hours to only a matter of minutes) and because the Nyquist frequency at this sample rate (256 Hz) is high enough to capture all relevant neural activity frequencies. Data was cleaned using Artefact Subspace Reconstruction from the $\mathrm{clean\_rawdata}$ extension (version 0.13), written by Miyakoshi and Kothe, available from the EEGLAB extensions web page (https://sccn.ucsd.edu/wiki/EEGLAB_Extensions_and_plug-ins). Parameters for cleaning were all left at default values except for Channel Noise detection and Line Noise detection which were both turned off in order to avoid channel rejection. Following cleaning of the EEG data the PREP pipeline (Bigdely-Shamlo et al., 2015) was run. Parameters provided to the PREP pipeline were:

detrendChannels = 1:30

detrendType = high pass

detrendCutoff = 1

lineNoiseChannels = 1:31

lineFrequencies = [12, 50, 100, 105, 135, 150, 200, 250]

referenceChannels = 1:30

evaluationChannels = 1:30

rereference = 1:31

rereferenceType = robust

ignoreBoundaryEvents = true

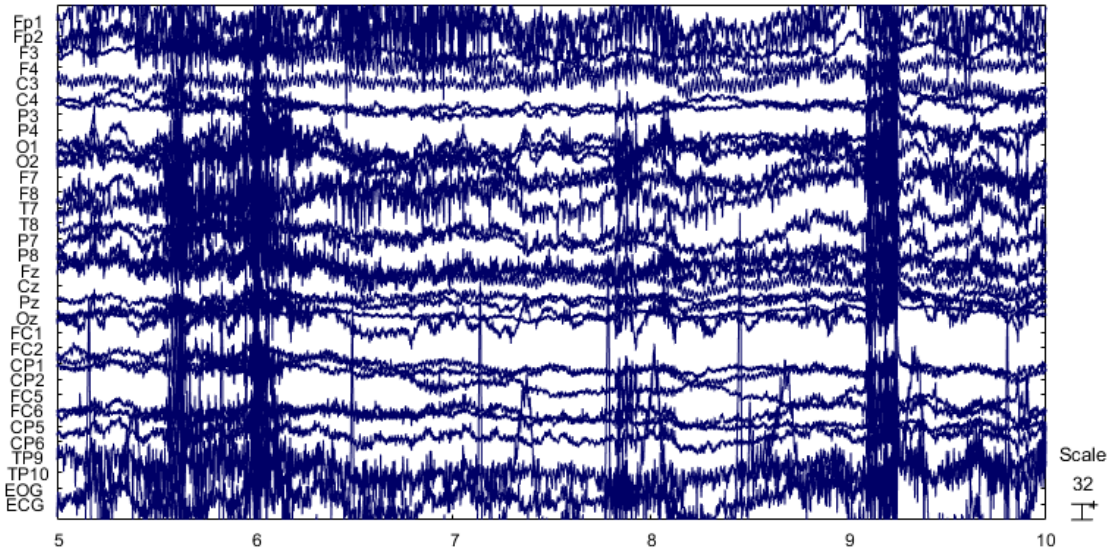
These parameters allowed the PREP pipeline to detrend all EEG channels (channel 31 was the EOG channel) with a high pass filter set at 1Hz. It also searched for and removed line noise in all channels at the frequencies listed in the lineFrequencies parameter. After completion of the above two steps, it performed a robust re-referencing of the data using only the EEG channels as reference and scanning only EEG channels for noise. However, the robust re-referencing was performed over all channels. The PREP pipeline was also asked to ignore boundary events since they serve no real purpose in this experiment other than to mark the onset of the stimulus presentation. The PREP pipeline allows for a number of pre-processing steps including high-pass filtering, line noise removal, signal referencing and the interpolation of bad channels to be automated.

Following pre-processing using the PREP pipeline, the data were then low-pass filtered at $40\,\mathrm{Hz}$ using a fourth order Butterworth filter on channels 1 to 30. This was performed using the $\mathrm{pop\_basicfilter}$ function from the ERPLab 5.0.0.0 plugin (Lopez-Calderon & Luck, 2014) for EEGLAB.

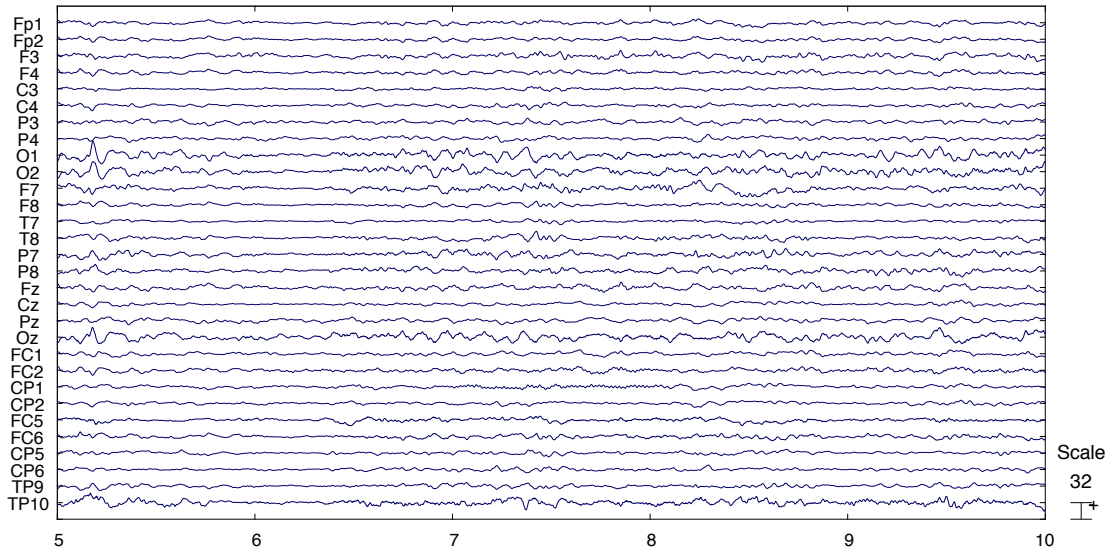
An ICA was performed on data to reduce data contamination from non-neural data sources (eye-blinks, muscle artefacts, electrical artefacts etc.). This was performed using the $\mathrm{pop\_runica}$ function from EEGLAB, as implemented in the infomax algorithm. Components representing data contamination sources were identified by visual inspection of component topographies, time-courses and amplitude spectra. Finally, the EOG channel was removed and the data were converted to text format in order to be processable by applications outside of MATLAB® (namely HTK and Python). Samples of the EEG time-course for both pre and post cleaning can be seen in Figure 4.4.

### Producing EEG Feature Vectors

LaBB-CAT was used to perform a search for all words uttered by each of the participants. The output of this search contained the time points for the onset of all the words that had been uttered by each participant. Because the audio and EEG data had been recorded at a low-latency and were synchronised by LSL, the same information used to identify the onset and duration of words could be used to extract corresponding segments of the EEG signal. In-house Python scripts were used to extract the corresponding sections of EEG that were aligned with both the words that the speaker uttered and the words that their interlocutor uttered. After these segments had been extracted the data were then transformed into feature vectors. MFCCs have been used to represent speech data in feature space, but this would




(a) EEG signal before cleaning




(b) EEG signal after cleaning

Figure 4.4: EEG signal samples for pre and post cleaning. Both samples are taken from the same period of participant AGY13's engagement in task BA_BB_1. Both samples have the same scale of $32\,\mu$V over 5 s of data. Figure 4.4b has had both the EOG and ECG channels removed. In figure 4.4a a lot of high frequency noise can be seen, as represented by rapidly fluctuating data. Additionally, it can be seen that there is a lot of signal drift in figure 4.4a, as represented by the data not centring around its associated channel. Both of these issues, along with some others, are not present in the cleaned data in figure 4.4b.

not be appropriate for the EEG signal. Although there is some precedence for using MFCCs for the vectorisation of EEG signals (Nguyen, Tran, Huang, & Sharma, 2012), it was considered to be inappropriate since the Mel filterbanks used when producing MFCCs are designed to represent the frequency scaling present in the human cochlea. In addition, when applying a novel analysis approach it makes sense to initially keep processing steps as simple as possible. For these reasons, the type of feature vectorisation selected was log power spectral density (PSD). The PSD describes how the energy contained in a signal is distributed with frequency. Other vectorisation methods may provide better coverage of the spectral characteristics of the EEG signal (see Gross, 2014 for an extensive review of appropriate methods for spectral analysis), the log PSD is a good starting point for applying this HMM-based approach. The log PSD considers the EEG signal in the frequency domain and given the literature surrounding the role that certain frequency bands might play in speech processing (Ghitza, 2011; Giraud & Poeppel, 2012; Doelling et al., 2014; Kösem et al., 2016), it is a sensible place to start.

The log PSD was extracted using the sciPy (Jones, Oliphant, Peterson, et al., 2016) ecosystem in Python (Python Software Foundation, 2016). Specifically, the welch function from the signal module was used. It requires the time series of a signal as input (in this case the EEG signal) along with the sampling rate of the signal, and the length of the segment, this was iteratively fed to the function based on the attributes of the current sample. The remainder of the function's parameters were left at default values. Hanning windows were used for the calculation of the PSD and the total number of sample windows used for calculation of the PSD was dependent on the length of the sample. This returned the PSD across the whole frequency range up to the Nyquist frequency (256 Hz for this data set) for all of the 30 EEG channels in each of the word-linked EEG samples. The data were then restricted to the 1 to 49 Hz range in order to exclude any line noise that might be present at the 50 Hz level and to exclude any activity that is unlikely to have a neurophysiological root (Iriarte et al., 2003; Olbrich, Jödicke, Sander, Himmerich, & Hegerl, 2011). The final step was to log-transform the resulting values. This step was performed simply to provide positive values to work with.

The process of creating the EEG feature vectors can be thought of as a transformation of the EEG data from the time dimension into the frequency dimension for each of the EEG channels. The final feature vectors used for each of the word-linked EEG samples contained 30 coefficients, each representing the power of the frequency that it is associated with. Where the speech feature vectors (MFCCs) summarised the information contained in the speech signal over time, the EEG feature vectors used here summarise the information in the EEG signal over frequency.

### Applying HMMs to the EEG Data

Since HMMs are non-domain specific, they can be applied to any signal that varies in relation to another dimension (eg. time). It is this property of HMMs that is utilised here to assess the change over time of the EEG signal in the participants. Upon completion of the EEG feature vectorisation, the remainder of the analysis follows the same pipeline as the application of HMMs to speech data.

Where the speech data analysis used the first 12 MFCC coefficients to maximise the parameters $\Lambda_A$ and $\Lambda_B$, here it is the full 30 coefficients of the $\log PSD$ (which relate to the 30 individual EEG channels) that are used. The speaker models for the EEG data were trained over all available data for each of the speakers such that $\Lambda_A$ and $\Lambda_B$ in the probabilities $p(X_A, S_A|\Lambda_A)$ and $p(X_B, S_B|\Lambda_B)$ are maximised. Dependent on the analysis being conducted, $X_A$ represents one of three sequences of EEG feature vectors drawn from the EEG signal of speaker $A$ (with $X_B$ representing the equivalent for speaker $B$):

- The sequence of all feature vectors extracted from the EEG signal occupying the same time frame as each word uttered by speaker $A$.
- The sequence of all feature vectors extracted from the EEG signal occupying the same time frame as each word uttered by speaker $B$.
- The sequence of all feature vectors extracted from the EEG signal occupying the same time frame as each word uttered by both speaker $A$ and speaker $B$.

It is important to highlight the fact that unlike the speech signal, the EEG signal continues across all of the utterances for both speakers. Therefore, feature vectors can be extracted from throughout the length of the interaction as long as there is a behavioural anchor to which it can be tied. Here the behavioural anchor being used are the words uttered by the speakers. The parameter maximisation for each of the above three sequences of feature vectors is implemented using the same Baum-Welch algorithm used in the speech analysis and is implemented in HTK.

For the EEG analysis only the Gaussian Mixture Model (GMM) and *left-right* models are produced. The exclusion of the word-dependent model for the EEG data was due to the comparatively low sampling rate of the EEG signal compared to the audio signal, $512\,\text{Hz}$ and $48\,000\,\text{Hz}$ respectively. Because the word-dependent approach develops models for each of the phonemes in a given word, it has to chunk the feature vectors associated with that word into groups analogous to each of the given phonemes in the word. This works given the high sampling rate of the audio signal because there will most likely be enough samples for each phoneme. However, with the $512\,\text{Hz}$ sampling rate of the EEG signal, the number of samples is greatly diminished. When this is coupled with the restriction of the samples to the $1$ to $49\,\text{Hz}$ range, the number of samples is reduced even further. This leaves little to no available data with which to estimate the probabilities associated with each phoneme.

These two models allow for the assessment of the general distribution of the neural evidence (GMMs) and the temporal patterning of the neural evidence (left-right).

After these models have been produced, the same log-likelihood ratio (Equation 4.1) used in the speech analysis can be applied to the EEG data for any given word. This can then be used to estimate whether the model for speaker $A$ or the model speaker $B$ best explains the neural activity produced by a speaker for any given word. Because the likelihood for the EEG feature vectors having been produced by either model $A$ or model $B$ can be estimated for every word the *Spearman Correlation Coefficient* defined in Equation 4.2 can also be used here. Thus, in the same way that change towards or away from their interlocutor over time can be estimated from the speech data, so too can this degree of change be estimated for EEG data.

## 4.3 Results

Results are presented for the speech data and EEG data. The results of all three HMMs (GMM, left-right and word-dependent) are provided for the speech data but only the results for the GMM and left-right HMMs are provided for the EEG data (as outlined in the final part of Section 4.2.6). This is due to the low sampling rate of the EEG signal meaning that there are a reduced number of samples available for use in each vectorisation window used to calculate the log PSD.

The different types of adaptation pattern are briefly recapped here. The adaptation patterns tested for can be grouped into four conditions:

- *Convergence*: Any statistically significant positive correlation.

  This constitutes the cases where both $L(A,B)$ and $L(B,A)$ are positive and statistically significant, where $L(A,B)$ is not statistically significant but $L(B,A)$ is positive and statistically significant and where $L(A,B)$ is positive and statistically significant but $L(B,A)$ is not statistically significant.

- *Divergence*: Any statistically significant negative correlation.

  This constitutes the cases where both $L(A,B)$ and $L(B,A)$ are negative and statistically significant, where $L(A,B)$ is not statistically significant but $L(B,A)$ is negative and statistically significant and where $L(A,B)$ is negative and statistically significant but $L(B,A)$ is not statistically significant.

- *Complementarity*: A statistically significant correlation for both speakers but in different directions.

  This constitutes the cases where $L(A,B)$ is negative and statistically significant but $L(B,A)$ is positive and statistically significant and where $L(A,B)$ is positive and statistically significant but $L(B,A)$ is negative and statistically significant.

- *Maintenance*: No statistically significant correlation for either of the speakers.

  This constitutes the cases where neither $L(A,B)$ or $L(B,A)$ are statistically significant.

Each of these possible adaptation cases is summarised in Table 4.2. Given that the classification of the convergence and divergence patterns both have the possibility of observing two instances of statistically significant correlations (i.e. $++$ or $--$), it is possible to have a greater number of overall significant correlations than the total number of tasks undertaken. For instance, if convergence is found in $12$ of the interactions undertaken by participants it is possible that the total number of statistically significant correlations leading to that result is $24$, provided that both participants converged to a statistically significant extent in all $12$ of those tasks. Equally, this result could have been brought about by just $12$ statistically significant correlations

if for each of the tasks only one of the participants expresses a statistically significant level of convergence. This is important to note since the interpretation of the ability of the HMM-based approach to detect accommodation patterns is determined by the number of statistically significant correlations that it detects. It is not determined by the number of tasks that it classifies into each accommodation pattern condition. The accommodation patterns consider the interaction of the dyad as a whole whereas the method for detecting if a significant change in the speech or EEG feature space has occurred is based on the speech or EEG signal of an individual in response to their interlocutor.

| | | $L(A,B)$ | | |
|---|---|---|---|---|
| | | $+$ | $=$ | $-$ |
| $L(B,A)$ | $+$ | Convergence | Convergence | Complementarity |
| | $=$ | Convergence | Maintenance | Divergence |
| | $-$ | Complementarity | Divergence | Divergence |

Table 4.2: This table provides an overview of the types of adaptation pattern that are identified in this work. $+$ indicates a statistically significant positive correlation, $-$ indicates a statistically significant negative correlation and $=$ indicates no statistically significant correlation.

### 4.3.1 Speech Data Results

The key questions that require answering here are:

- Whether the HMM-based approach is able to detect changes in speech patterning being expressed by the speakers.
- Whether speakers demonstrate the same trends that were observed in Section 3.6. In other words, can the results of experiment 1 be replicated?

In order to answer the first of the above two questions, the number of statistically significant correlations were counted and a binomial test was performed on the resulting figures. The results of which are presented in Table 4.3. The number of times at least one of the two correlations $L(A,B)$ and $L(B,A)$ is statistically significant with confidence level 0.05 (after Bonferroni correction) is $58$, $46$ and $36$ for the GMMs, left-right models and word-dependent models, respectively. According to the binomial test, the probability of achieving such a result by chance is lower than $10^{-7}$ in all cases. This suggests that the patterns detected by the HMM-based approach are related to changes in speech patterning being expressed by the individual speakers and are not the result of chance.

To answer the second of the two questions posed above, the results of the accommodation pattern classifications must be considered. These results are presented in Table 4.4 for each model type, along with the average durations of the tasks that they are associated with.

| Model Type | No. Significant Cases | $p$ |
|---|---|---|
| GMM | 58 | $< 0.001$ |
| Left-Right | 46 | $< 0.001$ |
| Word-Dependent | 36 | $< 0.001$ |

Table 4.3: Results of binomial test for the speech data in the neural experiment.

| | Model Type | Convergence | Divergence | Maintenance | Complementarity |
|---|---|---|---|---|---|
| Count | GMM | 15 | 18 | 30 | 9 |
| | Left-Right | 19 | 16 | 33 | 4 |
| | Word-Dependent | 19 | 9 | 42 | 2 |
| Avg. Duration (s) | GMM | $558 \pm 72$ | $405 \pm 36$ | $454 \pm 38$ | $575 \pm 95$ |
| | Left-Right | $616 \pm 62$ | $439 \pm 45$ | $440 \pm 36$ | $303 \pm 46$ |
| | Word-Dependent | $593 \pm 62$ | $530 \pm 64$ | $408 \pm 29$ | $642 \pm 257$ |

Table 4.4: This table reports the counts and average duration ($\pm$ the standard error) of tasks classified as either Convergence, Divergence, Maintenance or Complementarity. The Convergence condition (at least one of the two speakers converges towards the other to a statistically significant extent) is associated to tasks that require longer time to be addressed. Values have been rounded to the nearest second.

Looking at the table, it is clear that for all model types (GMM,left-right and word-dependent), convergence tends to take place in tasks with a longer average duration. When the average durations of the tasks associated with convergence are compared to each of the other accommodation patterns this statement holds for divergence and maintenance but not complementarity. However, the total number of tasks classified as complementarity are too few to draw definite conclusions from. The same could be said for the number of tasks classified as divergence in the word-dependent model. This might go some way to accounting for the comparatively high average duration seen in the divergence pattern for the word-dependent model.

Figure 4.5 offers a representation of the whole data set rather than just the means. The data are represented as a series of bubble plots, one for each of the three HMM types implemented in this experiment (GMM, left-right and word-dependent). The size of the bubbles is proportional to the length of the task that it represents. Overall, the bubbles representing tasks where convergence has taken place can be seen to be larger than the bubbles representing divergence and maintenance. The same can also be said about convergence in relation to complementarity but, as discussed above, there are a far smaller number of cases. Both Figure 4.5a and Figure 4.5b seem to show fairly similar trends with cases of convergence tending to be larger than the cases of divergence or maintenance. There are some instances where the size of the bubbles for convergence are far smaller than those seen in divergence and maintenance but the general trend still holds. Looking at Figure 4.5c, it can be seen that of the nine cases of divergence observed, three are substantially longer than the others. Given that this accounts for 33% of the cases of divergence observed they may very well be skewing the results towards higher duration values. Having said

that, a good number of cases of convergence can still be seen to be larger than that of both the divergence and maintenance patterns. Taken together, these plots support the assertions made based on the averaged data in Table 4.4. The data continue to suggest that different adaptation patterns would seem to take place during tasks with different average lengths.

The fact that this trend seems to hold across all of the models tested would indicate that the relationship between the length of time taken to complete the task and the type of accommodation demonstrated in that task holds across multiple aspects of the speech signal. Where the GMMs represent features in the general distribution of acoustic evidence in the feature space, the left-right models represent the temporal patterns demonstrated in the data and the word-dependent models representing the words uttered during the task. Having said that, it can also be seen that as the HMMs become more constrained (from time- and word- independent through to time- and word-dependent), the number of maintenance cases increases. This suggests that as accommodation is required to take place on more levels, patterns associated with accommodation become too weak to be observed, if any.

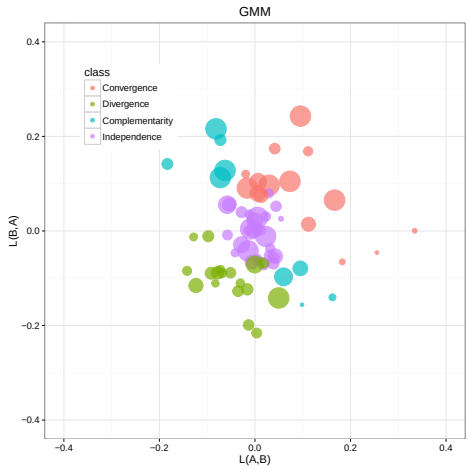
Taking this interpretation further unpaired, two-tailed t-tests were conducted to ascertain if differences between interaction length for accommodation classification groups are significant. The results for these can be seen in Figure 4.6 along with the standard error of the mean for each type of adaptation pattern. The dotted horizontal line represents the mean length of all tasks. These results are also presented in Table 4.5

For the GMM models, comparisons for interaction length were performed between the convergence and divergence conditions, the convergence and maintenance conditions and the divergence and maintenance conditions. No comparisons were made with the complementarity condition as there are too few cases to provide statistically reliable indications. For the comparisons that were made, no statistically significant differences were found at the 5% level. Having said this, the p-value for the comparison between convergence and divergence was approaching significance at $0.057$.

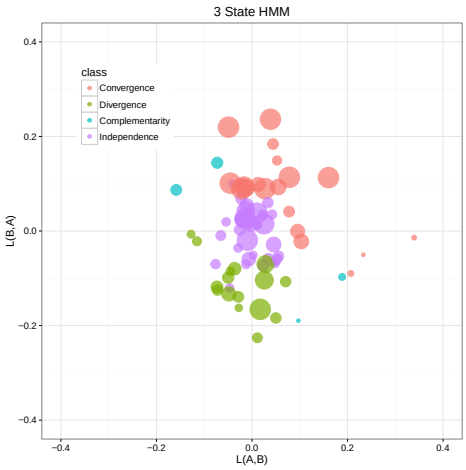
The same comparisons that were carried out for the GMM models were carried out for the left-right models. Here we find that the difference between convergence and divergence conditions to be statistically significant with $p < 0.05$. The difference between convergence and maintenance was also found to be statistically significant at the 1% level with $p = 0.01$.

For the word-dependent models only the convergence and maintenance conditions contained enough cases to provide statistically reliable indications, so this was the only comparison performed. The difference between convergence and maintenance was found to be highly significant at less that the 1% level with $p = 0.003$.

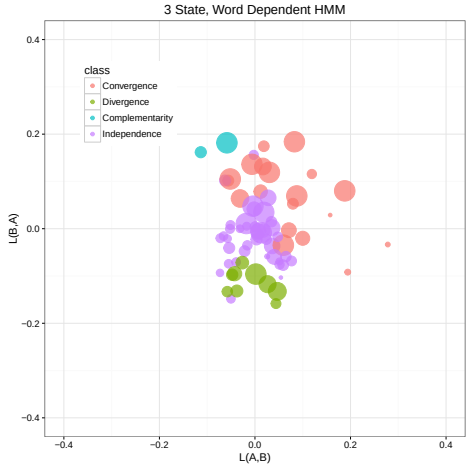
Whilst there are some differences between the results presented here and the results presented in Section 3.6, the broad trends appear to hold. Instances of con-




(a) Gaussian mixture model (GMM) for speech data




(b) Left-right HMM for speech data




(c) Word-dependent HMM for speech data

Figure 4.5: Bubble plots for the different HMMs used to classify accommodation in the speech data of the neural experiment. Each bubble represents a single interaction and the size of the bubbles are proportional to the length of the interaction. Colours indicate accommodation categories.

vergence still appear to occur in tasks with longer average durations. Whilst the comparison between convergence and divergence for the GMM models may not be significant at the 5% level, it is approaching significance. The patterns observed for the left-right models in Section 3.6 are replicated here. The fact that the difference between the maintenance cases and the convergence cases is so highly significant suggests that when accommodation is considered across a number of levels (as it is for a word-dependent HMM), greater task duration is more strongly related to convergent behaviour. The opposite could also be said for shorter task durations being more strongly linked to independent behaviour.

### 4.3.2 EEG Data Results

Here, the results of the HMM-based analyses of the EEG signal are presented. Again, it should be noted that only the GMM and left-right models are considered in these

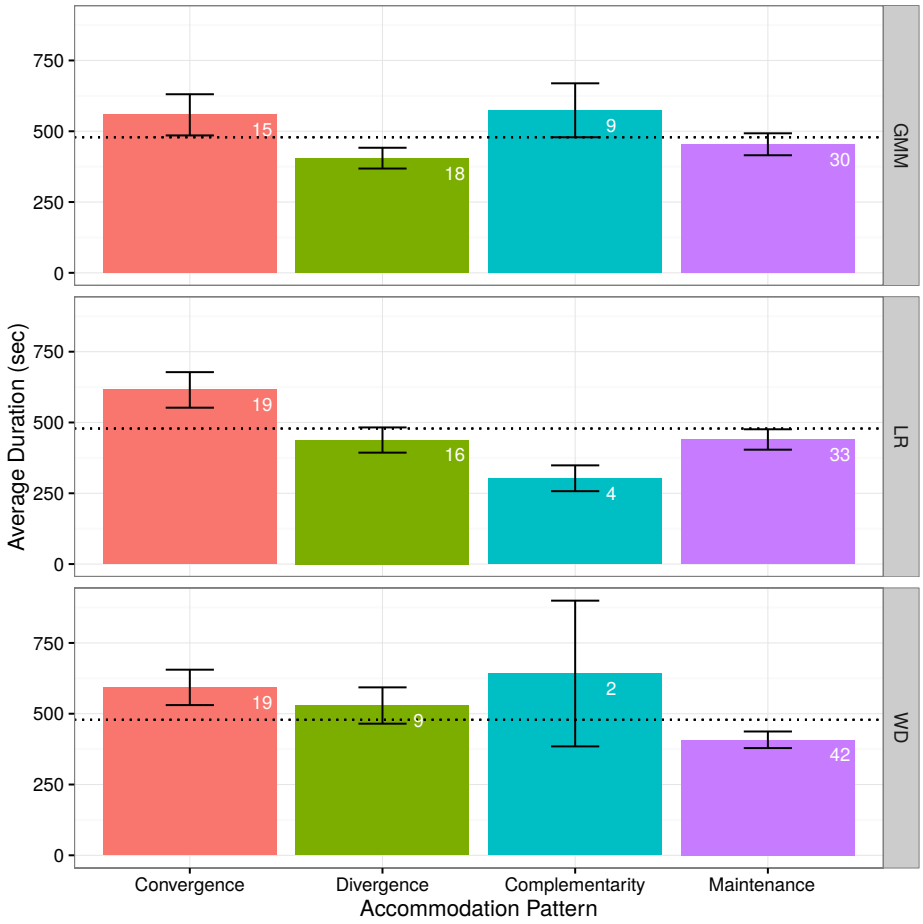



Figure 4.6: Average duration of tasks for each of the four possible adaptation patterns. Error bars represent the standard error of the mean. Numbers in each bar indicates the number of observations for that adaptation pattern. Dotted black line indicates the mean duration of all interactions. GMM = Gaussian mixture model (single state HMM), LR = left-right HMM, WD = word-dependent HMM.

| Model Type | Comparison | $p$ |
|---|---|---|
| GMM | convergence:divergence | $0.057^{+}$ |
| | convergence:maintenance | 0.17 |
| | divergence:maintenance | 0.4 |
| Left-Right | convergence:divergence | $0.03^{*}$ |
| | convergence:maintenance | $0.01^{**}$ |
| | divergence:maintenance | 0.98 |
| Word-Dependent | convergence:maintenance | $0.003^{***}$ |

Table 4.5: This table presents the t-test results for comparisons between accommodation category classification by the different HMM types and interaction length for speech data.

analyses for the reasons outlined in the final part of Section 4.2.6.

This section is separated into three parts, the first part evaluates the degree of change in the EEG feature space associated with the words uttered by the speaker themselves. The second part evaluates the degree of change in the EEG feature

space associated with the words uttered by the speaker's partner. The final part evaluates the degree of change in the EEG feature space associated with both the words uttered by the speaker and the words uttered by their partner. Aside from the exclusion of the word-dependent model, the results in each of the three parts of this section are presented in the same way as the results of the speech data.

The accommodation pattern types outlined at the beginning of this results section still hold for the EEG data.

#### EEG Data Results: Self

The data presented in this part of the EEG results section concern observed accommodation patterns in the EEG feature space associated with the word uttered by the speaker themselves. It should be noted that the data presented here are the data most likely to still contain noise from the muscular activity associated with speech production.

The key questions to be addressed here are:

- Whether the HMM-based approach is able to detect changes in the EEG signal of the speakers for the periods when they themselves are speaking.
- Whether speakers demonstrate the same trends in EEG patterning that they do in speech patterning.

To address the first question, the same approach as was taken for the speech data was applied. The number of statistically significant correlations were counted and a binomial test was performed on the resulting figures. The results of this test are presented in Table 4.6. The number of times at least one of the two correlations $L(A,B)$ and $L(B,A)$ is statistically significant with confidence level 0.05 (after Bonferroni correction) is $37$ and $43$ for the GMMs and left-right models, respectively. The probability of achieving such a result by chance is lower than $10^{-7}$ in both cases. This suggests that the patterns detected in the EEG signal by the HMM-based approach are related to changes in neural patterning being expressed by the individual speakers and are not the result of chance.

| Model Type | No. Significant Cases | $p$ |
|---|---|---|
| GMM | 37 | $< 0.001$ |
| Left-Right | 43 | $< 0.001$ |

Table 4.6: Results of binomial test for the speakers' own EEG data samples.

In order to address the second question, just as for the speech data, the results of the accommodation pattern classifications must be considered. These results are presented in Table 4.7 for both model types, along with the average durations of the tasks that they are associated with.

| | Model Type | Convergence | Divergence | Maintenance | Complementarity |
|---|---|---|---|---|---|
| Count | GMM | 18 | 13 | 40 | 1 |
| | Left-Right | 16 | 16 | 36 | 4 |
| Avg. Duration (s) | GMM | $561 \pm 62$ | $409 \pm 40$ | $467 \pm 37$ | $340 \pm NA$ |
| | Left-Right | $525 \pm 68$ | $407 \pm 49$ | $481 \pm 37$ | $554 \pm 129$ |

Table 4.7: This table reports the counts and average duration ($\pm$ the standard error) of the speakers' own EEG data samples providing a task classification of either Convergence, Divergence, Maintenance or Complementarity. The Convergence condition (at least one of the two speakers converges towards the other to a statistically significant extent) is associated with tasks that require longer time to be addressed. Values have been rounded to the nearest second.

As with the results of the speech data, the table demonstrates that for both model types (GMM and left-right), convergence appears to take place in tasks with a longer average duration. This trend appears to hold for both model types when convergence is compared to divergence and maintenance. Making comparisons against complementarity would not be valid since the number of cases is so small. The average duration for each of the accommodation pattern conditions (excluding complementarity) look to be somewhat more consistent between the two types of model than the average durations observed for accommodation pattern conditions in the speech data.

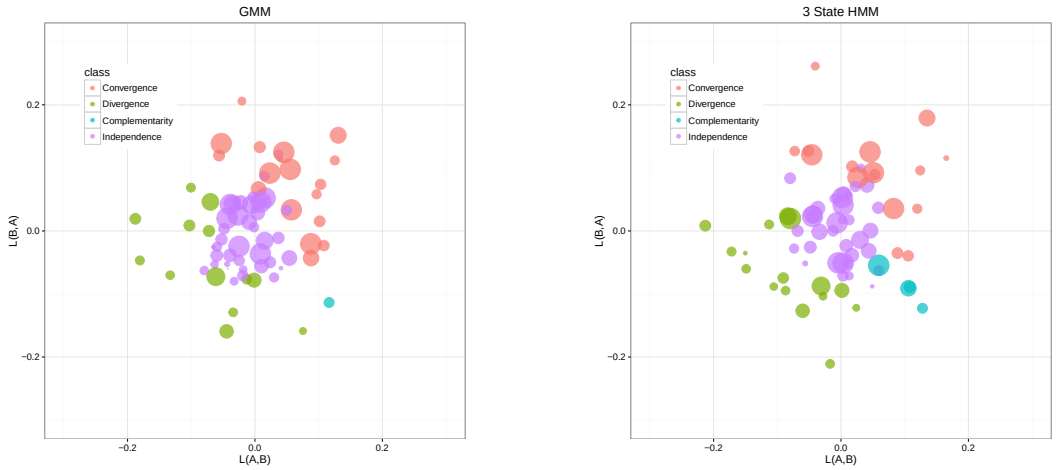
Broadly, the data presented in Table 4.7 resemble that of the results for the speech data in Table 4.4. One slight exception is that the number of cases for each accommodation pattern category follows the opposite pattern to that observed in the speech data. Upon moving from GMMs to left-right models, cases of convergence and maintenance decrease whilst cases of divergence and complementarity increase. However, the degree to which this change in patterning is indicative of changes in the observed patterns in the EEG signal is difficult to assess.

Again, the data presented in Table 4.7 concerns the average values for the obtained data. Figure 4.7 offers an insight into the distribution and trends for both the GMMs and left-right models in the EEG data.

In these bubble plots, the size of the bubbles represent the lengths of the tasks that they are associated with. Larger bubbles represent longer interactions, smaller bubbles represent shorter interactions.

In general, the bubble plots for both models appear to demonstrate convergence being related to tasks with longer durations. This appears to be especially true when the lengths of convergence instances are compared against that of the divergence instances. The degree to which this is true when convergence is compared to the maintenance conditions, is less clear.

There don't appear to be many major differences between the GMM data in Figure 4.7a and the left-right data presented in Figure 4.7b. The main observation about the difference between the two models is that the spread of the data across




(a) Gaussian mixture model (GMM) (b) Left-right HMM

Figure 4.7: Bubble plots for the speakers' own EEG data samples associated with the word uttered by the speaker. Each bubble represents a single DiapixUK task. The size of the bubbles is related to the length of the task they represent. Figure 4.7a presents the data for the GMMs and Figure 4.7b presents the data for the left-right models.

the correlation space seems to be larger for the left-right models. This is suggestive of the left-right models providing a slightly better categorisation of accommodative patterns in the EEG signal. The fact that the number of maintenance cases decreases when moving from GMMs to left-right models provides added support for this claim.

This would suggest that the accommodative patterns that are observed are consistent across both the general distribution of EEG signal evidence in the feature space and the temporal patterning present in the data associated with the words uttered by the speaker because the data pattern similarly for both model types.

Looking at Figure 4.8, the same general trends that were present in the speech data are also found here. Cases of convergence tend to occur in tasks with longer average durations than those observed in the divergence and maintenance conditions. This difference between the conditions does however seem to be less prominent than observed in the speech data. Results of t-tests between the values for each of the conditions are presented in Table 4.8 and confirm that the differences between conditions are not significant. Having said that, p-values for the difference between convergence and divergence are the closest to 0.05 of all the comparisons made for each of the model types. For the GMMs, the comparison between convergence and divergence conditions is approaching significance at $p = 0.068$. Given the potential level of noise in the data and the rather course nature of the feature vectorisation, this is rather suggestive that the patterns in the neural data may very well reflect those found in the speech data. On the other hand, because this data is drawn from sections of the EEG signal that are associated with the actual speech of the participant in question, these trends could be being brought about by correlations with the contamination in the EEG signal from muscular artefacts. By the same token though, this would suggest that patterns in the feature space of the muscular activ-

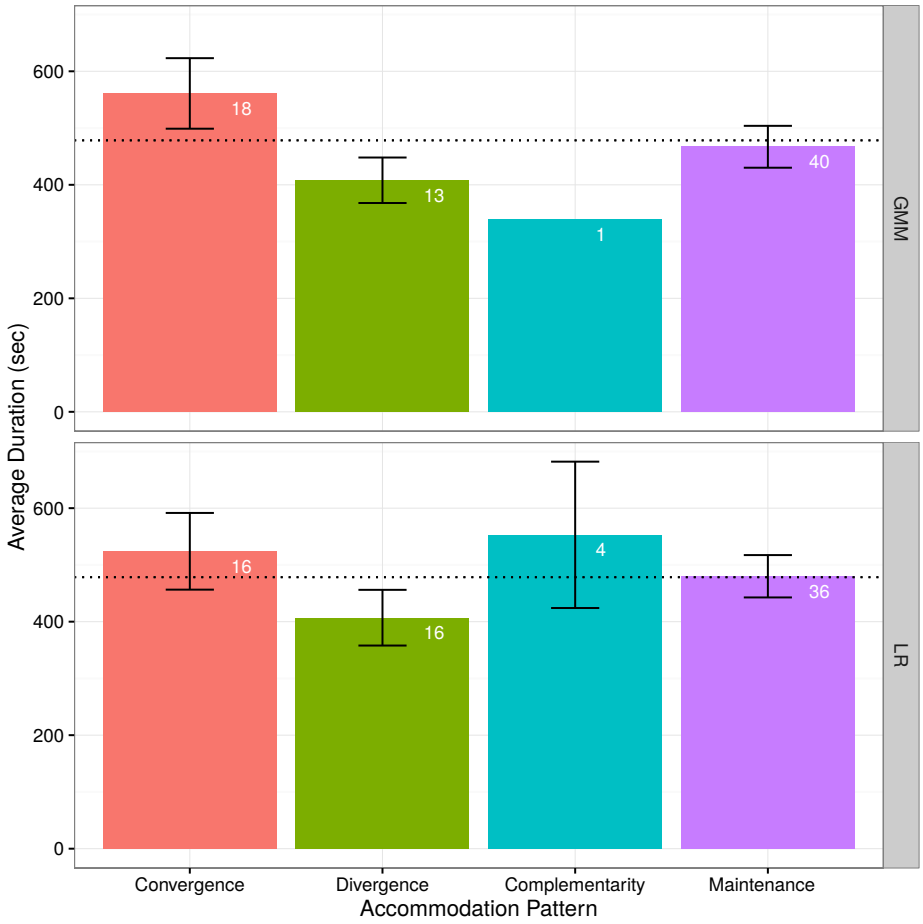



Figure 4.8: This graph presents the classification of EEG data from speakers' own EEG data samples in relation to the duration of the interactions. Error bars represent the standard error of the mean. Numbers in each bar indicates the number of observations for that adaptation pattern. Dotted black line indicates the mean duration of all interactions. GMM = Gaussian mixture model (single state HMM), LR = left-right HMM.

ity detected by the EEG electrodes demonstrate similar patterns to that of the speech data.

| Model Type | Comparison | $p$ |
|---|---|---|
| GMM | convergence:divergence | 0.068 |
| | convergence:maintenance | 0.176 |
| | divergence:maintenance | 0.399 |
| Left-Right | convergence:divergence | 0.169 |
| | convergence:maintenance | 0.545 |
| | divergence:maintenance | 0.261 |

Table 4.8: This table presents the t-test results for comparisons between accommodation category classification by the different HMM types and interaction length for the speakers' own EEG data samples.

**EEG Data Results: Partner**

The data presented in this part of the EEG data results section are taken from the segments of the EEG signal that are associated with the words uttered by the partner of a given participant. Given that turn-taking is generally quite consistent in the corpus, using these segments of the EEG signal should contain less of the noise associated with the muscular activity from speech production.

As with the previous section using the EEG signal associated with the speech of the speaker, there are two key questions to evaluate in this section:

- Whether the HMM-based method can detect patterns of accommodation in the EEG signal associated with the words uttered by a speaker's partner.
- Whether the adaptation patterns found for the EEG signal follow the trends of accommodation patterns found in the speech data.

Once again, the approach for evaluating whether the HMM-based approach is detecting real trends in the data and not random fluctuations is based on conducting a binomial test on the counts of statistically significant correlations detected. The number of times at least one of the two correlations $L(A,B)$ and $L(B,A)$ is statistically significant with confidence level 0.05 (after Bonferroni correction) is 57 for both the GMMs and the left-right models.

| Model Type | No. Significant Cases | $p$ |
|---|---|---|
| GMM | 57 | $< 0.001$ |
| Left-Right | 57 | $< 0.001$ |

Table 4.9: Results of binomial test for the partner's EEG data samples.

According to a binomial test, the probability of achieving such a result by chance is lower than $10^{-7}$ in both cases. Results can be found in Table 4.9 for both model types (GMM and left-right). This result suggests that the HMM-based approach is detecting real trends in the data that are not the result of chance.

Making an assessment of the similarity of trends in the EEG data to that found in the speech data requires an evaluation of the relationship between the number of observed instances of each accommodation pattern condition and the duration of associated tasks. The counts for each condition and the associated average durations are presented in Table 4.10.

The trend for convergence being more likely to be present in tasks with longer average durations looks to still be present in this data although to a much smaller degree. The difference between the average durations found for convergence and those found for maintenance are far smaller than those found in the previous analyses. For the difference in convergence and maintenance durations in the GMMs it is hard to say whether this difference really exists at all. Having said that, the number

| | Model Type | Convergence | Divergence | Maintenance | Complementarity |
|---|---|---|---|---|---|
| Count | GMM | 12 | 27 | 26 | 7 |
| | Left-Right | 19 | 20 | 27 | 6 |
| Avg. Duration (s) | GMM | $492 \pm 77$ | $418 \pm 35$ | $483 \pm 44$ | $669 \pm 110$ |
| | Left-Right | $489 \pm 59$ | $419 \pm 38$ | $437 \pm 39$ | $829 \pm 63$ |

Table 4.10: This table reports the counts and average duration ($\pm$ the standard error) of the partner's EEG data samples providing a task classification of either Convergence, Divergence, Maintenance or Complementarity. Values have been rounded to the nearest second.

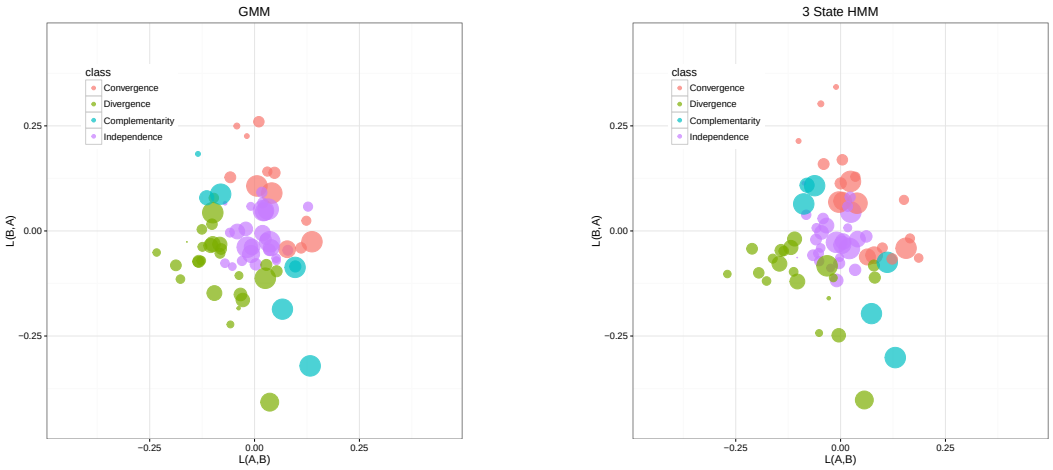
of cases of convergence does seem to be small in comparison to those found for divergence and maintenance for the GMMs. It may be possible that this is constraining the effect to some extent. However, the difference between the duration of convergence cases and divergence cases is held at a consistent distance which is consistent with previously observed trends. Interestingly, the durations related to the divergence cases can also be seen to tend to remain below the durations associated with maintenance cases. Again, the cases of complementarity are too few to sensibly interpret. One trend that holds between this data and the EEG data associated with the speech of the participant is the pattern of increase and decrease in counts when moving from GMMs to left-right models. Just as in the previous EEG data, convergence and maintenance counts increase whilst divergence and complementarity counts decrease. Admittedly, the change in maintenance and complementarity counts is only a step of one and this is may not be meaningful.

The number of counts themselves is also worth noting. There are fewer instances of maintenance observed here than in previous analyses. There has also been an increase in the number of divergence cases observed. The fact that the number of maintenance cases is comparatively few might suggest that the data obtained from the sections of EEG signal associated with the speech of the speech partner might be more conducive to detecting trends in the data.

Data representing the individual tasks that make up the data set can be found presented as a bubble plot in Figure 4.9. Each bubble represents a given task completed in the experiment, the size of the bubbles is relative to the length of the task. Larger bubbles indicate longer tasks and smaller bubbles indicate shorter tasks.

Looking at the data presented in the bubble plot, it appears far harder to conclude that convergence can generally be observed to be taking place in tasks with longer durations. For both of the models presented, although there are some instances of longer tasks observed for convergence, it is hard to conclude that the lengths of the tasks where convergence takes place are consistently longer than those of divergence or maintenance. Indeed, even the cases of complementarity, although few, appear to be mostly equal to or longer than the cases of convergence.

Patterns for the data in both of the models look remarkably similar, this may indicate that findings are consistent across the two models. Finding consistency




(a) eeg scatter GMM (b) eeg scatter LR

Figure 4.9: Bubble plots for the speakers' own EEG data samples associated with the word uttered by the speaker. Each bubble represents a single DiapixUK task. The size of the bubbles is related to the length of the task they represent. Figure 4.9a presents the data for the GMMs and Figure 4.9b presents the data for the left-right models.

between the model representing the general distribution of evidence in the EEG feature space (GMM) and the model representing temporal patterns in the data (left-right) suggests a distribution of trend evidence across the feature space. Figure 4.10 provides a graphical representation of the general trends in the data.

In general, there doesn't appear to be any large differences between the accommodation pattern conditions aside from the complementarity condition. Although the complementarity condition only has a few cases for each of the models, they do appear to have a consistently higher average duration than the other conditions.

Of the differences that are observable, divergence does tend to have a lower average duration associated with it than the convergence cases. The divergence cases also demonstrate consistently lower average durations than the overall average duration of the tasks. Values for the average duration associated with convergence appear to pattern around the average duration across all tasks. This suggests that whilst cases of convergence may take place in tasks with a longer length than that observed in divergence cases, the overall average duration of convergence cases does not differ from the average task length.

Cases of maintenance in the GMMs carry a similar average duration to convergence whereas in the left-right models they look more comparable to the cases of divergence. Given the good number of cases of maintenance for each of the models and the consistency of the convergence and divergence cases across the models, this may suggest a greater level of variation in the data trends leading to maintenance classification.

The trends present in the data are similar to that previously observed but are somewhat more modest. The t-tests conducted on the data generally confirm this and are presented in Table 4.11.

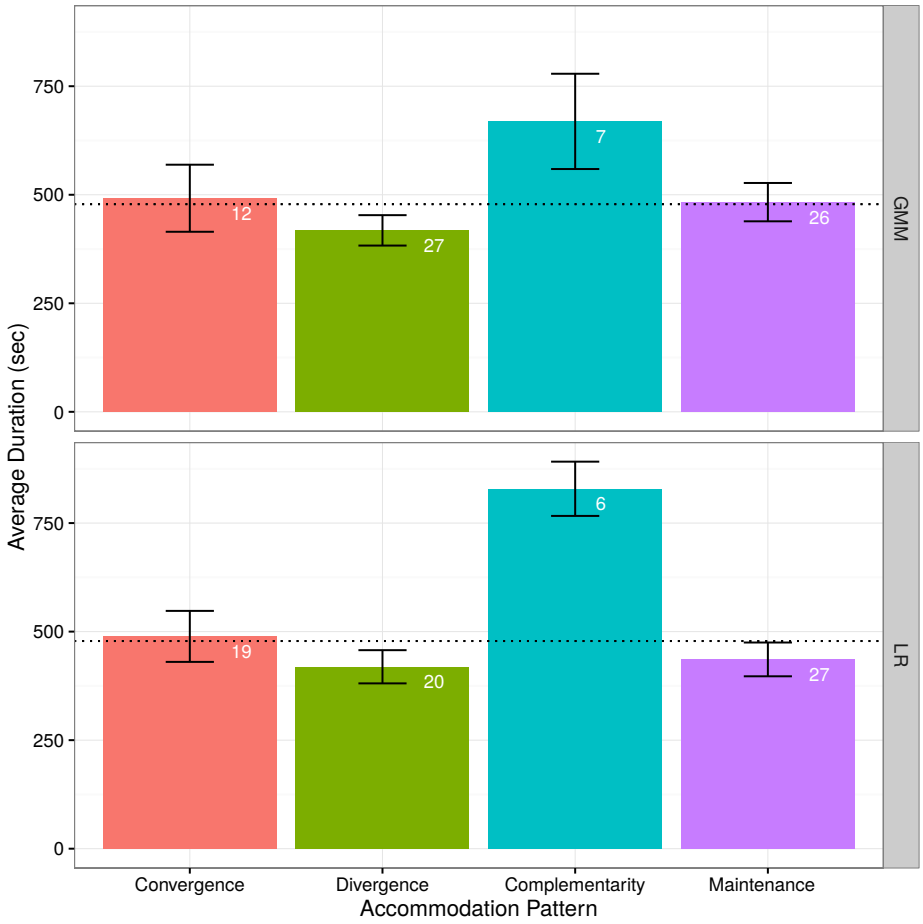



Figure 4.10: This graph presents the classification of EEG data from the partners' EEG data samples in relation to the duration of the interactions. Error bars represent the standard error of the mean. Numbers in each bar indicates the number of observations for that adaptation pattern. Dotted black line indicates the mean duration of all interactions. GMM = Gaussian mixture model (single state HMM), LR = left-right HMM.

| Model Type | Comparison | $p$ |
|---|---|---|
| GMM | convergence:divergence | 0.319 |
| | convergence:maintenance | 0.913 |
| | divergence:maintenance | 0.253 |
| Left-Right | convergence:divergence | 0.318 |
| | convergence:maintenance | 0.441 |
| | divergence:maintenance | 0.753 |

Table 4.11: This table presents the t-test results for comparisons between accommodation category classification by the different HMM types and interaction length for the partners' EEG data samples.

None of the differences between conditions tested return any significant values. The trend for the convergence-divergence comparison having the closest to significant value that is seen in the previous analyses is only seen here in the left-right

models. For the GMMs the comparison between divergence and maintenance returns the closest to significant value.

Overall, it would seem that there is only a weak indication that the trends seen in the EEG data associated with the speech of the participant's partner demonstrate similar trends to that observed in the speech data.

#### EEG Data Results: Both

Here, the data from both the EEG signal associated with the speech produced by the participant and the EEG signal associated with the speech produced by the participant's partner is considered together. Combining these two sources of EEG signal data allows for a greater amount of the EEG signal to be utilised. Using the data associated with the speech of just one of the speakers excludes a large amount of the available data.

The key questions to be addressed here are:

- Whether the HMM-based approach can detect patterns of adaptation in the segments of the EEG signal from a single participant that are associated with the words uttered by both speakers.
- Whether adaptation patterns found for the EEG signal follow the trends of accommodation found in the speech data.

To asses whether the HMM-based approach can detect real trends in the data and that results are not the result of chance, a binomial test was conducted on the total count for the number of statistically significant correlations. Results for this test are presented in Table 4.12.

| Model Type | No. Significant Cases | $p$ |
|---|---|---|
| GMM | 66 | $< 0.001$ |
| Left-Right | 66 | $< 0.001$ |

Table 4.12: Results of binomial test for both the speakers' own and the partner's EEG data samples.

The number of times at least one of the two correlations $L(A,B)$ and $L(B,A)$ is statistically significant with confidence level 0.05 (after Bonferroni correction) is 66 for both the GMMs and the left-right models. The probability of achieving such a result by chance is lower than $10^{-7}$ in both cases. Given the low probability that this results was the result of chance, it can be said that the trends detected by the HMM-based approach represent real events in the data.

To investigate the types of trend found in the data and to assess whether they pattern with that of the speech data, the counts for each of the accommodation pattern conditions and the average durations associated with these conditions must be considered. This data is presented in Table 4.13.

| | Model Type | Convergence | Divergence | Maintenance | Complementarity |
|---|---|---|---|---|---|
| Count | GMM | 15 | 29 | 22 | 6 |
| | Left-Right | 15 | 27 | 22 | 8 |
| Avg. Duration (s) | GMM | $545 \pm 71$ | $427 \pm 39$ | $484 \pm 44$ | $541 \pm 119$ |
| | Left-Right | $487 \pm 61$ | $455 \pm 45$ | $475 \pm 47$ | $551 \pm 94$ |

Table 4.13: This table reports the counts and average duration ($\pm$ the standard error) of both the speakers' own and the partner's EEG data samples providing a task classification of either Convergence, Divergence, Maintenance or Complementarity. Values have been rounded to the nearest second.

The average duration of convergence cases remains higher than the average durations of the divergence and maintenance cases for both model types. However, the total number of tasks classified as having convergence is somewhat smaller than seen in previous analyses. It is also smaller than the total number of tasks classified as divergence or maintenance.

The number of tasks classified as divergence on the other hand, has increased in comparison to previous analyses. It also remains consistently higher than the counts for both convergence and maintenance.

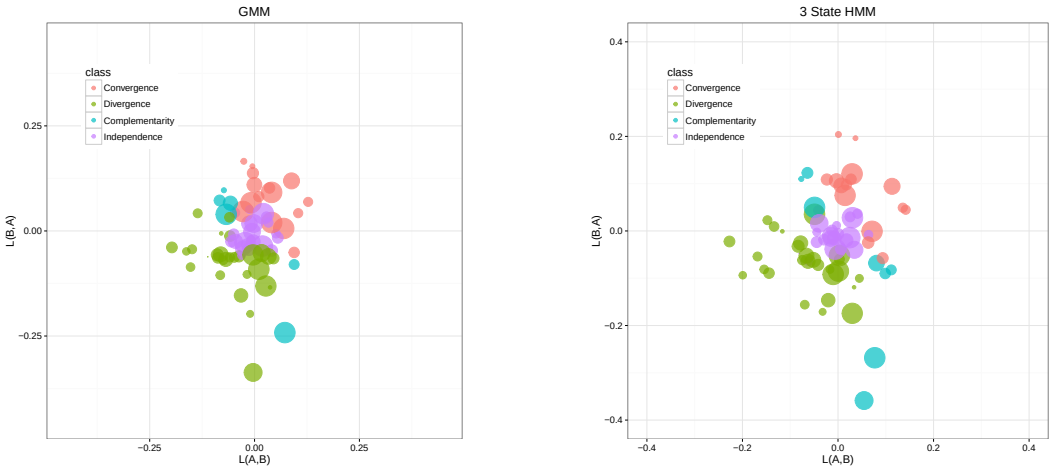
Although complementarity remains the accommodation pattern condition with the fewest number of cases, it also demonstrates the only instance with a higher average duration than convergence. Complementarity shows the closest average durations to that expressed by convergence for both model types. Again though, it is difficult to know how reliable this trend is given the small counts.

The average duration associated with the cases of maintenance tend to pattern more closely to the average durations associated with convergence than that of divergence. This trend is in keeping with previous analyses.

The individual results associated with these averages can be seen in Figure 4.11, where they are presented as a bubble plot. Each bubble represents a DiapixUK interaction and the size of the bubble is proportional to the length of that given interaction. Larger bubbles indicate longer tasks.

Overall, these plots look like they confirm that the tasks associated with convergence tend to have longer durations than those that are associated with divergence. The difference between the lengths of the tasks associated with convergence and the length of the tasks associated with maintenance appears to be less clear.

The spread of the data is greater for the left-right models presented in Figure 4.11b. Aside from the clear examples of complementarity towards the negative end of the y-axis, there is also a tendency for greater dispersion in the group of convergence cases in the upper right quadrant. In addition, the divergence cases in the lower left quadrant appear to have more of a tendency towards the negative end of the x-axis. One interpretation of these differences between the GMMs and the left-right models could be that there are greater changes in the EEG signal to be detected when the temporal domain is accounted for. Looking more closely at




(a) eeg scatter GMM (b) eeg scatter LR

Figure 4.11: Bubble plots for both the speakers' own and the partners' EEG data samples associated with the word uttered by the speaker. Each bubble represents a single DiapixUK task. The size of the bubbles is related to the length of the task they represent. Figure 4.11a presents the data for the GMMs and Figure 4.11b presents the data for the left-right models.

Figure 4.11a, it would appear that the divergence values tend to pattern close to the zero values of each axis. Given that the more extreme values for divergence tend to track along one dimension of the plotting space (either along the x-axis or the y-axis) rather than in both dimensions at once (moving diagonally away from the centre) may suggest that the patterning of one participant may be dominant in the GMMs. However, this is a very tentative interpretation and would need additional verification.

Figure 4.12 is a graphical representation of the average durations for each accommodation pattern condition for both model types.

For the GMMs, convergence can be seen to occur in tasks with longer average durations whilst divergence has shorter average durations. The same cannot really be said for the left-right models. Although the mean for the average durations of the convergence accommodation pattern condition does still provide a greater value than the mean for divergence, the standard errors for the conditions overlap considerably. They are unlikely to show a true difference between the conditions.

The maintenance condition appears to demonstrate a degree of separation from the divergence case in the GMMs but not in the left-right models. The average durations for the maintenance condition are consistent across both model types. the same can be said for the complementarity condition, although the number of cases is again rather small.

In sum, the only comparison between conditions that is likely to prove to be a real difference is the difference between convergence and divergence for the GMMs. The results of t-tests for comparisons between the conditions can be found in Table 4.14.

The results of these t-tests demonstrate no significant differences between the convergence, divergence and maintenance conditions. The trends seen in the order

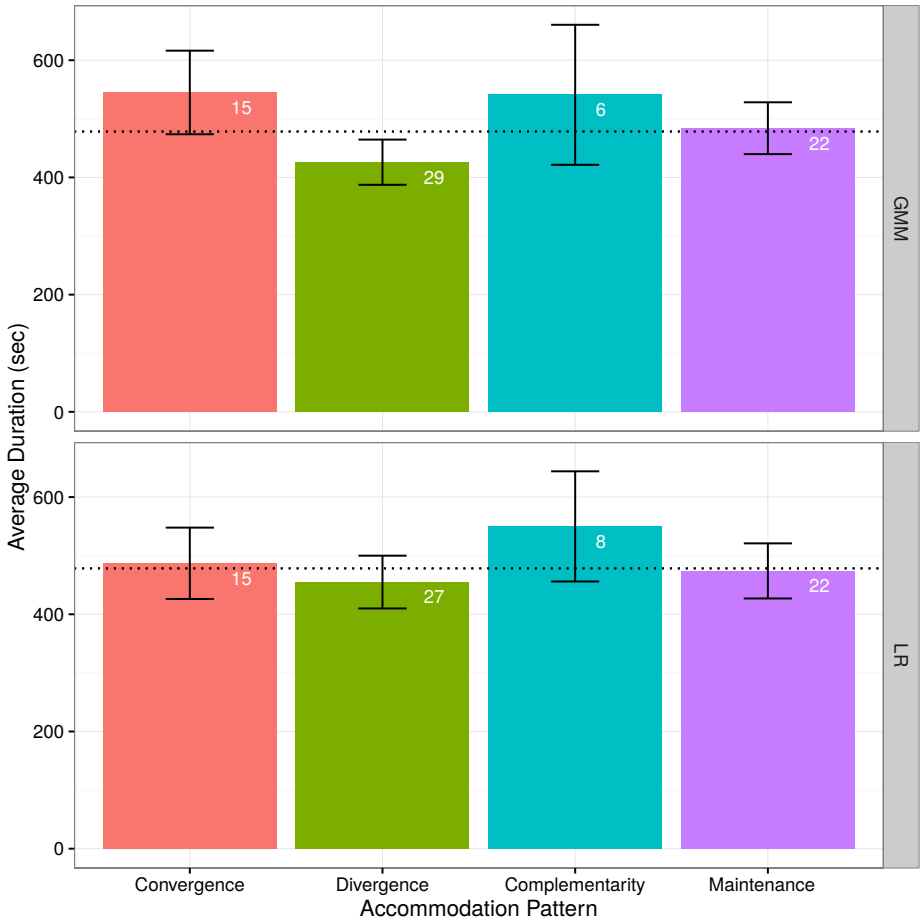



Figure 4.12: This graph presents the classification of EEG data from both the speakers' own and the partners' EEG data samples in relation to the duration of the interactions. Error bars represent the standard error of the mean. Numbers in each bar indicates the number of observations for that adaptation pattern. Dotted black line indicates the mean duration of all interactions. GMM = Gaussian mixture model (single state HMM), LR = left-right HMM.

of most to least significant do not pattern like the other analyses. Whilst the GMMs do show the difference between convergence and divergence to be the closest to significance, the left-right models do not. For the left-right models,the comparison that is closest to significance is between divergence and maintenance. This is interesting because it suggests that, for the left-right models the difference between the duration of the tasks where no accommodation is detected and the duration of tasks where movement away from one's partner is detected is greater than the difference between detection of the two opposing accommodation trends (convergence and divergence). This is further compounded by the high p-value found for the difference between convergence and maintenance, suggesting no difference in the durations of tasks demonstrating convergence and no accommodative behaviour. So, given the EEG data associated with the words uttered by both the speaker and their partner, the temporal patterning in the EEG feature space suggests no difference between the length of interactions for each accommodation classification and offers no suggestion of a trend.

| Model Type | Comparison | $p$ |
|---|---|---|
| GMM | convergence:divergence | 0.117 |
| | convergence:maintenance | 0.449 |
| | divergence:maintenance | 0.331 |
| Left-Right | convergence:divergence | 0.674 |
| | convergence:maintenance | 0.868 |
| | divergence:maintenance | 0.437 |

Table 4.14: This table presents the t-test results for comparisons between accommodation category classification by the different HMM types and interaction length for both the speakers' own and the partners' EEG data samples.

Having said that, the results of the GMMs show different patterns. Here, the difference between the durations of the tasks associated with convergence and those associated with divergence is the closest to significance. This suggests that the task durations of cases classed as convergence, based on the general distribution of evidence in the EEG feature space, are more likely to be longer than those observed for divergence although not to a statistically significant extent.

The lack of statistical significance for the comparisons drawn is not surprising but the results nevertheless suggest potential avenues for further investigation.

## 4.4 Discussion

At the outset of this chapter, the key findings and conclusions from chapter 3 were recounted and a series of three main aims for this chapter were presented. These key aims were:

1. To replicate the findings of the HMM based approach in the behavioural experiment.

2. To determine if an HMM based approach can detect shifting trends in brain activity patterns relative to an interlocutor.

3. To determine if there is a relationship between accommodation patterns and brain activity patterns.

The outcome of each of these aims will be discussed here. Interpretations of the relationship between the results presented in chapter 3 and the results presented in this chapter will not be discussed, except for when answering the first main aim of this chapter (subsection 4.4.1). Comparison between studies and consideration of the thesis as a whole is the concern of the following chapter, chapter 5. Generally speaking, subsection 4.4.1 addresses the first main aim of this chapter, subsection 4.4.2 addresses the second main aim of this chapter and subsection 4.4.3 deals with the third main aim for this chapter. Each remaining subsection can be thought of as

performing the following tasks, subsection 4.4.1 interprets the findings of the results presented in subsection 4.3.1, subsection 4.4.2 interprets the findings of the results presented in subsection 4.3.2 and finally, subsection 4.4.3 will tie together the findings from both the speech and EEG analyses.

### 4.4.1 Speech Data

The main aim that the series of results relating to this speech data pertain to is to determine if the results of the behavioural experiment had been replicated. Broadly speaking, it can be said that the results presented for this experiment do replicate those found in the behavioural experiment. The detection of convergence still remains at a consistent rate across all levels of representation offered by the HMMs, this is again true when compared against the numbers of interactions classified as divergence, which decease as the HMMs become more specific. Whilst the levels of significance for each of the comparisons may not present as strongly as seen in the behavioural experiment, they still tend to demonstrate similar trends.

The lesser degree of significance in these result could be due to a number of factors. The first of these concerns the participants themselves. Given that accommodation is thought to be affected by subtle social cues, the results for this experiment may have been impacted by individual differences in the participant groups. One such factor could be the age differences between participants within a pair. For the behavioural experiment, the mean age difference between pairs was $9.17$ years with the minimum age difference being $1$ year and the maximum age difference being $29$ years. In this experiment, the values associated with age difference were higher. The mean age difference between pairs was $13.00$ years with the minimum age difference being $4$ years and the maximum age difference being $33$ years. Whilst accommodation in general may be unlikely to be impacted by age, age differences between participants engaged in a collaborative task may have an impact. For instance, it could be the case that whilst participants accommodate generally, there may be particular words or phrases used by an older speaker that a younger speaker may not use at all. This may lead to a hyperarticulation of a new word by the younger speaker which is neither converging or diverging from the form used by the older speaker but is rather being used as calibration. There may be some scope here to investigate the role that diachronic language change has on accommodation and how this interaction might fit into synchronic language change models.

The main difference here between the behavioural experiment and this experiment, could be considered to be the application of EEG caps to the participants. It may be the case that the application of these caps had some form of effect on the way in which the participants chose to speak. Perhaps, the application of the EEG caps made the experiment feel more clinical and therefore inadvertently encouraged the participants to use more prestige forms of speech throughout the experiment. The

process of application and adjustment of the caps may have also had an impact on the speech of the participants. This is because there was necessarily more contact with the experimenters, questions needed to be asked regarding the health of the participants, regular communication was required during cap application to ensure a safe and appropriate fit and there were many more checks on the participants during the experiment in order to ensure a clean EEG signal throughout the experiment and to assess the welfare of the participants. All of this additional contact with the experimenters may have introduced a confound that meant participants partly accommodated towards the experimenters as well as towards one another. This confound was, however, unavoidable since the necessity for a clean EEG signal and for ensuring participant welfare were paramount.

Finally, with regards to the results for the word-dependent HMMs, the lack of significance for comparisons with divergence may be related to the low classification numbers for that accommodation pattern. The loss in classification counts for divergence as the HMMs become more targeted was also a feature of the HMM results for the behavioural experiment and it is heartening to see this trend carried through to this experiment. However, in this instance the number of divergence cases drops to a level that is comparable to that found for patterns of complementarity, which consistently attracts low classification counts. This remarkably low count for divergence may be having an impact on the power of the t-tests used to compare classification groups.

In general, even when accounting for the possible impact of confounding factors, the results can be said to mirror those of the behavioural experiment. As was mentioned in the discussion for the HMM results of the behavioural experiment, the approach presented here is a first pass attempt at applying a machine learning approach to the detection of accommodation and there is still much room for improvement of the approach. Indeed, further testing and expansion of the approach used here is suggested and encouraged. However, it is promising that even with the use of a relatively crudely implemented approach, the results between experiments demonstrate a good degree of reproducibility. This addresses the first main aim of this chapter and concludes that this experiment has been able to replicate the findings of the HMM based approach in the behavioural experiment.

### 4.4.2 EEG Data

This subsection looks to address the second aim of this chapter, namely to determine if an HMM based approach can detect shifting trends in brain activity patterns relative to an interlocutor. Across all of the three different EEG analyses performed, the data tend to track with the findings of the speech data, especially for the speakers' own EEG data. Although there were no significant results, a number were approaching significance. However, this is not to say that this approach should

be abandoned, there are a number of considerations that could lead to improved performance. These considerations as well as some possible interpretations of the findings themselves will be presented here.

The possible adjustments that might lead to improved performance will first be discussed so that the interpretations of findings can project some suggestions for future implementations. As with the development of any analysis tool, initial attempts may not prove to be overly fruitful but with adaptations and developments, future versions can be very powerful. Indeed, the development of speech recognition tools has taken a considerable amount of time and has seen many different iterations (Juang & Rabiner, 2005). Much in the same vein as the development of speech recognition tools, the development of tools to detect brain activity linked to ongoing behavioural phenomenon will require careful consideration and a good deal of time. The HMM based approach presented here represents only a first pass attempt at integrating EEG measurements with ongoing speech data. It is also a comparatively crude method of signal representation given the currently available computational methods that are being applied in the field of BCI (eg. Müller et al., 2008). As such, there are a number of technical consideration that could be applied to improve the ability of this HMM based approach to classify brain activity in relation to the speech signal.

### Vectorisation parameters

The choice of power spectral density (PSD) as the vectorisation parameter in the EEG analysis was made based on the the principle that initial trials with the application of this HMM based approach should kept as simple as possible. The PSD offers a snapshot of the brain activity in the frequency domain, which is in keeping with the findings regarding neural entrainment to speech (Peelle & Davis, 2012; Giraud & Poeppel, 2012). So, the GMMs would characterise the general characteristics of the PSD feature space for a speaker irrespective of time whilst the left-right HMMs would provide an equivalent PSD feature space representation with a time dimension. This restriction was due to the down sampling of the EEG signal in the pre-processing stage. Although the original EEG data was recorded at a higher sample rate, the signal was downsampled to 512 Hz in order to speed up processing time and because the Nyquist frequency associated with this sampling rate was high enough to capture all known frequencies of brain activity. In hindsight, retaining a higher sample rate may have allowed for extension of the PSD vectorisation technique to include word-dependent HMMs. The reason why this was not possible is covered in subsection 4.2.6 and relates to the lack of samples to calculate the PSD when drawing from data sampled at 512 Hz. Retaining a higher sample rate for pre-processing may have allowed for inclusion of the word-dependent models. However, given that there were no significant results at the broader GMM and left-right levels, it is not clear if further specification in the measure would aid or hinder detection. A

potentially more promising approach may be to change the vectorisation parameter from PSD to some other format.

As discussed in subsection 4.2.6 there are a number of other possible techniques for the vectorisation of EEG signals and the reader was pointed to the work of Gross (2014) for a review of such techniques. There are a wide number of potential techniques that might be more appropriate for this sort of analysis which draw on many different forms of transform based on signal filters such at the Fourier transform, the Hilbert transform and transforms based on wavelet analysis. Wavelet analysis based approaches, specifically those using a Morlet wavelet, have been argued by Gross (2014) to provide 'an optimal trade-off between time and frequency resolution' (pp.66) and may be a potential option to provide a better characterisation of changes in the frequency domain of EEG signals. The application of different vectorisation parameters, including wavelet based analyses, to the data would be a reasonably straightforward task to perform and could be achieved by simply editing part of the signal processing pipeline. This would be roughly equatable to speech recognition technologies cycling through different forms of representation for the speech signal. In speech recognition, the use of MFCCs has been shown to be effective only through the testing of a variety of different vectorisation parameters, the same must be done in order to apply a process such as HMMs to EEG data.

#### HMM structures

Further to considerations regarding the vector parametrisations, the structure of the HMMs used must also be considered. In the same vein as the discussion regarding the appropriate number of states to use for speech recognition in subsection 3.6.3, this too must be considered for the application the HMMs to EEG data. For word-dependent HMMs, the states can broadly be considered to be representing the underlying phonemes of the word in question. Thus as the word-dependent HMMs used in this thesis have 3 states, they may under represent some phonemes in longer words and over represent some phonemes in shorter words. For the left-right models, the use of a 3 state HMM is more appropriate since ii doesn't distinguish between words, a 3 state model can be thought of as a reasonable compromise across the words used to train the model. Likewise, if the EEG signal is being chunked based on the segmentation of the speech signal into words, then the number of states that the EEG signal is being modelled with should represent some meaningful underlying content. Assuming a direct link between the EEG signal and perceived phonemes may be wishful thinking. So, trialling different numbers of states for the modelling of the signal should allow for an exploration of an appropriate coverage of the signal in relation to the behavioural samples that the EEG data are linked to. However, the consequence of shifting the number of states used for EEG analysis away from that used for speech analysis mean that the underlying features that the states tied to the EEG signal represent will need to be carefully considered.

Through a consideration of the likely features underlying the states used in HMMs for EEG data, it may be possible to improve the performance. In the work presented in this thesis, the observations for the HMMs are derived from the EEG signal whereas the states are dictated by the word segmentation of the speech signal. The way in which this was performed was by using the start times of words from the speech signal. What this did not account for was the lag between the production of the word by a speaker and the perception of the word in the partner. This means that the brain activity associated with that particular chunk of speech may not represent the processing or interpretation of the word but rather some feature present in the brain prior to processing. Accounting for this lag between the production and perception of speech may be able to be resolved through the application of a constant lag value but a dynamic solution would probably be more accurate. A system for assessing sections of brain activity that are maximally predicted by the available speech signal would be a sensible avenue to explore. This would require the integration of both speech vector forms and EEG vector forms into an HMM in order to estimate the joint probabilities of the two vector forms occurring together. This option was explored in the creation of the method employed but is not presented in this thesis because it proved to not be possible to implement in HTK (Young & Young, 1993). If this is something that could reasonably be taken forward, other tools for HMM creation and maximisation would need to be explored. Some possible starting points might include the HiddenMarkov package (Harte, 2016) in R (R Core Team, 2016) and the Markov package (Bleackley, 2016) for Python (Python Software Foundation, 2016).

### Signal cleaning and preprocessing

In addition to considerations regarding the implementation of HMMs on EEG data, it is also important to discuss the signal cleaning during pre-processing. For the analysis performed on the speaker's own data, there clearly is the possibility that the HMMs are detecting both the information pertaining to the word in question produced by the brain but also that produced by the muscular activity of the speech articulators. Whilst, considerable efforts were taken to remove speech artefacts, it is not possible to say that they were completely removed. For the comparison with the speech of the partner there is less chance of contamination from the speech articulators, although there may still be some small artefacts from the eye movements. The artefacts in the data for the analysis of all available segments will naturally fall somewhere between these two analyses. The approaches used to remove any non-neurophysiological information from the EEG data stream was based on well documented and widely used EEG preprocessing tools (Delorme et al., 2011; Bigdely-Shamlo et al., 2015). It is not the intention of this thesis to comment on or critique these tools but rather to suggest that the way they were implemented may have been sub-optimal. Although every care was taken to follow the recom-

mendations for implementation of the EEG data cleaning tools, much like the many considerations for the implementation of HMMs, there are a number of parameters that are able to be adapted to suit specific purposes. In this case, a number of choices were made regarding data cleaning including the removal of certain frequencies in the EEG data and the reduction of contamination in the data through evaluation of ICA topographies, time-courses and amplitude spectra. It may be the case that the cleaning that was performed on the data may have been too stringent or perhaps too lenient. As with the use of vectorisation parameters, some calibration of the use of these tools may provide more reliable and clear results.

**Final considerations**

With the proviso that the suggestions above might allow for some improvement of the methods employed in this thesis, the results of the experiments can now be considered. There were no significant effects found between brain activity and the accommodation pattern classifications. However, given the results of the binomial test, it can be concluded that the results were tracking real trends in the data and were not being produced by chance. Taken broadly across all of the EEG analyses, the most straightforward interpretation of the findings would be that there is no relationship between the length of interactions and the EEG data accommodation pattern classifications. Alternatively, a more nuanced interpretation can be found in the tendencies towards significance found in the results. For instance, the difference between convergence and divergence classifications in relation to interaction length for the EEG data relating to the speech produced by the speaker was approaching significance at $p = 0.068$. Admittedly, this is the closest that any of the comparisons comes to significance and it is based on the general distribution of evidence in the feature space with no time dimension but the fact that it is approaching significance in light of the potential hurdles, as outlined earlier in this subsection, suggests that significance may be achieved through some adaptation of the approach.

It is interesting that the results from the data drawn from the speech produced by the speaker presents the strongest case for possible detection of similar activity patterns in the brains of two speakers. What this could potentially suggest is that there may have been an influence of the muscular activity from speaking on the data used in this analysis. This is possible because the electrical activity related to the muscles controlling the speech articulator movements is likely to be greater than that detected from the brain. This is something that could be remedied by adjusting the approach to include a lag to account for processing or by improving the data cleaning process (both mentioned above). However, if it is assumed that the speech artefact removal was successful, it could also be possible that the EEG patterns for the participants are tending towards each other when producing words. This could suggest that there is some adaptation in the manner in which the brain is producing instructions for production, at least within the interactions that took

longer to complete. It might suggest that the ongoing activity of the brain activity produced by the participants became more similar in the general distribution of the feature space in interactions that the participants found harder. This would be in keeping with the findings of the speech data and would also provide support for theories that suggest an alignment between mental states when engaging in a joint process. However, whilst this kind of interpretation of the data would be possible if the results were significant, they cannot be concluded given the findings of the work presented here.

Again, although the findings are not significant, it is interesting to look at the significance values across each of the EEG analyses. For convenience, the values have been placed side-by-side in table 4.15. When comparing the significances of the GMM and left-right HMMs within each comparison for each type of EEG analysis, six of the nine comparisons see the significance value increase as the HMM moves from GMM to left-right. This could indicate that the left-right model is less able to characterise the EEG signal in terms that relate to the adaptations of the speech signal or it could indicate that the left-right models are classifying based on some other behavioural aspect that has not been tested for here. Given that the left-right models provide some interpretation of the time dimension in their representations, it is somewhat difficult to fully interpret the findings given the issue that was raised above regarding the lag between production and processing that was not accounted for. This lag could very well be playing a role in the comparatively poor relationship between interaction length and the left-right accommodation type classifications. Further to this consideration, it is interesting to see that the differences between GMM and left-right significances move in the same direction for both convergence based comparisons but hold a different pattern for the divergence-maintenance comparison. Differences in significance for the convergence comparisons tend to increase for both the EEG self and EEG full analyses but decrease for the EEG partner analyses. This may suggest some slight improved ability of the left-right models performance when classifying brain activity based on the partner's utterances. This is an extremely tentative interpretation (especially considering the $0.001$ difference between the GMM and left-right values in EEG Partner) and at this point should not be taken as much more than speculation but it might suggest that during the processing of a partner's speech, a deeper level of processing occurs when in a challenging situation. The states that the left-right model is presumably basing its predictions on are more fine-grained than those used by the GMMs and the fact that this decrease in significance values only occurs in convergence cases may point to a possible effect. However, as mentioned, these are merely interpretations of what may be detectable in the data given the proper adjustments being made to the HMM based tools used in this thesis. As such, these suggestions should only be taken as speculation that would require significant further investigation to validate.

The data presented demonstrate indication that an HMM based approach might

| | Model Type | Conv:Div | Conv:Main | Div:Main |
|---|---|---|---|---|
| EEG Self | GMM | 0.068 | 0.176 | 0.399 |
| | Left-Right | 0.169 | 0.545 | 0.261 |
| EEG Partner | GMM | 0.319 | 0.913 | 0.253 |
| | Left-Right | 0.318 | 0.441 | 0.753 |
| EEG Full | GMM | 0.117 | 0.449 | 0.331 |
| | Left-Right | 0.674 | 0.868 | 0.437 |

Table 4.15: This table reports the significance values for each of the EEG analyses in order to aid comparison. All values presented are the $p$-values associated with the t-tests reported in subsection 4.3.2. Conv = convergence, Div = divergence, Main = maintenance and the colon marks that a comparison was made between the interaction lengths associated with these accommodation type classifications.

be appropriate for EEG analysis in a context such as the one in this thesis. However, the approach will need some fine tuning before complete validation can be offered. The use of HMMs and MFCCs took some time to be reliably applied to speech recognition. The EEG signal can be considered to be even more complex and diluted than the speech signal. The approach presented here is a first pass attempt to characterise the EEG signal in relation to the behavioural phenomenon of speech with an early version of a sophisticated tool. These results suggest more refinement and calibration is needed to extend the HMM based approach to the analysis of EEG signals.

### 4.4.3 General

The last main aim of this chapter asks if it is possible to detect a relationship between accommodation patterns and brain activity patterns. For the neural experiment presented in this thesis, the answer to this is that it has not been possible to detect a relationship. If a relationship had been present, it would have been expected that the patterns observed in the speech data would also have been seen in the EEG data. Whilst it is possible to theorise that, given some modifications to the approach taken, some of the trends shown in the EEG data may have been significant, verification of these theories would require further investigation. Nevertheless, if these theorisations can inform future research directions, then they should be discussed. This is the purpose of this subsection, to provide some context for the outcome of the EEG analyses in relation to the speech analyses and to suggest how to sensibly move forward with this approach.

Whilst this thesis has mainly focused on measuring and interpreting brain activity data in the frequency domain, it is worth noting that the analyses were not restricted to particular brain regions. Part of the reasoning for this was the poor spatial resolution that would have been achieved using a 32-channel EEG system but an

additional reason was to keep the analysis relatively simple. Given the number of pre-processing and analysis techniques employed throughout the work of this thesis, additional constraints on the data were avoided. However, given the results that were obtained, it may be sensible to consider a restriction of analyses to particular brain regions. By not isolating to a region that is known to engage in a particular process, this analysis of brain activity is also picking up on information pertaining to other ongoing activity in the brain such as the processing incoming visual and tactile information, ongoing emotional state tracking and general brain activity related to bodily functioning. If the analyses presented in this thesis could be restricted to targeted areas of the brain, it could further improve performance whilst also allowing for some more specific interpretations regarding the specific use of oscillatory activity for ongoing speech tracking to be evaluated. Much of the research surrounding the relationship between oscillatory activity in the brain and the speech signal has focused on the primary auditory region of the brain (eg. Giraud et al., 2007; Giraud & Poeppel, 2012; Peelle et al., 2013; O'connell et al., 2015). Restricting analyses to an area such as the primary auditory cortex may prove to be a sensible direction for future work but this comes with an additional issue. The primary auditory cortices are located bilaterally on either side of the head, this places them close to some of the larger muscles controlling jaw movement. By restricting analyses to these areas, it may prove to decrease the signal to noise ratio, making it harder to remove potential sources of muscular artefacts. Some adaptation of the experimental protocol may be able to work around this issue but this would not help to advance the tools being developed to allow for collection of EEG data during speech production.

This experiment has explored the link between brain activity and speech through comparison of analyses from two different signal sources in relation to a behavioural outcome (interaction length). The interaction length of tasks signifies when participants found a task more difficult to complete, however it does also introduce a potential confound of different sample numbers for the analyses (ie. longer tasks will likely provide more potential samples). Whilst this is mitigated to a good degree through HMM implementation, the development or use of a collaborative task that meets the same requirements as the DiapixUK task but that also holds other elements of the experiment constant when assessing task difficulty would prove to be a valuable addition to this research. Such an experiment would allow for verification of these findings in direct relation to an explicit measure of difficulty. This would allow for a clearer judgement to be made regarding both the use of accommodation as a tool to aid in clarification of a difficult joint task/concept and for the relationship between accommodation and brain activity. Indeed, coupling an experiment such as the one suggested with the inclusion of an HMM based approach that drew direct links between speech and brain activity would provide even more insight.

An additional topic to consider is the types of analyses that were undertaken in this thesis. Three different analyses were presented, one which used EEG data

associated with a participant's own speech, one which used EEG data associated with the partner's speech and one which combined these two. They can broadly be considered as attempts to determine the degree to which brain activity relating to one's own utterances changes, the degree to which brain activity relating to a partner's utterances changes and the degree to which brain activity relating to all speech during an interaction changes. Whilst each of these analyses were aimed at assessing a particular type of trend in the neural signal, in hindsight, some additional analyses might also have proved useful. One such analysis would have been to use EEG data from the speaker and participant during the same period of time, for the same word. Modelling data from the same time would have allowed for an evaluation of the degree of similarity in the neural signals during production and perception of the same word. This is something that cannot be done with the speech signal since participants are not always speaking. However, participants are always thinking and this information could prove to shed some additional light on the relationship between speech and brain activity.

It is difficult to make concrete statements about the relationship between speech and neural signals with the data presented in this thesis. Having said that, although further investigation is needed, there are some promising routes of exploration that may prove to be fruitful in the future. It can be concluded that whilst the exact process of investigation used in this thesis may not have returned statistically significant results, there are trends that suggest a machine learning approach will lead to valuable findings in the future.

# Chapter 5

# Discussion

As a whole, this thesis covers a broad range of topics, from the roots of CAT (Giles et al., 1973; Giles et al., 1991b) through experimental approaches for detecting accommodation (Shockley et al., 2004; Babel, 2009b) and the cognitive theories underpinning it (Pardo et al., 2016a) to the possible role of neural entrainment in speech processing (Giraud & Poeppel, 2012; Peelle & Davis, 2012) and the role of machine learning in resolving the technicalities of capturing these phenomenon (Rabiner & Juang, 1986; Dumas et al., 2011). As such, many sections have their own dedicated discussions. This discussion section will return to the overall goals and aims of the thesis. This final chapter to the thesis aims to draw together considerations from across the thesis and explore the general findings as well as providing an indication of the wider implications that this thesis has for future work in the field.

Before entering into the main discussion, it is worth providing a summary of the main aims and findings of the work presented.

**Behavioural experiment:**

- Main aims
  - To determine if a segmental acoustic-phonetic analysis approach can be used to detect accommodation across a continuous interaction
  - To determine if a holistic analysis approach can be used to detect accommodation in a continuous interaction
- Key findings
  - Detection of accommodation during short-term, continuous interactions using phonetic measures returns a few effects, specifically linked to interaction length.
  - Holistic approaches to detection of accommodation during short-term, continuous interactions are able to classify interactions by accommodation pattern. They return results that show a significant relationship between local interactional contexts and convergence of speaker vocalisations.

- Conclusions
  - Accommodation occurs across a number of acoustic features. Assessing them individually does not account for any interactions between the features.
  - Considering accommodation on a continuous basis may allow for this gap in the assessment of this phenomenon to be filled.
  - An HMM (or machine learning) based approach presents a potential way to evaluate accommodation as a continuous holistic process.
  - Whilst time is undeniably a factor in accommodation, the impact of other behavioural factors may have been overlooked due to a lack of sensitivity in investigation methods.
  - There may be some suggestion that a shift towards investigating accommodation in relation to behavioural triggers will uncover more insight about the driving factors behind accommodation.

**Neural experiment:**

- Main aims
  - To replicate the findings of the HMM approach in the behavioural experiment.
  - To determine if an HMM-based approach can detect shifting trends in brain activity patterns relative to an interlocutor.
  - To determine if there is a relationship between accommodation patterns and brain activity patterns.
- Key findings
  - The HMM based approach was able to broadly replicate the findings of the behavioural experiment.
  - The data suggest that the HMM based approach was able to detect some trends in the EEG signal in brain activity pattern relative to the interlocutor.
  - There data indicate that the HMM based approach was detecting something related to accommodation since the EEG patterns tracked with that of the speech data. However, there were no statistically significant results to suggest a relationship between accommodation patterns and brain activity patterns.
- Conclusions
  - Further evidence for the efficacy of holistic approaches for the detection of speech accommodation in continuous live interactions is provided. The holistic, HMM based approach replicated the results from the behavioural experiment.

- HMM based approaches may still be able to detect shifts in brain activity in relation to accommodation but only after improvements have been trialled for the approach.

## 5.1 Accommodation in live, continuous interactions

The traditional methodologies of assessing accommodation in the speech signal have generally not considered accommodation in live, continuous speech from a holistic viewpoint (see section 2.2). Whilst traditional methods have proven to be excellent tools for the investigation of general trends in accommodation through perceptual studies and of general changes in specific acoustic-phonetic features, this type of approach does not capture the relational nature of an interaction between two speakers in a continuous and dynamic manner.

The ongoing relationship between speakers in relation to the context of the conversation is likely to have an important role in determining the level and type of accommodation that takes place during natural human communication. The findings presented in this thesis support this view, to a certain extent. The fact that convergence was classified as the most common accommodation pattern in interactions that took longer to complete, meaning that participants found them more difficult, suggests that alignment of speech patterns can be used as a mechanism to improve communication under difficult circumstances. The holistic, HMM based approach makes an assessment of the degree of accommodation that either speaker shows in relation to the other through an assessment of virtually every word uttered during an interaction. This allows for a judgement of accommodation within an interaction to be made based on ongoing trends across the interaction. It would have been difficult to find a result such as this using more traditional methods since the traditional approach to testing for accommodation would rely on sampling the early and late portions of each interaction and making a comparison between them. Not only would this reduce the amount of data being used to measure accommodation but it would also not be able to capture any trends that occur during the interaction itself. The continually unfolding relationship between the speakers in relation to the conversational context would not factor in to the measure of accommodation. At any given point during a spoken interaction a speaker could enter a particular internal state in relation to what the partner had said and this may lead to particular accommodative effects. For example, a speaker may express disagreement or frustration with what the partner had said, potentially leading to some divergent accommodative behaviour. On the other hand, a speaker may express agreement or happiness in relation to what the partner had said, potentially leading to some convergent accommodative behaviour. In traditional approaches these potential variations would not factor in to the overall classification of an interaction as convergence, divergence, maintenance or complementarity. Rather, the classification

would be the result of the average absolute difference between an early speech sample and the late speech sample. The fact that an HMM based approach deals with the data in a more continuous manner allows for the integration of these more nuanced accommodative behaviours into classification of interactions.

## 5.2 Efficacy of holistic approaches

As mentioned throughout the thesis, the application of HMMs in this work is a somewhat crude use of machine learning approaches such as HMMs. This was a deliberate choice in order to test the premise of applying machine learning to the problem of accommodation detection in both the speech and neural signals. The reasoning being that if it is possible to detect some effects using a crude implementation of a simple machine learning tool, then it would provide the basis for further development of a machine learning approach with more advanced techniques. This section of the discussion offers some considerations of the efficacy of an HMM based approach to accommodation detection based on the findings presented in this thesis.

### 5.2.1 Holistic approaches for speech signals

This thesis has presented a preliminary method for the detection of accommodation in a continuous interaction which utilised HMMs to model speaker adaptation. It can be concluded that the use of the HMM based method was successful in the classification of accommodation. However, as the presented method is only preliminary, there may yet still be some work that can be done to improve performance.

Something that was clear upon reviewing the current literature in accommodation detection was that there is a complex relationship between accommodation and the acoustic-phonetic features that are used by speakers to produce the phenomenon. Whilst the human perceptual system may be able to detect similarity between speaker voices, the process remains a 'black box'. The relationship between the acoustic-phonetic measures that are used to evaluate accommodation and the perceptual correlates of these measures is not clear. Indeed, studies such as Pardo et al. (2010) and Babel and Bulatov (2012) conclude that there is a non-superficial link between acoustic-phonetic measures and perceptual measures. What an HMM based approach offers is a step towards being able to interpret the speech signal in a way that is more akin to perceptual measures but that is also rooted in the spectral properties of the speech signal. It is not being suggested that HMM based approaches represent the way in which the human perceptual system interprets and adapts to speech. All that is being suggested is that HMM based approaches consider the speech signal in a more holistic and continuous way than traditional acoustic-phonetic methods. This move towards a more holistic and continuous method of measurement is suggested to be more akin to human speech signal processing than

the interpretation of single (or even multiple) acoustic-phonetic features.

An additional benefit of an HMM based approach to accommodation detection is that an assessment can be made in relation to the behaviour of both speakers. Accommodation is an adaptive phenomenon that is contingent on exposure to the speech of another speaker. All studies of accommodation rely on a speaker producing a speech response upon hearing the speech of another speaker (eg. Goldinger, 1998; Pardo, 2006; Babel, 2009b; Casasanto et al., 2010; Yu & Abrego-collier, 2011; Bailly & Martin, 2014; Pardo et al., 2016b). However, thus far, these studies do not asses the ongoing speech adaptations that are produced by both speakers throughout an interaction. They instead generally rely on comparisons between pre and post exposure samples. By modelling the general form of a speaker's speech characteristics and comparing every word against models for both speakers in an interaction, HMM based methods allow for not only a continuous assessment of a single speaker's accommodation but the relational accommodation between speakers.

Taken as a whole, HMM based approaches for accommodation detection and classification provide a potential avenue for the development of sophisticated tools for continuous, spectrally based measures of interactional speech behaviours. The use of HMMs is only a first pass attempt at applying machine learning techniques to improve the detection and classification of accommodation. There are many other machine learning approaches that may also be appropriate for this sort of problem and that would be well worth exploring. This is a major contribution of this thesis, it has been demonstrated that HMM or machine learning based approaches to the investigation of accommodation outperform even a statistically sophisticated phonetic analysis of accommodation. The pursuit of this approach is encouraged as it could prove to be a significant aid in understanding accommodation.

### 5.2.2 Holistic approaches for EEG signals

The use of HMM based approaches for the assessment of similarity between continuous EEG signals has not proven as fruitful as their use in the identification and classification of accommodation in the speech signal. However, there are signs that the approach could be useful if refined. The fact that the HMMs have been shown to not be classifying random fluctuations in the EEG signal suggests that they are making assessment of actual trends in the data. This is an indication that it may simply be the case that the approach just needs to be fine tuned to better represent the neural traits of the behavioural phenomenon of interest. What can be concluded from this is that whilst HMM based approaches may not be a 'one-size-fits-all' answer to the classification of continuous EEG signals, with some targeted adaptations it may prove to be useful in task specific contexts.

Concerning the use of machine learning approaches more generally in the interpretation of continuous EEG signals and EEG signals more broadly, there are

already a number of applications being explored (Müller et al., 2008; Shoeb & Guttag, 2010; Shi & Lu, 2013; Johannesen, Bi, Jiang, Kenney, & Chen, 2016). The potential applications for machine learning approaches in EEG and brain activity measuring/imaging in general are broad and span the whole pipeline of analysis form pre-processing to final analyses. The application of these tools is only likely to become more common as research in BCI becomes more advanced. What is presented in this thesis is only an initial suggestion of a potential route for exploration.

In terms of the possible implications for the use of holistic approaches on the EEG signal, there are a number of ways in which the work in this thesis could prove useful. Whilst continuous EEG measures are used in clinical settings to monitor patients for indications of seizure activity etc. (eg. Claassen, Mayer, Kowalski, Emerson, & Hirsch, 2004; Arndt et al., 2013) they are not as commonplace in research considering cognitive functioning where ERP based approaches are favoured (eg. Campanella et al., 2002; Davis et al., 2015). The work presented in this thesis presents a potential alternative that might allow for some relaxation of the experimental restrictions of ERP analyses. By using a HMM to characterise the general form of a participant's ongoing brain activity and then comparing specific segments of data to the established HMM (or even potentially a baseline HMM that is built from a large and robust EEG data corpus), it might be possible to move away from the need for multiple trials. Additionally, the fact that HMMs can be built based on a number of different vectorisation parameters, there is scope to develop parameters that focus on specific theorised features of the neural signal. The approach requires refinement but these are all feasible applications of this type of signal analysis on EEG data.

## 5.3 Accommodation and brain activity

At the beginning of this thesis it was stated that the main theoretical research question that was being asked was:

> Is speech accommodation linked to the alignment of mental representations as accounted for through observable brain activity?

In terms of the results presented in this thesis, it can be concluded that there is no evidence provided that supports a link between speech accommodation and brain activity. However, a more pragmatic answer to this question would be that the findings are inconclusive. The method presented here for the evaluation of EEG signals in relation to the speech signal is a preliminary design that will require a good degree of fine tuning before it can concluded that (a) the HMM based measure is measuring accommodation and (b) that the manner in which the HMM based measure is employed to evaluate EEG signals is optimal.

This thesis has attempted to take advantage of a machine learning approach to signal analysis in order to classify the subtle behavioural phenomenon of accommodation. It has then tried to apply this technique to explore the possible link between accommodation and brain activity as measured through a comparatively weak neurophysiological signal that is likely to contain a high degree of noise. Whilst the patterns observed in the results may not provide conclusive evidence of a relationship between accommodation and brain activity, it is promising that what is found in the neural analyses are not simply the classifications of noise. It is also positive that this classification has been possible through the use of simple HMMs as opposed to more complex machine learning approaches such as recurrent neural networks (Rajan, Abbott, & Sompolinsky, 2010). Having said this, this thesis cannot be said to have directly addressed the relationship between accommodation and joint brain activity since the relationship was inferred through a joint behavioural response (interaction time). The most appropriate way to expand on the work presented here in order to directly interpret the link between accommodation and joint brain activity would be to implement a series of HMMs that are defined over the joint probabilities of both speech and brain data. This is something that was mentioned in subsection 4.4.2 and which was not possible to address within this thesis due to the restraints of HTK (Young & Young, 1993).

Further to this, the thesis provides work furthering the consideration of brain activity interacting as part of a joint process. It expands on previous work (Babiloni et al., 2007; Lindenberger et al., 2009; Dumas et al., 2011) by making links between joint brain activity and speech. Previous works that have looked at joint brain activity in humans tend to focus on either joint action behaviours (Lindenberger et al., 2009; Dumas et al., 2010) or neuroeconomics (Montague et al., 2002; King-Casas et al., 2005). In general there isn't much work on joint brain activity during spoken interactions. As such, this thesis provides some considerations and contributions for advancing research into joint brain activity during speech. This is important because speech is one of the most complex behavioural processes that humans undertake and yet it comes naturally to most. Having the capacity to evaluate both the speech signal and brain activity in terms of dyadic interactions could provide valuable insights into the communicative process that analyses of single speakers could not provide.

Taking this work forward may provide more evidence to aid in the understanding of the dynamics of accommodation during an interactional engagement. Adaptation in relation to another speaker is naturally dependent on an number of social and contextual factors (Namy et al., 2002; Evans & Iverson, 2007; Purnell, 2009; Pardo et al., 2010). Perhaps an approach that utilises a HMM or machine learning approach to the continuous evaluation of accommodation might be able to shed some light on the relationship between some of these factors through tightly controlling social and contextual factors but allowing for speech to unfold more naturally. This

would allow for more ecological capture of speech forms whilst isolating some factors that are theorised to contribute to accommodation. This could then be taken further to investigate any potential relationships between theorised areas of brain activity related to specific cognitive processes during a spoken interaction. Work such as this could offer some key insights into both the continuous and adaptive nature of accommodation during short-term interactions as well as providing some concrete neurophysiological evidence for theories of cognitive systems involved in both accommodation and possibly speech processing more generally. However, this proposed work would be contingent on the development and fine tuning of the approach employed in this thesis.

## 5.4 Future research directions

The work presented in this thesis is far from conclusive and there are a wide range of potential routes for the advancement of the work presented here. It is the aim of this section to provide some suggestions for avenues of research that might prove fruitful in advancing both the theoretical and technical aspects of the work presented here.

A key outcome of what is presented in this thesis is that accommodation is reactive to the environment and local context in which it is being used. These contexts may be related to the emotional or internal state of the speaker. In order to further investigate this proposal, a series of experiments evaluating the degree to which speakers converge and/or diverge from a target partner when positive versus negative contexts arise would prove useful. However, the construction of such experiments would be somewhat difficult to construct since switching between emotional states regularly within one participant might lead to a levelling of the accommodative effect. Running a between speakers experiment where different groups are treated in either positive or negative conditions could be one way around this and may even serve to compound the effect. In this case though, a ceiling effect might be reached if the con/divergence does indeed compound or it might emerge that excessive con/divergence leads to acclimatisation effect leading to a return to maintenance. Additionally, for experiments such as these, to maintain an interactive setting, a confederate would have to be used. This would not be ideal as the confederate may unknowingly introduce some variation into their speech patterning that influences results. Perhaps an virtual reality approach such as the one taken by Casasanto et al. (2010) might be appropriate in this situation to eliminate experimenter bias. Whatever the structure of the experiment, the findings would provide valuable information on the possible role that context and emotional state play in the production of accommodative behaviours. They could also go some way to making suggestions about the role of accommodation in language change more generally. If it could be demonstrated that positive affect was associated with greater convergence, it could be argued that propagation of certain phonetic variants are linked

to positive reinforcement through pleasurable joint activities.

Although the core of this thesis focuses on holistic, HMM based approaches to the detection and classification of accommodation, it would also be sensible to explore GAM based approaches. Some accommodative effects were visible through the use of GAM based approaches. It would be wise to pursue this method of analysis both as a method for detecting accommodation in its own right and as a complementary measure to any future holistic, HMM (or machine learning) based approaches.

Finally, one potential application of a holistic, HMM (or machine learning) based approach could be the construction of a 'real time' accommodation tool. Admittedly, this might stretch the technology a little but it is theoretically possible. Given a period of calibration to individual speakers or possibly calibration based on an existing corpus of speech HMMs could be used on-the-fly to determine likelihood ratios. Automatic speech recognition technologies already provide functions for real time speech recognition (eg. Apple's Siri, Microsoft's Cortana, the Amazon Echo) applying this to accommodation would involve recognising speaker similarity rather than individual words or sentences. It may take some inventive thinking to apply a real time display of accommodation into a research setting but the resulting likelihood ratios could certainly be used to produce evaluations such as those provided in this thesis, albeit after interaction rather than in real-time.

## 5.5 General conclusions

The work presented in this thesis has aimed to investigate the relationship between accommodation and brain activity during a live interaction. In order to do this, it has presented a novel holistic acoustic-phonetic measurement approach based around HMMs. Findings suggest that the approach suggested is able to detect accommodation in the speech signal and could have possible applications for use in the analysis of brain activity. However, the thesis acknowledges the limitations of the approach presented here and suggests that in order for the approach to be more widely applied, further refinement is required. Suggestions are made for ways in which the approach could be improved and for directions of future research.

Results concerning the detection of accommodation in speech provide support for theories that suggest accommodation, as a phenomenon occurs across multiple phonetic features at once. Standard analyses of individual phonetic features are likely to lack the power necessary to capture accommodation during a live interaction. A machine learning or HMM based approach is able to contribute to accommodation detection since it considers the speech signal as continuous and constrained by time. It is suggested that an HMM or machine learning based approach to the investigation of accommodation could prove to be a major contribution to the field. The analyses in this thesis demonstrate that a rudimentary HMM based approach is able to outperform a statistically sophisticated phonetic analysis in the detec-

tion of accommodation whilst also being able to categorise accommodation more accurately in terms of its potential adaptive directions (convergence, divergence, complimentarity and maintenance). This insight into the measurement of accommodation through the use of an HMM based approach is a major contribution of this thesis.

Results concerning the detection of a link between accommodation and joint brain activity provided limited support. However, a number of trends in the data and the identification of some technical errors suggest that a link may be present but that an improvement in the methodology employed would be required to detect it.

This thesis presents findings that suggest a function for accommodation that has not been explored in the past. It advances tools for measurement which need to be employed more widely in the field in order to provide further evidence for this function of accommodation.

# Bibliography


Abercrombie, D. (1964). *English phonetic texts*. Faber.

Adank, P. & Janse, E. (2010). Comprehension of a novel accent by young and older listeners. *Psychology and aging, 25*(3), 736.

Aguilar, L., Downey, G., Krauss, R., Pardo, J., Lane, S., & Bolger, N. (2015). A dyadic perspective on speech accommodation and social connection: Both partners' rejection sensitivity matters. *Journal of Personality*, n/a–n/a.

Alpaydin, E. (2014). *Introduction to machine learning* (3rd ed.). Cambridge, Massachusetts: The MIT Press.

Alshangiti, W. & Evans, B. (2011). Regional accent accommodation in spontaneous speech: Evidence for long-term accent change? (pp. 224–227).

Anderson, A. H., Bader, M., Bard, E. G., Boyle, E., Doherty, G., Garrod, S., …, Miller, J., et al. (1991). The hcrc map task corpus. *Language and speech*, *34*(4), 351–366.

Arndt, D. H., Lerner, J. T., Matsumoto, J. H., Madikians, A., Yudovin, S., Valino, H., …, Buxey, F., et al. (2013). Subclinical early posttraumatic seizures detected by continuous eeg monitoring in a consecutive pediatric cohort. *Epilepsia*, *54*(10), 1780–1788.

Astolfi, L., Toppi, J., Borghini, G., Vecchiato, G., He, E. J., Roy, A., …, He, B., et al. (2012). Cortical activity and functional hyperconnectivity by simultaneous eeg recordings from interacting couples of professional pilots. In *Engineering in medicine and biology society (embc), 2012 annual international conference of the ieee* (pp. 4752–4755). IEEE.

Astolfi, L., Toppi, J., De Vico Fallani, F., Vecchiato, G., Cincotti, F., Wilke, C., … Babiloni, F. (2011). Imaging the social brain by simultaneous hyperscanning during subject interaction. *Intelligent Systems, IEEE*, *26*(5), 38–45.

Astolfi, L., Cincotti, F., Mattia, D., Fallani, F. D. V., Salinari, S., Marciani, M., …, He, B., et al. (2009). Estimation of the cortical activity from simultaneous multi-subject recordings during the prisoner's dilemma. In *2009 annual international conference of the ieee engineering in medicine and biology society* (pp. 1937–1939). IEEE.

Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics using r*. Cambridge University Press.

Babel, M. (2009a). Selective vowel imitation in spontaneous phonetic accommodation. *UC Berkeley Phonology Lab Annual Report (2009)*, 163–194.

Babel, M. (2010). Dialect divergence and convergence in new zealand english. *Language in Society*, *39*(04), 437–456.

Babel, M. (2012). Evidence for phonetic and social selectivity in spontaneous phonetic imitation. *Journal of Phonetics*, *40*(1), 177–189.

Babel, M. E. (2009b). *Phonetic and social selectivity in speech accommodation* (Doctoral dissertation, University of California).

Babel, M. & Bulatov, D. (2012). The role of fundamental frequency in phonetic accommodation. *Language and speech*, *55*(2), 231–248.

Babiloni, F., Cincotti, F., Mattia, D., Mattiocco, M., De Vico Fallani, F., Tocci, A., … Astolfi, L. (2006). Hypermethods for eeg hyperscanning. In *Engineering in medicine and biology society, 2006. embs'06. 28th annual international conference of the ieee* (pp. 3666–3669). IEEE.

Babiloni, F., Cincotti, F., Mattia, D., Fallani, F. D. V., Tocci, A., Bianchi, L., … Astolfi, L. (2007). High resolution eeg hyperscanning during a card game. In *2007 29th annual international conference of the ieee engineering in medicine and biology society* (pp. 4957–4960). IEEE.

Babiloni, F. & Astolfi, L. (2014). Social neuroscience and hyperscanning techniques: Past, present and future. *Neuroscience & Biobehavioral Reviews*, *44*, 76–93. Applied Neuroscience: Models, methods, theories, reviews. A Society of Applied Neuroscience (SAN) special issue.

Bailly, G., Lelong, A. et al. (2010). Speech dominoes and phonetic convergence. *Proceedings of Interspeech 2010*, 1153–1156.

Bailly, G. & Martin, A. (2014). Assessing objective characterizations of phonetic convergence. In *15th annual conference of the international speech communication association (interspeech 2014)* (P–19).

Baker, R. & Hazan, V. (2011). Diapixuk: Task materials for the elicitation of multiple spontaneous speech dialogs. *Behavior research methods*, *43*(3), 761–770.

Baker, S. C., Gallois, C., Driedger, S. M., & Santesso, N. (2011). Communication accommodation and managing musculoskeletal disorders: Doctors' and patients' perspectives. *Health communication*, *26*(4), 379–388.

Ball, P., Giles, H., Byrne, J. L., & Berechree, P. (1984). Situational constraints on the evaluative significance of speech accommodation: Some australian data. *International Journal of the Sociology of Language*, *1984*(46), 115–130.

Ball, P., Giles, H., & Hewstone, M. (1985). Interpersonal accommodation and situational construals: An integrative formalisation. *Recent advances in language, communication, and social psychology*, 263–286.

Bandettini, P. A. (2009). What's new in neuroimaging methods? *Annals of the New York Academy of Sciences*, *1156*(1), 260–293.

Bane, M., Graff, P., & Sonderegger, M. (2010). Longitudinal phonetic variation in a closed system. *Proc. CLS*, *46*, 43–58.

Bates, D., Mächler, M., & Bolker, B. (2011). Lme4, r package version 0.999375–38.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48.

Belin, P., Zatorre, R. J., Lafaille, P., Ahad, P., & Pike, B. (2000). Voice-selective areas in human auditory cortex. *Nature*, *403*(6767), 309–312.

Benet-Martínez, V. & John, O. P. (1998). Los cinco grandes across cultures and ethnic groups: Multitrait-multimethod analyses of the big five in spanish and english. *Journal of personality and social psychology*, *75*(3), 729.

Beňuš, Š. (2014). Social aspects of entrainment in spoken interaction. *Cognitive Computation*, *6*(4), 802–813.

Berry, D. S. (1991). Attractive faces are not all created equal: Joint effects of facial babyishness and attractiveness on social perception. *Personality and Social Psychology Bulletin*, *17*(5), 523–531.

Bieniek, M. M., Pernet, C. R., & Rousselet, G. A. (2012). Early erps to faces and objects are driven by phase, not amplitude spectrum information: Evidence from parametric, test-retest, single-subject analyses. *Journal of Vision*, *12*(13), 12.

Biesmans, W., Vanthornhout, J., Wouters, J., Moonen, M., Francart, T., & Bertrand, A. (2015). Comparison of speech envelope extraction methods for eeg-based auditory attention detection in a cocktail party scenario. In *2015 37th annual international conference of the ieee engineering in medicine and biology society (embc)* (pp. 5155–5158). IEEE.

Bigdely-Shamlo, N., Mullen, T., Kothe, C., Su, K.-M., & Robbins, K. A. (2015). The prep pipeline: Standardized preprocessing for large-scale eeg analysis. *Frontiers in neuroinformatics*, *9*.

Bigham, D. S. (2010). Mechanisms of accommodation among emerging adults in a university setting. *Journal of English Linguistics*, *38*(3), 193–210.

Blank, H. & Davis, M. H. (2016). Prediction errors but not sharpened signals simulate multivoxel fmri patterns during speech perception. *PLoS Biology*, *14*(11), e1002577.

Bleackley, P. J. (2016). Markov. [Online; accessed 2017-01-03].

Boersma, P. & Weenink, D. (2016). Praat: Doing phonetics by computer [computer program, version 6.0.21]. http://www.praat.org/.

Bourhis, R. Y. & Giles, H. (1977). Language, ethnicity, and intergroup relations. In H. Giles (Ed.), (Chap. The language of intergroup distinctiveness, pp. 119–136). London: Academic Press.

Bradlow, A. R., Baker, R. E., Choi, A., Kim, M., & Van Engen, K. J. (2007). The wildcat corpus of native- and foreign-accented english. *The Journal of the Acoustical Society of America*, *121*(5), 3072–3072.

Bronstad, P. M. & Russell, R. (2007). Beauty is in the 'we' of the beholder: Greater agreement on facial attractiveness among close relations. *Perception*, *36*(11), 1674–1681.

Brooker, B. H. & Donald, M. W. (1980). Contribution of the speech musculature to apparent human eeg asymmetries prior to vocalization. *Brain and Language*, *9*(2), 226–245.

Bulatov, D. (2009). The effect of fundamental frequency on phonetic convergence. *Berkeley Phonology Lab Annual Report*, *2009*, 404–434.

Burgoon, J. K., Stern, L. A., & Dillman, L. (1995). *Interpersonal adaptation: Dyadic interaction patterns*. Cambridge University Press.

Byrne, D. E. (1971). *The attraction paradigm*. Academic Pr.

Campanella, S., Quinet, P., Bruyer, R., Crommelinck, M., & Guerit, J.-M. (2002). Categorical perception of happiness and fear facial expressions: An erp study. *Journal of cognitive neuroscience*, *14*(2), 210–227.

Campbell-Kibler, K., Walker, A., Elward, S., & Carmichael, K. (2014). Apparent time and network effects on long-term cross-dialect accommodation among college students. *University of Pennsylvania Working Papers in Linguistics*, *20*(2), 4.

Casasanto, L. S., Jasmin, K., & Casasanto, D. (2010). Virtually accommodating: Speech rate accommodation to a virtual interlocutor. In *Proceedings of the 32nd annual conference of the cognitive science society* (pp. 127–132).

Chambers, J. K. (1995). *Sociolinguistic theory*. Blackwell.

Chan, W., Jaitly, N., Le, Q., & Vinyals, O. (2016). Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *2016 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 4960–4964). IEEE.

Cho, T. & Ladefoged, P. (1999). Variation and universals in vot: Evidence from 18 languages. *Journal of Phonetics*, *27*(2), 207–229.

Claassen, J., Mayer, S., Kowalski, R., Emerson, R., & Hirsch, L. (2004). Detection of electrographic seizures with continuous eeg monitoring in critically ill patients. *Neurology*, *62*(10), 1743–1748.

Clark, H. H. (1996). *Using language*. Cambridge University Press Cambridge.

Collins, B. (1998). Convergence of fundamental frequencies in conversation: If it happens, does it matter? In *Proceedings of icslp* (Vol. 98).

Coupland, N. & Giles, H. (1988). *Communicative accommodation: Recent developments*. Pergamon Press.

Cui, G., Wong, M. L., & Lui, H.-K. (2006). Machine learning for direct marketing response models: Bayesian networks with evolutionary programming. *Management Science*, *52*(4), 597–612.

Cummins, F. (2009). Rhythm as entrainment: The case of synchronous speech. *Journal of Phonetics*, *37*(1), 16–28.

Cummins, F. (2012). Oscillators and syllables: A cautionary note. *Frontiers in psychology*, *3*(October), 364.

Davidson, P. R. & Wolpert, D. M. (2005). Widespread access to predictive models in the motor system: A short review. *Journal of Neural Engineering*, *2*, S313–19.

Davis, J. D., Winkielman, P., & Coulson, S. (2015). Facial action and emotional language: Erp evidence that blocking facial feedback selectively impairs sentence comprehension. *Journal of cognitive neuroscience*.

Davis, M. H. & Johnsrude, I. S. (2007). Hearing speech sounds: Top-down influences on the interface between audition and speech perception. *Hearing research*, *229*(1), 132–147.

Davis, S. & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, *28*(4), 357–366.

De Leon, P. L., Pucher, M., Yamagishi, J., Hernaez, I., & Saratxaga, I. (2012). Evaluation of speaker verification security and detection of hmm-based synthetic speech. *IEEE Transactions on Audio, Speech, and Language Processing*, *20*(8), 2280–2290.

de Ruiter, L. E. (2015). Information status marking in spontaneous vs. read speech in story-telling tasks – evidence from intonation analysis using {gtobi}. *Journal of Phonetics*, *48*, 29–44. The Impact of Stylistic Diversity on Phonetic and Phonological Evidence and Modeling.

Debener, S., Minow, F., Emkes, R., Gandras, K., & Vos, M. (2012). How about taking a low-cost, small, and wireless eeg for a walk? *Psychophysiology*, *49*(11), 1617–1621.

Delorme, A. & Makeig, S. (2004). Eeglab: An open source toolbox for analysis of single-trial eeg dynamics including independent component analysis. *Journal of neuroscience methods*, *134*(1), 9–21.

Delorme, A., Mullen, T., Kothe, C., Acar, Z. A., Bigdely-Shamlo, N., Vankov, A., & Makeig, S. (2011). Eeglab, sift, nft, bcilab and erica: New tools for advanced eeg processing. *Computational intelligence and neuroscience*, *2011*, 10.

Delorme, A., Sejnowski, T., & Makeig, S. (2007). Enhanced detection of artifacts in eeg data using higher-order statistics and independent component analysis. *Neuroimage*, *34*(4), 1443–1449.

Delvaux, V. & Soquet, A. (2007). The influence of ambient speech on adult speech productions through unintentional imitation. *Phonetica*, *64*(2-3), 145–173.

Deng, L., Hinton, G., & Kingsbury, B. (2013). New types of deep neural network learning for speech recognition and related applications: An overview. In *2013 ieee international conference on acoustics, speech and signal processing* (pp. 8599–8603). IEEE.

Deng, L. & Li, X. (2013). Machine learning paradigms for speech recognition: An overview. *IEEE Transactions on Audio, Speech, and Language Processing*, *21*(5), 1060–1089.

Dijksterhuis, A. & Bargh, J. A. (2001). The perception-behavior expressway: Automatic effects of social perception on social behavior. (Vol. 33, pp. 1–40). Advances in Experimental Social Psychology. Academic Press.

Ding, N., Melloni, L., Zhang, H., Tian, X., & Poeppel, D. (2016). Cortical tracking of hierarchical linguistic structures in connected speech. *Nature neuroscience*, *19*(1), 158–164.

Ding, N. & Simon, J. Z. (2014). Cortical entrainment to continuous speech: Functional roles and interpretations. *Frontiers in Human Neuroscience*, *8*, 311.

Doelling, K. B., Arnal, L. H., Ghitza, O., & Poeppel, D. (2014). Acoustic landmarks drive delta–theta oscillations to enable speech comprehension by facilitating perceptual parsing. *NeuroImage*, *85, Part 2*, 761–768. New Horizons for Neural Oscillations.

Donald, M. (1991). *Origins of the modern mind: Three stages in the evolution of culture and cognition*. Harvard University Press.

Drager, K. & Hay, J. (2012). Exploiting random intercepts: Two case studies in sociophonetics. *Language Variation and Change*, *24*(01), 59–78.

Dumas, G., Lachat, F., Martinerie, J., Nadel, J., & George, N. (2011). From social behaviour to brain synchronization: Review and perspectives in hyperscanning. *{IRBM}*, *32*(1), 48–53.

Dumas, G., Nadel, J., Soussignan, R., Martinerie, J., & Garnero, L. (2010). Inter-brain synchronization during social interaction. *PLoS ONE*, *5*(8), e12166.

Eckert, P. (1998). Language and gender: A reader. In J. Coates (Ed.), (Chap. Gender and sociolinguistic variation, pp. 64–75). Oxford: Blackwell.

Engel, A. K., Fries, P., & Singer, W. (2001). Dynamic predictions: Oscillations and synchrony in top–down processing. *Nature Reviews Neuroscience*, *2*(10), 704–716.

Evans, B. G. & Iverson, P. (2004). Vowel normalization for accent: An investigation of best exemplar locations in northern and southern british english sentences. *The Journal of the Acoustical Society of America*, *115*(1), 352–361.

Evans, B. G. & Iverson, P. (2007). Plasticity in vowel perception and production: A study of accent change in young adults. *The Journal of the Acoustical Society of America*, *121*, 3814.

Ewing, L., Caulfield, F., Read, A., & Rhodes, G. (2015). Perceived trustworthiness of faces drives trust behaviour in children. *Developmental Science*, *18*(2), 327–334.

Fitt, S. (2002). Unisyn lexicon release. *The Center for Speech Technology Research, University of Edinburgh, http://www. cstr. ed. ac. uk/projects/unisyn*.

Foulkes, P. & Docherty, G. J. (1999). *Urban voices: Accent studies in the british isles*. A Hodder Arnold Publication.

Fowler, C. A. (1986). An event approach to the study of speech perception from a direct-realist perspective. *Status Report on Speech Research, edited by IG Mattingly and N. O'Brien (Haskins Laboratories, New Haven, CT)*, 139–169.

Fowler, C. A. (2014). Talking as doing: Language forms and public language. *New ideas in psychology*, *32*, 174–182.

Fox, M. D. & Raichle, M. E. (2007). Spontaneous fluctuations in brain activity observed with functional magnetic resonance imaging. *Nature Reviews Neuroscience*, *8*(9), 700–711.

Fromont, R. & Hay, J. (2008). Onze miner: The development of a browser-based research tool. *Corpora*, *3*(2), 173–193.

Galantucci, B., Fowler, C., & Turvey, M. (2006). The motor theory of speech perception reviewed. *Psychonomic bulletin & review*.

Gales, M. & Young, S. (2008). The application of hidden markov models in speech recognition. *Foundations and trends in signal processing*, *1*(3), 195–304.

Gallois, C. & Giles, H. (2015). Communication accommodation theory. In *The international encyclopedia of language and social interaction*. John Wiley & Sons, Inc.

Gallois, C., Franklyn-Stokes, A., Giles, H., & Coupland, N. (1988). Communication accommodation in intercultural encounters. *Theories in intercultural communication*, *158*, 185.

Ganushchak, L., Christoffels, I., & Schiller, N. (2011). The use of electroencephalography in language production research: A review. *Frontiers in Psychology*, *2*, 208.

Ghitza, O. (2011). Linking speech perception and neurophysiology: Speech decoding guided by cascaded oscillators locked to the input rhythm. *Frontiers in psychology*, *2*(June), 130.

Ghitza, O. (2013). The theta-syllable: A unit of speech information defined by cortical function. *Frontiers in Psychology*, *4*, 138.

Ghitza, O., Giraud, A.-L., & Poeppel, D. (2013). Neuronal oscillations and speech perception: Critical-band temporal envelopes are the essence. *Frontiers in Human Neuroscience*, *6*(340).

Giles, H. (1973). Accent mobility: A model and some data. *Anthropological linguistics*, 87–105.

Giles, H., Bourhis, R. Y., & Taylor, D. M. (1977). Towards a theory of language in ethnic group relations. *Language, ethnicity and intergroup relations*, *307348*.

Giles, H., Coupland, J., & Coupland, N. (1991a). *Contexts of accommodation: Developments in applied sociolinguistics*. Cambridge University Press.

Giles, H., Coupland, N., & Coupland, I. (1991b). Accommodation theory: Communication, context, and. *Contexts of accommodation: Developments in applied sociolinguistics*, 1.

Giles, H., Mulac, A., Bradac, J. J., & Johnson, P. (1987). Speech accommodation theory: The first decade and beyond. *Communication yearbook*, *10*(13-48).

Giles, H., Taylor, D. M., & Bourhis, R. (1973). Towards a theory of interpersonal accommodation through language: Some canadian data. *Language in society*, *2*(2), 177–192.

Giraud, A.-L., Kleinschmidt, A., Poeppel, D., Lund, T. E., Frackowiak, R. S. J., & Laufs, H. (2007). Endogenous cortical rhythms determine cerebral specialization for speech perception and production. *Neuron*, *56*(6), 1127–1134.

Giraud, A.-L. & Poeppel, D. (2012). Cortical oscillations and speech processing: Emerging computational principles and operations. *Nature neuroscience*, *15*(4), 511–7.

Goldinger, S. D. (1998). Echoes of echoes? an episodic theory of lexical access. *Psychological review*, *105*(2), 251.

Grech, R., Cassar, T., Muscat, J., Camilleri, K. P., Fabri, S. G., Zervakis, M., … Vanrumste, B. (2008). Review on solving the inverse problem in eeg source analysis. *Journal of NeuroEngineering and Rehabilitation*, *5*(1), 25.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology*, *74*, 1464–80.

Gregory Jr, S. W. & Webster, S. (1996). A nonverbal signal in voices of interview partners effectively predicts communication accommodation and social status perceptions. *Journal of personality and social psychology*, *70*(6), 1231.

Gregory, S. W., Green, B. E., Carrothers, R. M., Dagan, K. A., & Webster, S. W. (2001). Verifying the primacy of voice fundamental frequency in social status accommodation. *Language & Communication*, *21*(1), 37–60.

Gross, J. (2014). Analytical methods and experimental approaches for electrophysiological studies of brain oscillations. *Journal of Neuroscience Methods*, *228*, 57–66.

Gwin, J. T., Gramann, K., Makeig, S., & Ferris, D. P. (2010). Removal of movement artifact from high-density eeg recording during walking and running. *Journal of neurophysiology*, *103*(6).

Hari, R. & Kujala, M. V. (2009). Brain basis of human social interaction: From concepts to brain imaging. *Physiological Reviews*, *89*(2), 453–479.

Harte, D. (2016). *Hiddenmarkov: Hidden markov models*. R package version 1.8-7. Statistics Research Associates. Wellington.

Hassan, M. R. & Nath, B. (2005). Stock market forecasting using hidden markov model: A new approach. In *Intelligent sysytems design and applications, 2005. isda'05. proceedings. 5th international conference on* (pp. 192–196).

Hastie, T. & Tibshirani, R. (1986). Generalized additive models. *Statistical Science, 1*(3), 297–318.

Hay, J. B., Pierrehumbert, J. B., Walker, A. J., & LaShell, P. (2015). Tracking word frequency effects through 130years of sound change. *Cognition, 139*, 83–91.

Heider, F. (1958). *The psychology of interpersonal relations*. Psychology Press.

Heldner, M. & Edlund, J. (2010). Pauses, gaps and overlaps in conversations. *Journal of Phonetics*, *38*(4), 555–568.

Henrich, K., Alter, K., Wiese, R., & Domahs, U. (2014). The relevance of rhythmical alternation in language processing: An erp study on english compounds. *Brain and language*, *136C*, 19–30.

Henry, M. J. & Obleser, J. (2012). Frequency modulation entrains slow neural oscillations and optimizes human listening behavior. *Proceedings of the National Academy of Sciences*, *109*(49), 20095–20100.

Hickok, G. & Poeppel, D. (2007). The cortical organization of speech processing. *Nature Reviews Neuroscience*, *8*(5), 393–402.

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A.-r., Jaitly, N., …, Sainath, T. N., et al. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, *29*(6), 82–97.

Hintzman, D. (1986). Schema abstraction in a multiple-trace memory model. *Psychological Review*, *93*(4), 411–428.

Hirata, M., Ikeda, T., Kikuchi, M., Kimura, T., Hiraishi, H., Yoshimura, Y., & Asada, M. (2014). Hyperscanning meg for understanding mother–child cerebral interactions. *Frontiers in human neuroscience*, *8*, 118.

Hirose, K. & Kawanami, H. (2002). Temporal rate change of dialogue speech in prosodic units as compared to read speech. *Speech Communication, 36*(1), 97–111.

Hlavac, M. (2015). Stargazer: Well-formatted regression and summary statistics tables. r package version 5.2.

Holmes, J. (1997). Setting new standards: Sound changes and gender in new zealand english. *English World-Wide*, *18*(1), 107–142.

Howard, D. & Angus, J. (2001). *Acoustics and psychoacoustics (2nd edition)* (F. Rumsey, Ed.). Oxford: Focal Press.

Howell, P. & Kadi-Hanifi, K. (1991). Comparison of prosodic properties between read and spontaneous speech material. *Speech Communication*, *10*(2), 163–169.

Iacoboni, M., Woods, R. P., Brass, M., Bekkering, H., Mazziotta, J. C., & Rizzolatti, G. (1999). Cortical mechanisms of human imitation. *Science*, *286*(5449), 2526–2528.

Iriarte, J., Urrestarazu, E., Valencia, M., Alegre, M., Malanda, A., Viteri, C., & Artieda, J. (2003). Independent component analysis as a tool to eliminate artifacts in eeg: A quantitative study. *Journal of Clinical Neurophysiology*, *20*(4), 249–257.

James, D., Hutter, H.-P., & Bimbot, F. (1996). Cave–speaker verification in banking and telecommunications. In *Proceedings of the ubilab conference* (Vol. 96). Citeseer.

Jensen, K. M., Borrie, S. A., Studenka, B. E., & Gillam, R. B. (2016). *Conversational alignment: A study of neural coherence and speech entrainment* (All Graduate Plan B and other Reports. Paper 779. Utah State University).

Jockers, M. L. & Witten, D. M. (2010). A comparative study of machine learning methods for authorship attribution. *Literary and Linguistic Computing*.

Johannesen, J. K., Bi, J., Jiang, R., Kenney, J. G., & Chen, C.-M. A. (2016). Machine learning identification of eeg features predicting working memory performance in schizophrenia and healthy adults. *Neuropsychiatric electrophysiology*, *2*(1), 3.

John, O. P., Naumann, L. P., & Soto, C. J. (2008). Paradigm shift to the integrative big five trait taxonomy. *Handbook of personality: Theory and research*, *3*, 114–158.

John, O., Donahue, E., & Kentle, R. (1991). The big five inventory: Versions 4a and 54, institute of personality and social research. *University of California, Berkeley, CA*.

Jones, E. E. & Davis, K. E. (1965). From acts to dispositions the attribution process in person perception. *Advances in experimental social psychology*, *2*, 219–266.

Jones, E., Oliphant, T., Peterson, P., et al. (2016). SciPy: Open source scientific tools for Python. [Online; accessed 2016-01-30].

Juang, B.-H. & Rabiner, L. R. (2005). Automatic speech recognition–a brief history of the technology development. *Georgia Institute of Technology. Atlanta Rutgers University and the University of California. Santa Barbara*, *1*, 67.

Jung, T.-P., Makeig, S., Humphries, C., Lee, T.-W., Mckeown, M. J., Iragui, V., & Sejnowski, T. J. (2000). Removing electroencephalographic artifacts by blind source separation. *Psychophysiology*, *37*(2), 163–178.

Kawasaki, M., Yamada, Y., Ushiku, Y., Miyauchi, E., & Yamaguchi, Y. (2013). Inter-brain synchronization during coordination of speech rhythm in human-to-human social interaction. *Scientific reports*, *3*, 1692.

Kelley, H. H. (1973). The processes of causal attribution. *American psychologist*, *28*(2), 107.

Kemper, S., Othick, M., Warren, J., Gubarchuk, J., & Gerhing, H. (1996). Facilitating older adults' performance on a referential communication task through speech accommodations. *Aging, Neuropsychology, and Cognition*, *3*(1), 37–55.

Kendall, T. & Thomas, E. R. (2014). *Vowels: Vowel manipulation, normalization, and plotting*. R package version 1.2-1.

Kim, J., Lee, S.-K., & Lee, B. (2014). Eeg classification in a single-trial basis for vowel speech perception using multivariate empirical mode decomposition. *Journal of Neural Engineering*, *11*(3), 36010–36021.

Kim, M., Horton, W. S., & Bradlow, A. R. (2011). Phonetic convergence in spontaneous conversations as a function of interlocutor language distance. *Laboratory Phonology*, *2*(1), 125–156.

King-Casas, B., Tomlin, D., Anen, C., Camerer, C. F., Quartz, S. R., & Montague, P. R. (2005). Getting to know you: Reputation and trust in a two-person economic exchange. *Science*, *308*(5718), 78–83.

Klatt, D. H. (1979). Speech perception: A model of acoustic-phonetic analysis and lexical access. *Journal of phonetics*, *7*(3), 279–312.

Koelsch, S., Gunter, T. C., Wittfoth, M., & Sammler, D. (2005). Interaction between syntax processing in language and in music: An erp study. *Journal of cognitive neuroscience*, *17*(10), 1565–1577.

Koiso, H., Horiuchi, Y., Tutiya, S., Ichikawa, A., & Den, Y. (1998). An analysis of turn-taking and backchannels based on prosodic and syntactic features in japanese map task dialogs. *Language and speech*, *41*(3-4), 295–321.

Konvalinka, I. & Roepstorff, A. (2012). The two-brain approach: How can mutually interacting brains teach us something about social interaction? *Frontiers in Human Neuroscience*, *6*, 215.

Kösem, A., Bosker, H. R., Meyer, A. S., Jensen, O., & Hagoort, P. (2016). Neural entrainment reflects temporal predictions guiding speech comprehension. In *The eighth annual meeting of the society for the neurobiology of language (snl 2016)*.

Krauss, R. M. & Pardo, J. S. (2006). Speaker perception and social behavior: Bridging social psychology and speech science. *Bridging Social Psychology: Benefits of Transdisciplinary Approaches. Lawrence Erlbaum Associates, Mahwah, NJ, USA*.

Kreiman, J. & Sidtis, D. (2011). *Foundations of voice studies: An interdisciplinary approach to voice production and perception*. John Wiley & Sons.

Kuhlen, A. K., Allefeld, C., & Haynes, J.-D. (2012). Content-specific coordination of listeners' to speakers' eeg during communication. *Frontiers in Human Neuroscience*, *6*(October), 1–15.

Kulesza, W., Dolinski, D., Wicher, P., & Huisman, A. (2015). The conversational chameleon: An investigation into the link between dialogue and verbal mimicry. *Journal of Language and Social Psychology*.

Laan, G. P. M. (1997). The contribution of intonation, segmental durations, and spectral features to the perception of a spontaneous and a read speaking style. *Speech Communication*, *22*(1), 43–65.

Laarne, P. H., Tenhunen-Eskelinen, M. L., Hyttinen, J. K., & Eskola, H. J. (2000). Effect of eeg electrode density on dipole localization accuracy using two realistically shaped skull resistivity models. *Brain Topography*, *12*(4), 249–254.

Labov, W. (1966). *The social stratification of english in new york city*. Washington DC, Center for Applied Linguistics.

Lakatos, P., Karmos, G., Mehta, A. D., Ulbert, I., & Schroeder, C. E. (2008). Entrainment of neuronal oscillations as a mechanism of attentional selection. *Science*, *320*(5872), 110–113.

Lan, Y., Hu, Z., Soh, Y. C., & Huang, G.-B. (2013). An extreme learning machine approach for speaker recognition. *Neural Computing and Applications*, *22*(3), 417–425.

Latinus, M., McAleer, P., Bestelmeyer, P. E., & Belin, P. (2013). Norm-based coding of voice identity in human auditory cortex. *Current Biology*, *23*(12), 1075–1080.

Lawson, E., Scobbie, J. M., & Stuart-Smith, J. (2013). Bunched /r/ promotes vowel merger to schwar: An ultrasound tongue imaging study of scottish sociophonetic variation. *Journal of Phonetics*, *41*(3–4), 198–210.

Lee, B. & Cho, K.-H. (2016). Brain-inspired speech segmentation for automatic speech recognition using the speech envelope as a temporal reference. *Scientific Reports*, *6*.

Lee, D. N. & Reddish, P. E. (1981). Plummeting gannets: A paradigm of ecological optics. *Nature*.

Lewandowski, N. (2012). *Talent in nonnative phonetic convergence* (Doctoral dissertation).

Liberman, A. M. & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition*, *21*(1), 1–36.

Lien, Y.-A. S., Gattuccio, C. I., & Stepp, C. E. (2014). Effects of phonetic context on relative fundamental frequency. *Journal of Speech, Language, and Hearing Research*, *57*(4), 1259–1267.

Lin, J. & Kolcz, A. (2012). Large-scale machine learning at twitter. In *Proceedings of the 2012 acm sigmod international conference on management of data* (pp. 793–804). SIGMOD '12. Scottsdale, Arizona, USA: ACM.

Lindenberger, U., Li, S.-C., Gruber, W., & Müller, V. (2009). Brains swinging in concert: Cortical phase synchronization while playing guitar. *BMC neuroscience*, *10*(1), 22.

Lisker, L. & Abramson, A. S. (1967). Some effects of context on voice onset time in english stops. *Language and Speech*, *10*(1), 1–28. PMID: 6044530.

Little, A. C., Burt, D. M., & Perrett, D. I. (2006). What is good is beautiful: Face preference reflects desired personality. *Personality and Individual Differences*, *41*(6), 1107–1118.

Lobanov, B. M. (1971). Classification of russian vowels spoken by different speakers. *The Journal of the Acoustical Society of America*, *49*(2B), 606–608.

Lopez-Calderon, J. & Luck, S. J. (2014). Erplab: An open-source toolbox for the analysis of event-related potentials. *Frontiers in Human Neuroscience*, *8*, 213.

Macafee, C. (1983). *Glasgow*. John Benjamins Publishing.

Macaulay, R. K. (1976). Social class and language in Glasgow. *Language in Society*, *5*(02), 173–188.

MacFarlane, A. E. & Stuart-Smith, J. (2012). 'one of them sounds sort of glasgow uniish'. social judgements and fine phonetic variation in glasgow. *Lingua*, *122*(7), 764–778.

Manson, J. H., Bryant, G. A., Gervais, M. M., & Kline, M. A. (2013). Convergence of speech rate in conversation predicts cooperation. *Evolution and Human Behavior*, *34*(6), 419–426.

McCroskey, J. C. & McCain, T. A. (1974). The measurement of interpersonal attraction. *Speech Monographs*, *41*(3), 261–266.

McCroskey, L. L., McCroskey, J. C., & Richmond, V. P. (2006). Analysis and improvement of the measurement of interpersonal attraction and homophily. *Communication Quarterly*, *54*(1), 1–31.

Mees, I. (1987). Glottal stop as a prestigious feature in Cardiff English. *English World-Wide*, *8*(1), 25–39.

Mehta, G. & Cutler, A. (1988). Detection of target phonemes in spontaneous and read speech. *Language and Speech*, *31*(2), 135–156.

Ménard, L., Schwartz, J.-L., & Aubin, J. (2008). Invariance and variability in the production of the height feature in French vowels. *Speech communication*, *50*(1), 14–28.

Miller, J. L., Green, K. P., & Reeves, A. (1986). Speaking rate and segments: A look at the relation between speech production and speech perception for the voicing contrast. *Phonetica*, *43*(1-3), 106–115.

Miller, R. M., Sanchez, K., & Rosenblum, L. D. (2013). Is speech alignment to talkers or tasks? *Attention, Perception, & Psychophysics*, *75*(8), 1817–1826.

Milroy, J. & Milroy, L. (1985). Linguistic change, social network and speaker innovation. *Journal of linguistics*, *21*(02), 339–384.

Milroy, L. & Gordon, M. (2008). *Sociolinguistics: Method and interpretation*. John Wiley & Sons.

Minguillon, J., Lopez-Gordo, M. A., & Pelayo, F. (2017). Trends in eeg-bci for daily-life: Requirements for artifact removal. *Biomedical Signal Processing and Control*, *31*, 407–418.

Mistry, D. S. & Kulkarni, A. V. (2013). Overview: Speech recognition technology, mel-frequency cepstral coefficients (mfcc), artificial neural network (ann). In *International journal of engineering research and technology* (Vol. 2, *10 (October-2013)*). ESRSA Publications.

Montague, P., Berns, G. S., Cohen, J. D., McClure, S. M., Pagnoni, G., Dhamala, M., … Fisher, R. E. (2002). Hyperscanning: Simultaneous fmri during linked social interactions. *NeuroImage*, *16*(4), 1159–1164.

Moreno, P. J. (1996). *Speech recognition in noisy environments* (Doctoral dissertation, Carnegie Mellon University).

Mullen, T., Kothe, C., Chi, Y. M., Ojeda, A., Kerth, T., Makeig, S., … Jung, T.-P. (2013). Real-time modeling and 3d visualization of source dynamics and connectivity using wearable eeg. In *Engineering in medicine and biology society (embc), 2013 35th annual international conference of the ieee* (pp. 2184–2187). IEEE.

Müller, K.-R., Tangermann, M., Dornhege, G., Krauledat, M., Curio, G., & Blankertz, B. (2008). Machine learning for real-time single-trial eeg-analysis: From brain–computer interfacing to mental state monitoring. *Journal of neuroscience methods, 167*(1), 82–90.

Namy, L. L., Nygaard, L. C., & Sauerteig, D. (2002). Gender differences in vocal accommodation:: The role of perception. *Journal of Language and Social Psychology, 21*(4), 422–432.

Neagu, A. (1997). *Analyse articulatoire du signal de parole: Caractérisation des syllabes occlusive-voyelle en français* (Doctoral dissertation, Doctoral dissertation, Signal-Image-Word specialization, Grenoble, INPG).

Nguyen, P., Tran, D., Huang, X., & Sharma, D. (2012). A proposed feature extraction method for eeg-based person identification. In *International conference on artificial intelligence*.

Nielsen, K. (2011). Specificity and abstractness of vot imitation. *Journal of Phonetics, 39*(2), 132–142.

Nilsson, J. (2015). Dialect accommodation in interaction: Explaining dialect change and stability. *Language & Communication*, *41*, 6–16. Recent developments in Communication Accommodation Theory: Innovative contexts and applications.

Nozaradan, S., Zerouali, Y., Peretz, I., & Mouraux, A. (2015). Capturing with eeg the neural entrainment and coupling underlying sensorimotor synchronization to the beat. *Cerebral Cortex*, *25*(3), 736–747.

Nunez, P. L. & Srinivasan, R. (2006). *Electric fields of the brain: The neurophysics of eeg*. Oxford University Press, USA.

Nygaard, L. C. & Queen, J. S. (2000). The role of sentential prosody in learning voices. *The Journal of the Acoustical Society of America, 107*(5), 2856.

Obleser, J., Herrmann, B., & Henry, M. J. (2012). Neural oscillations in speech: Don't be enslaved by the envelope. *Frontiers in Human Neuroscience*, *6*.

O'connell, M., Barczak, A., Ross, D., McGinnis, T., Schroeder, C., & Lakatos, P. (2015). Multi-scale entrainment of coupled neuronal oscillations in primary auditory cortex. *Frontiers in human neuroscience*, *9*, 655.

Ofli, F., Meier, P., Imran, M., Castillo, C., Tuia, D., Rey, N., …, Parkan, M., et al. (2016). Combining human computing and machine learning to make sense of big (aerial) data for disaster response. *Big Data, 4*(1), 47–59.

Öhman, S. E. (1966). Coarticulation in vcv utterances: Spectrographic measurements. *The Journal of the Acoustical Society of America, 39*(1), 151–168.

Ojeda, A., Bigdely-Shamlo, N., & Makeig, S. (2014). Mobilab: An open source toolbox for analysis and visualization of mobile brain/body imaging data. *Frontiers in human neuroscience*, *8*, 121.

Olbrich, S., Jödicke, J., Sander, C., Himmerich, H., & Hegerl, U. (2011). Ica-based muscle artefact correction of eeg data: What is muscle and what is brain? comment on mcmenamin et al. *NeuroImage*, *54*, 1–3.

Pannese, E. (2015). *Neurocytology: Fine structure of neurons, nerve processes, and neuroglial cells*. Springer.

Pardo, J. (2013). Measuring phonetic convergence in speech production. *Frontiers in psychology*, *4*, 559.

Pardo, J. S. (2006). On phonetic convergence during conversational interaction. *The Journal of the Acoustical Society of America*, *119*, 2382.

Pardo, J. S., Gibbons, R., Suppes, A., & Krauss, R. M. (2012). Phonetic convergence in college roommates. *Journal of Phonetics*, *40*(1), 190–197.

Pardo, J. S., Jay, I. C., Hoshino, R., Hasbun, S. M., Sowemimo-Coker, C., & Krauss, R. M. (2013a). Influence of role-switching on phonetic convergence in conversation. *Discourse Processes*, *50*(4), 276–300.

Pardo, J. S., Jay, I. C., & Krauss, R. M. (2010). Conversational role influences speech imitation. *Attention, Perception, & Psychophysics*, *72*(8), 2254–2264.

Pardo, J. S., Jordan, K., Mallari, R., Scanlon, C., & Lewandowski, E. (2013b). Phonetic convergence in shadowed speech: The relation between acoustic and perceptual measures. *Journal of Memory and Language*, *69*(3), 183–195.

Pardo, J. S., Urmanche, A., Wilman, S., & Wiener, J. (2016a). Phonetic convergence across multiple measures and model talkers. *Attention, Perception, & Psychophysics*, 1–23.

Pardo, J., Urmanche, A., Wilman, S., & Wiener, J. (2016b). Phonetic convergence and talker sex: It's complicated. *The Journal of the Acoustical Society of America*, *139*(4), 2105–2106.

Peelle, J. E. & Davis, M. H. (2012). Neural oscillations carry speech rhythm through to comprehension. *Frontiers in Psychology*, *3*.

Peelle, J. E., Gross, J., & Davis, M. H. (2013). Phase-locked responses to speech in human auditory cortex are enhanced during comprehension. *Cerebral cortex (New York, N.Y. : 1991)*, *23*(6), 1378–87.

Perrault, T. J., Vaughan, J. W., Stein, B. E., & Wallace, M. T. (2003). Neuron-specific response characteristics predict the magnitude of multisensory integration. *Journal of Neurophysiology*, *90*(6), 4022–4026.

Pfitzinger, H. R. (1998). Local speech rate as a combination of syllable and phone rate. In *Fifth international conference on spoken language processing*.

Piazza, C., Miyakoshi, M., Akalin-Acar, Z., Cantiani, C., Reni, G., Bianchi, A. M., & Makeig, S. (2016). An automated function for identifying eeg independent components representing bilateral source activity. In *Xiv mediterranean confer-*

*ence on medical and biological engineering and computing 2016* (pp. 105–109). Springer International Publishing.

Pickering, M. J. & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and brain sciences*, *27*(02), 169–190.

Pickering, M. J. & Garrod, S. (2013). An integrated theory of language production and comprehension. *Behavioral and Brain Sciences*, *36*(4), 329–347.

Pisoni, D. B. (1993). Long-term memory in speech perception: Some new findings on talker variability, speaking rate and perceptual learning. *Speech communication*, *13*(1), 109–125.

Poeppel, D. (2003). The analysis of speech in different temporal integration windows: Cerebral lateralization as 'asymmetric sampling in time'. *Speech communication*, *41*(1), 245–255.

Porcaro, C., Medaglia, M. T., & Krott, A. (2015). Removing speech artifacts from electroencephalographic recordings during overt picture naming. *Neuroimage*, *105*, 171–180.

Port, D. K. & Preston, M. S. (1972). Early apical stop production: A voice onset time analysis. *Haskins Laboratories Status Report on Speech Research, SR-29*, *30*, 125–149.

Powesland, P. F. & Giles, H. (1975). *Speech style and social evaluation*. Academic Press, London.

Pratto, F. & John, O. P. (1991). Automatic vigilance: The attention-grabbing power of negative social information. *Journal of personality and social psychology*, *61*(3), 380.

Prehn, K., Korn, C. W., Bajbouj, M., Klann-Delius, G., Menninghaus, W., Jacobs, A. M., & Heekeren, H. R. (2015). The neural correlates of emotion alignment in social interaction. *Social cognitive and affective neuroscience*, *10*(3), 435–443.

Purnell, T. C. (2009). Convergence and contact in milwaukee: Evidence from select african american and white vowel space features. *Journal of Language and Social Psychology*, *28*(4), 408–427.

Python Software Foundation. (2016). *Python language, version 2.7*.

R Core Team. (2016). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria.

Rabiner, L. & Juang, B. (1986). An introduction to hidden markov models. *IEEE ASSP Magazine*, *3*(1), 4–16.

Rajan, K., Abbott, L. F., & Sompolinsky, H. (2010). Stimulus-dependent suppression of chaos in recurrent neural networks. *Physical Review E*, *82*(1), 11903.

Rammstedt, B. & John, O. P. (2007). Measuring personality in one minute or less: A 10-item short version of the big five inventory in english and german. *Journal of research in Personality*, *41*(1), 203–212.

Reynolds, D. A., Quatieri, T. F., & Dunn, R. B. (2000). Speaker verification using adapted gaussian mixture models. *Digital signal processing*, *10*(1), 19–41.

Rice, S. (1973). Distortion produced by band limitation of an fm wave. *Bell System Technical Journal*, *52*(5), 605–626.

Riedner, B. A., Vyazovskiy, V. V., Huber, R., Massimini, M., Esser, S., Murphy, M., & Tononi, G. (2007). Sleep homeostasis and cortical synchronization: Iii. a high-density eeg study of sleep slow waves in humans. *Sleep*, *30*(12), 1643.

Rilling, J. K., Gutman, D. A., Zeh, T. R., Pagnoni, G., Berns, G. S., & Kilts, C. D. (2002). A neural basis for social cooperation. *Neuron*, *35*(2), 395–405.

Rosenfelder, I., Fruehwald, J., Evanini, K., & Jiahong, Y. (2011). Fave (forced alignment and vowel extraction) program suite. Retrieved from http://fave.ling.upenn.edu

Rousselet, G. A., Husk, J. S., Bennett, P. J., & Sekuler, A. B. (2008). Time course and robustness of erp object and face differences. *Journal of Vision*, *8*(12), 3.

Rousselet, G. A., Ince, R. A. A., van Rijsbergen, N. J., & Schyns, P. G. (2014). Eye coding mechanisms in early human face event-related potentials. *Journal of Vision*, *14*(13), 7.

Sancier, M. L. & Fowler, C. A. (1997). Gestural drift in a bilingual speaker of brazilian portuguese and english. *Journal of Phonetics*, *25*(4), 421–436.

Savoy, R. L. (2001). History and future directions of human brain mapping and functional neuroimaging. *Acta Psychologica*, *107*(1–3), 9–42. Beyond the decade of the brain: Towards functional neuronanatomy of the mind.

Scheuer, M. L. (2002). Continuous eeg monitoring in the intensive care unit. *Epilepsia*, *43*(s3), 114–127.

Schilbach, L., Wohlschlaeger, A. M., Kraemer, N. C., Newen, A., Shah, N. J., Fink, G. R., & Vogeley, K. (2006). Being with virtual others: Neural correlates of social interaction. *Neuropsychologia*, *44*(5), 718–730.

Schweitzer, A. & Lewandowski, N. (2014). Social factors in convergence of f1 and f2 in spontaneous speech. In *Proceedings of the 10th international seminar on speech production, cologne*.

Seneta, E. (1996). Markov and the birth of chain dependence theory. *International Statistical Review / Revue Internationale de Statistique*, *64*(3), 255–263.

Shi, L.-C. & Lu, B.-L. (2013). Eeg-based vigilance estimation using extreme learning machines. *Neurocomputing*, *102*, 135–143.

Shockley, K., Sabadini, L., & Fowler, C. A. (2004). Imitation in shadowing words. *Perception & Psychophysics*, *66*(3), 422–429.

Shoeb, A. H. & Guttag, J. V. (2010). Application of machine learning to epileptic seizure detection. In *Proceedings of the 27th international conference on machine learning (icml-10)* (pp. 975–982).

Simard, L. M. L., Taylor, D. M. D., & Giles, H. (1976). Attribution processes and interpersonal accommodation in a bilingual setting. *Language and Speech*, *19*(4), 374–387.

Simons, H. W., Berkowitz, N. N., & Moyer, R. J. (1970). Similarity, credibility, and attitude change: A review and a theory. *Psychological Bulletin*, *73*(1), 1.

Smith, J. O. & Abel, J. S. (1995). The bark bilinear transform. In *Applications of signal processing to audio and acoustics, 1995., ieee assp workshop on* (pp. 202–205). IEEE.

Sohoglu, E., Peelle, J. E., Carlyon, R. P., & Davis, M. H. (2012). Predictive top-down integration of prior knowledge during speech perception. *Journal of Neuroscience*, *32*(25), 8443–8453.

Sonderegger, M. (2015). Trajectories of voice onset time in spontaneous speech on reality tv. In *Proceedings of the 18th international congress of phonetic sciences*.

Sonderegger, M., Bane, M., & Graff, P. (Accepted). The medium-term dynamics of accents on reality television. *to appear in Language*.

Sonderegger, M. & Keshet, J. (2012). Automatic measurement of voice onset time using discriminative structured predictiona). *The Journal of the Acoustical Society of America*, *132*(6), 3965–3979.

Sotiras, A., Gaonkar, B., Eavani, H., Honnorat, N., Varol, E., Dong, A., & Davatzikos, C. (2016). Chapter 10 - machine learning as a means toward precision diagnostics and prognostics. In G. Wu, D. Shen, & M. R. Sabuncu (Eds.), *Machine learning and medical imaging* (pp. 299–334). Academic Press.

Spaak, E., de Lange, F. P., & Jensen, O. (2014). Local entrainment of alpha oscillations by visual stimuli causes cyclic modulation of perception. *Journal of Neuroscience*, *34*(10), 3536–3544.

Stephens, G. J., Silbert, L. J., & Hasson, U. (2010). Speaker -listener neural coupling underlies successful communication. *Proceedings of the National Academy of Sciences*, *107*, 14425–14430.

Stuart-Smith, J. (2004). Scottish english: Phonology. In B. Kortmann, K. Burridge, E. Schneider, R. Mesthrie, & C. Upton (Eds.), *A handbook of varieties of english: 1: Phonology* (Vol. 1, pp. 47–67). Berlin, Germany: Mouton de Gruyter.

Stuart-Smith, J. (1999). Glasgow: Accent and voice quality. *Urban voices: Accent studies in the British Isles*, 203–222.

Stuart-Smith, J., Rathcke, T., Sonderegger, M., & Macdonald, R. (2015). A real-time study of plosives in glaswegian using an automatic measurement algorithm. In *Language variation-european perspectives v: Selected papers from the seventh international conference on language variation in europe (iclave 7), trondheim, june 2013* (Vol. 17, p. 225). John Benjamins Publishing Company.

Studdert-Kennedy, M. (1986). Two cheers for direct realism ə w. *Journal of Phonetics*, *14*, 99–104.

Swaab, T., Ledoux, K., Camblin, C. C., & Boudewyn, M. A. (2012). Language-related erp components. *Oxford handbook of event-related potential components*, 397–440.

Sykes, C. B. R. & Giard, P. (2015). Sox - sound exchange (version 14.4.2). Retrieved from http://sox.sourceforge.net/

Theodoridis, S. (2015). *Machine learning: A bayesian and optimization perspective.* San Diego;London: Elsevier.

Thut, G., Veniero, D., Romei, V., Miniussi, C., Schyns, P., & Gross, J. (2011). Rhythmic {tms} causes local entrainment of natural oscillatory signatures. *Current Biology*, *21*(14), 1176–1185.

Todorov, A., Said, C. P., Engell, A. D., & Oosterhof, N. N. (2008). Understanding evaluation of faces on social dimensions. *Trends in Cognitive Sciences*, *12*(12), 455–460.

Tokuda, K., Kobayashi, T., Masuko, T., & Imai, S. (1994). Mel-generalized cepstral analysis-a unified approach to speech spectral estimation. In *Icslp* (Vol. 94, pp. 18–22).

Toppi, J., Borghini, G., Petti, M., He, E. J., De Giusti, V., He, B., … Babiloni, F. (2016). Investigating cooperative behavior in ecological settings: An eeg hyperscanning study. *PloS one*, *11*(4), e0154236.

Tran, Y., Craig, A., Boord, P., & Craig, D. (2004). Using independent component analysis to remove artifact from electroencephalographic mesured during stuttered speech. *Medical and Biological Engineering and Computing*, *42*(5), 627–633.

Trendafilov, V., Sarasso, S., Moeller, J., Staedler, C., Kaelin-Lang, A., Galati, S., et al. (2016). Synaptic homeostasis in parkinson's disease: An high-density eeg study in different stage of the disease. *Parkinsonism & Related Disorders*, *22*, e163.

Trudgill, P. (1972). Sex, covert prestige and linguistic change in the urban british english of norwich. *Language in society*, *1*(02), 179–195.

Trudgill, P. (2008). Colonial dialect contact in the history of european languages: On the irrelevance of identity to new-dialect formation. *Language in Society*, *37*(02), 241–254.

Van Engen, K. J., Baese-Berk, M., Baker, R. E., Choi, A., Kim, M., & Bradlow, A. R. (2010). The wildcat corpus of native- and foreign-accented english: Communicative efficiency across conversational dyads with varying language alignment profiles. *Language and speech*, *53*, 510–540.

Van Petten, C. & Luka, B. J. (2012). Prediction during language comprehension: Benefits, costs, and erp components. *International Journal of Psychophysiology*, *83*(2), 176–190.

Vos, D. M., Riès, S., Vanderperren, K., Vanrumste, B., Alario, F.-X., Huffel, V. S., & Burle, B. (2010). Removal of muscle artifacts from eeg recordings of spoken language production. *Neuroinformatics*, *8*(2), 135–150.

Walker, V. G. (1988). Durational characteristics of young adults during speaking and reading tasks. *Folia Phoniatrica et Logopaedica*, *40*(1), 12–20.

Wan, V. & Campbell, W. M. (2000). Support vector machines for speaker verification and identification. In *Neural networks for signal processing x, 2000. proceedings of the 2000 ieee signal processing society workshop* (Vol. 2, pp. 775–784). IEEE.

Weiss, S. D. & Houser, M. L. (2007). Student communication motives and interpersonal attraction toward instructor. *Communication Research Reports*, *24*(3), 215–224.

Wells, J. C. (1982). *Accents of english*. Cambridge University Press.

West, P. (1999a). Perception of distributed coarticulatory properties of english /l/ and /ɹ/. *Journal of Phonetics*, *27*, 405–426.

West, P. (1999b). The extent of coarticulation of english liquids: An acoustic and articulatory study. In *Proceedings of the international conference of phonetic sciences* (pp. 1901–1904). Citeseer.

West, P. (2000). *Long-distance coarticulatory effects of english/l/and/r* (Doctoral dissertation, University of Oxford).

Wheeler, T. J. & Eddy, S. R. (2013). Nhmmer: Dna homology search with profile hmms. *Bioinformatics*, btt403.

Wickham, H. & Francois, R. (2016). *Dplyr: A grammar of data manipulation*. R package version 0.5.0.

Wilson, S. M., Saygin, A. P., Sereno, M. I., & Iacoboni, M. (2004). Listening to speech activates motor areas involved in speech production. *Nature neuroscience*, *7*(7), 701–702.

Wolpert, D. M. (1997). Computational approaches to motor control. *Trends in Cognitive Sciences*, *1*, 209–16.

Wood, S. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society (B)*, *73*(1), 3–36.

Young, B. G., Jordan, K. G., & Doig, G. S. (1996). An assessment of nonconvulsive seizures in the intensive care unit using continuous eeg monitoring an investigation of variables associated with mortality. *Neurology*, *47*(1), 83–89.

Young, S. J., Evermann, G., Gales, M. J. F., Hain, T., Kershaw, D., Moore, G., … Woodland, P. C. (2006). *The htk book, version 3.4*. Cambridge, UK: Cambridge University Engineering Department.

Young, S. J. & Young, S. (1993). *The htk hidden markov model toolkit: Design and philosophy*. Citeseer.

Yu, A. C. L., Abrego-Collier, C., Sonderegger, M., Yu Alan, C. L., Abrego-Collier, C., & Sonderegger, M. (2013). Phonetic imitation from an individual-difference perspective: Subjective attitude, personality and "autistic" traits. *PLoS ONE*, *8*(9), e74746.

Yu, A. & Abrego-collier, C. (2011). Speaker attitude and sexual orientation affect phonetic imitation. *17*(1).

Yun, K., Chung, D., & Jeong, J. (2008). Emotional interactions in human decision making using eeg hyperscanning. In *International conference of cognitive science*.

Zatorre, R. J., Belin, P., & Penhune, V. B. (2002). Structure and function of auditory cortex: Music and speech. *Trends in cognitive sciences*, *6*(1), 37–46.

Zatorre, R. J., Evans, A. C., Meyer, E., & Gjedde, A. (1992). Lateralization of phonetic and pitch discrimination in speech processing. *Science*, *256*, 846–849.

Zhang, Y., Zhao, Q., Zhou, G., Jin, J., Wang, X., & Cichocki, A. (2016). Removal of eeg artifacts for bci applications using fully bayesian tensor completion. In *Acoustics, speech and signal processing (icassp), 2016 ieee international conference on* (pp. 819–823). IEEE.

Zhdanov, A., Nurminen, J., Baess, P., Hirvenkari, L., Jousmäki, V., Mäkelä, J. P., … Parkkonen, L. (2015). An internet-based real-time audiovisual link for dual meg recordings. *PloS one*, *10*(6), e0128485.

Zwicker, E. (1961). Subdivision of the audible frequency range into critical bands (frequenzgruppen). *The Journal of the Acoustical Society of America*, *33*(2), 248–248.

# Appendices

# Appendix A

# Transcription Protocol

The transcription protocol that was followed, barring some exceptions as noted in subsection 3.3.3, is reported here. This transcription protocol was developed for the 'Sounds of the City' project within the department (see: http://soundsofthecity.arts.gla.ac.uk/) and is reported verbatim, all spelling errors etc. were present in the original protocol.

## A.1 Transcription protocol

**Orthography**

The major aim of all transciption work is to annotate orthographically what is being said and to time-align orthography and sound using the Transcriber software.

While transcribing, you should use standard British English orthography except where Scots forms are used (more on this below.) Your strategy should be to try to satisfy the English-language spell-checker of an imaginary word processor as much as possible, but knowing that you can't do anything to stop it from sounding the alarm if it encounters a non-English, i.e. Scots, form. So, for example, words ending in -ing that are pronounced -in' should, nonetheless, be transcribed with the full -ing ending. Similarly, any omissions or deletions of word-final consonants should be transcribed using standard spelling conventions (e.g. understand even if pronounced as understan, myself even if pronounced as mysel, etc). Function words like and, of, with, etc will often be reduced, i.e., pronounced an' or 'n', o' and wi', etc. Please always use standard orthography if you hear such cases. Transcribing because as 'cause (but not cos) is fine.

Apostrophes should be used as appropriate (e.g., o'clock, can't; wasn't; also for possessives, e.g., Peter's). Please do not use a hyphen in complex numerals such as ninety-five or compounds like sub-contract or upper-crust. Write the compound either as two words (upper crust, ninety five or as a single word subcontract). We use hyphens only to mark some cases of dysfluencies as described below. Do not transcribe compounds like sea boots or school bag as one word. Neologisms of any kind can be transcribed as they are pronounced (e.g. oncet instead of once).

**Scots orthography**

If Scots forms are used, please transcribe them in Scots orthography, e.g. doon, cannae. For Scots words and their orthography, please consult the Scots lexicon provided at the end of this document for your reference. If a Scots form occurs that you know to be Scots, but you are unsure about how to transcribe it and it is missing from the list, please mark those items as e.g. [?Scots?-]gonnae[-?Scots?] in the transcript using <ctrl><d> and keep a list of those cases so that you can draw TR's and BJ's attention to the items you identified when submitting your transcript. The Scots lexicon below is probably not an exhaustive list of all Scots forms that you might encounter but it should provide you with a core vocabulary.

Words or morphemes (e.g., past tense verbal markers and/or other Scots endings) that are recognised as distinctly Scots but that are not included in the lexicon below, should also be transcribed using Scots orthography (e.g., callt for English called, blamt for Engl blamed, feart for Engl feared, phont for Engl phoned, pult for English pulled etc).

Note that the negative marker 'nae' is used fairly productively in Scots (e.g., cannae, disnae, naebody, wasnae) and, so, you may encounter it in forms other than those that are listed in the lexicon below. The same is true for 'oo' pronunciations of 'ou' and 'ow' words (e.g., aboot for about, doon for down, hoose for house, moose for mouse, etc). Therefore, if you encounter such a form, it is not necessary to tag it as [?Scots?] in your transcript.

Many words spelled with an "o" in English (corresponding to /a/ and /o/-like sounds) will be pronounced with /o/- or /e/-vowel in Scots lexica (Scots /o/ for English /a/: e.g. boattle for bottle, boat for bought, boax for box and Scots /e/ for English /o/ e.g. lane for loan, bane for bone, hame for home, stane for stone, claes for clothes, ain for own). Please write them in Scots orthography. In cases of ambiguity with English homographs (boat, lane, bane etc.), additionally tag them as [Scots].

**Punctuation**

It should be borne in mind that the identification of grammatical sentences in spoken language is, in a sense, an arbitrary judgement made by the transcriber, and two transcribers might identify grammatical boundaries in different places.

The transcriber should use their speaker intuition in deciding whether a full-stop or a comma should be used, and where none is necessary. Please avoid using colon or semi-colon. Do not use ellipsis.

Turns and pauses Try to keep one speaker's turns as short as possible, shorter than a line in Transcriber stretching across your whole computer screen (but do not set a turn boundary without a plausible syntactic or prosodic break). To insert a new turn by a new speaker, press <ctrl> + <t>. Please note that it is better to put boundaries in gaps in speech rather than at ends of grammatical units.

Turns are often followed by, or interspersed with, pauses. Longer pauses should

be given a separate turn. The turn won't have any orthographic transcription but will be marked as a noise event. To create a noise event, press <ctrl> + <d>, type in pause and press <rtn>. Please make sure that a new spoken turn starts exactly where the words begin (tip: look out for the onset of the sound wave in the lower panel of Transcriber).

To adjust an existing boundary, go to the green and blue panels in the lower part of the main Transcriber window, place your cursor on the boundary to be adjusted, press and hold <ctrl> while dragging the boundary to the desired location with the left mouse button. It is important not to put the boundary in too early, meaning that you will hear the final sound of a turn in the next turn. We have noticed that sometimes, Transcriber does not save the changes in a boundary placement. This is a major issue, especially in transcript revisions. It seems that after the temporal adjustment you undertake on a particular turn, you have to play the turns surrounding the boundary that you moved, otherwise Transcriber won't store the information about temporal changes in the turns. Another way of adjusting boundaries is to delete an existing boundary and to insert a new one. Precise boundary setting is an essential part of the revision task, so please make sure that your boundary adjustments are saved.

If your task is to revise an old transcript, it would be a good idea to check out all instances of turns with typed [pause] in them and to replace [pause] by a noise event [pause].

**Quasi-linguistic phenomena (hesitations, back-channel particles, interjections)**

Following particles should be used: aah, aha, aw, eh, ehm, er, erm, hmm, huh, mmm, mmhm, oh, ooh, oops, ouch, phew, tsh, tsk, uh, uh-huh, uh-uh, um, urgh, yup This list can be expanded if necessary.

**Modified Speech**

Quoted speech should be indicated using single quotation marks. If some words are laughed, sung, shouted, whispered or the like, they should be marked as such by creating a noise event: highlight the word(s), press <ctrl> + <d>, specify the modification as laughed, sung, shouted, whispered or whatever it is, press <entr>.

**Spelled words and acronyms**

Where an informant spells out a word, each letter should be written as a capital, followed by a blank. For example, where the name 'Mary Smith' is spelled out by a speaker, it should be transcribed as: M A R Y S M I T H

**Dysfluencies**

There is a whole range of dysluencies which may happen in spontaneous speech. Truncated words like in goodb- will be marked by a hyphen no!. Similarly, false starts of a word (e.g. He d- died two years ago.) should be marked by a hyphen no! and followed by a <blnk> (NB please do not forget to press blank space! Do not type in d-died, cf. the table summarising old vs. new annotation conventions in the section on transcription corrections below). If there is a false start of an identi-

fiable word but there is a discrepancy between its orthography and pronunciation, transcribe the segments pronounced not the first letters of the words, e.g. write hw- instead of wh-. REVISION / UPDATE: Rather than using a hyphen (-) for false starts, use a tilde/squiggly ( ) instead. For the simple sake of consistency, also use for truncated words.

Do not use any notation if a word was uttered completely but repeated several times (e.g. I I think so.) Words spoken with an audible hesitation (e.g. pronounced as Maybeee tommmmorow?) should be marked as such by creating a noise event: select the words maybe and tomorrow, press <ctrl> + <d>, type in dysfluent, press <rtn>. Please do not use this tag for any other cases of dysfluency. If the speaker mumbles something you cannot understand, use the [unclear] tag (see also table below).

Non word vocalisations and events (can be expanded) Non-lexical vocalisations (such as burp, click, cough, exhale, giggle, gulp, inhale, laugh, sneeze, sniff, snort, sob, swallow, throat, yawn) will all be transcribed as noise-events by pressing <ctrl> + <d>, typing in the specification of the vocalisation and pressing return. Please note that you should not type “burp”, “click”, “cough”, etc into the transcript.

It is not necessary to indicate all ‘events’ (like a door bell, dog barking); only those which appear to affect the language. Therefore, if a student cough in a lecture makes the speaker pause or rephrase him/herself, it should be indicated, but, say, a mobile phone in the background, or a plane going over, which does not affect the speech, need not be marked.

# Appendix B

# Descriptives for BFI and IA data

## B.1 BFI descriptives

| | n | mean | sd | median | min | max | range | skew |
|---|---|---|---|---|---|---|---|---|
| Extraversion | 12 | 3.73 | 0.70 | 3.50 | 2.88 | 4.88 | 2.00 | 0.60 |
| Agreeableness | 12 | 4.23 | 0.48 | 4.05 | 3.56 | 4.89 | 1.33 | 0.27 |
| Conscientiousness | 12 | 4.00 | 0.67 | 4.22 | 3.00 | 5.00 | 2.00 | -0.34 |
| Neuroticism | 12 | 2.65 | 0.64 | 2.56 | 1.50 | 4.13 | 2.63 | 0.55 |
| Openness | 12 | 3.72 | 0.56 | 3.85 | 2.90 | 4.50 | 1.60 | -0.01 |

Table B.1: Descriptive statistics for BFI data.

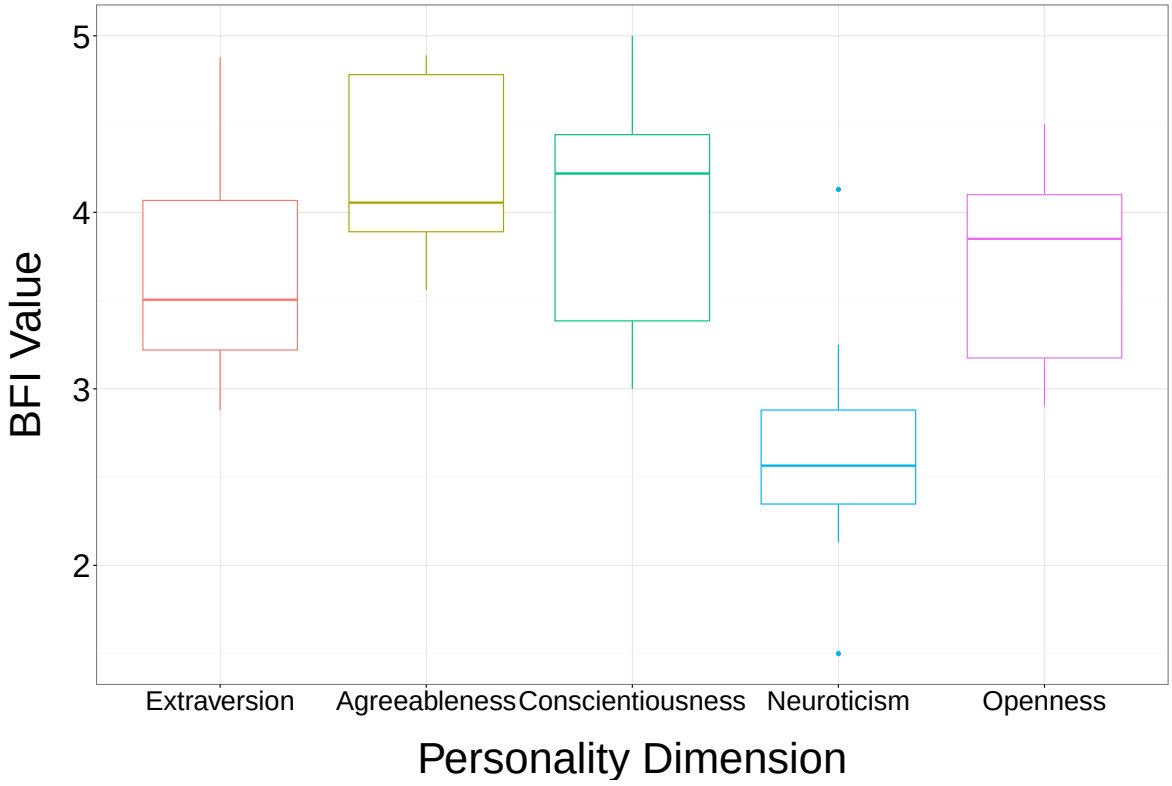



Figure B.1: Boxplots of the BFI data.

## B.2 IA Descriptives

| IA Dimension | n | mean | sd | median | min | max | range | skew |
|---|---|---|---|---|---|---|---|---|
| Social | 11 | 24.27 | 4.10 | 23 | 18 | 30 | 12 | -0.04 |
| Physical | 11 | 20.27 | 2.65 | 20 | 16 | 26 | 10 | 0.47 |
| Task | 11 | 27.36 | 2.69 | 28 | 22 | 30 | 8 | -0.83 |

Table B.2: Descriptive statistics of Interpersonal Attraction results for all participants except ARA14.

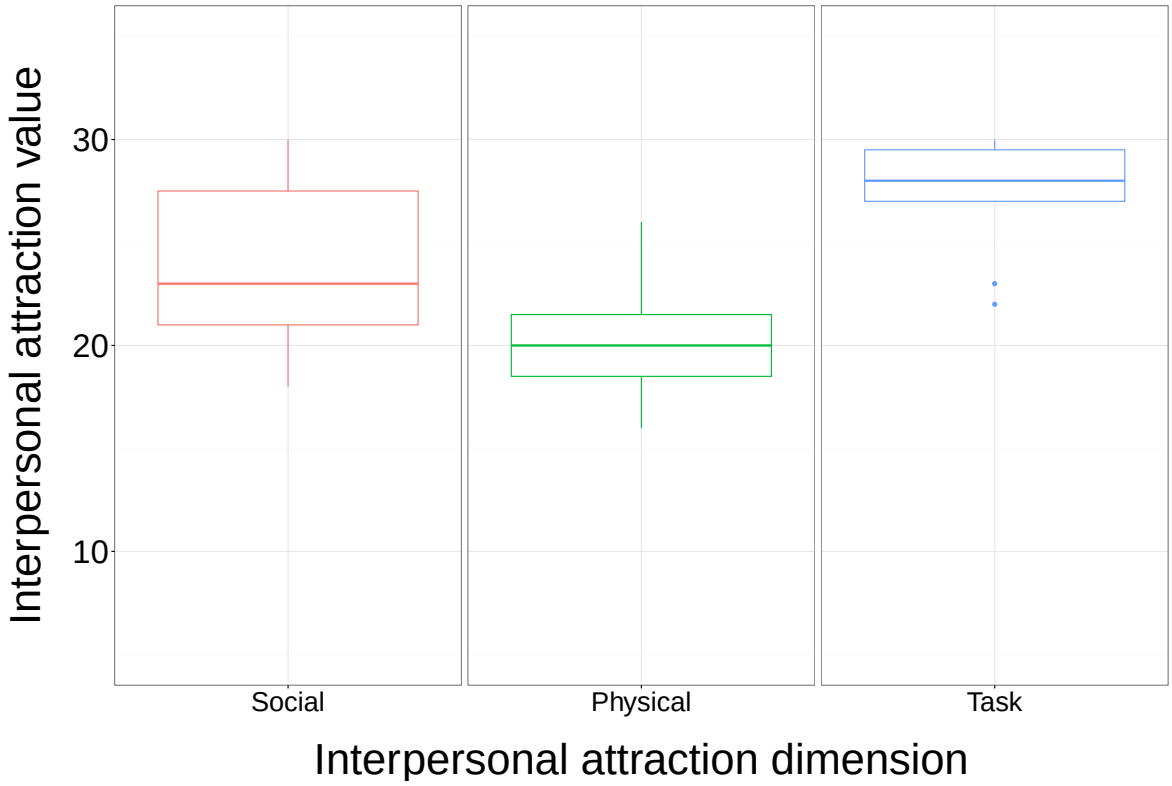



Figure B.2: Box plot of interpersonal attraction results. Data are separated by interpersonal attraction dimension. n = 11 for each dimension.