Speech Imagery as Corollary Discharge

by

Mark Scott

B.A. Linguistics, Memorial University of Newfoundland, 1997
M.A. Linguistics, Memorial University of Newfoundland, 2001

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF

Doctor of Philosophy

in

THE FACULTY OF GRADUATE STUDIES
(Linguistics)

The University Of British Columbia
(Vancouver)

April 2012

© Mark Scott, 2012
Abstract

This thesis tests the theory that the sensory content of inner speech is constituted by corollary discharge. Corollary discharge is a signal generated by the motor system and is a “prediction” of the sensory consequences of the motor system’s actions. Corollary discharge normally functions in the nervous system to segregate self-caused sensations from externally-caused sensations. It does this, partially, by attenuating the nervous system’s response to self-caused sensations. This thesis argues that corollary discharge has been co-opted in humans to provide the sensory content of speech imagery. The thesis further tests the claim that the sensory detail contained in speech imagery is sufficiently rich and sufficiently similar to the representations of external speech sounds that the perception of external speech sounds can be influenced by inner speech. This thesis claims that the perception of external speech is altered because corollary discharge prepares the auditory system to hear those sensory features which the corollary-discharge signal carries.

These claims were tested experimentally by having participants engage in specific forms of speech imagery while categorizing external sounds. In one set of experiments, when external sound and speech imagery were in synchrony and were similar in content, the perception of the external sound was altered — the external sound came to be heard as matching the content of the speech imagery. In a second set of experiments, the presence of corollary discharge in speech imagery was tested. When a sensation matches a corollary discharge signal, the sensation tends to have an attenuated impact. This attenuation is a hallmark of corollary discharge. In this set of experiments, when participants’ speech imagery matched an external sound, the perceptual impact of the external sound was attenuated. Proper controls ensured that it was the degree of match between the speech imagery and the ex-
ternal sound that was responsible for this attenuation, rather than some extraneous factor.
Preface

Both experiments reported in chapter 2 were co-authored with Henny Yeung, Bryan Gick and Janet Werker.

Henny and I worked very closely on all aspects of design, implementation and analysis. I would say that our contributions are essentially equal in all those areas. There were multiple pilot studies for those experiments and Henny took the lead for all of those pilot studies. In the event, circumstances led to me taking the lead in implementing and running the versions that are included in this thesis, and thus also being responsible for their analysis and write-up. So, while I am technically first author, the work is very much a partnership and contributions are essentially impossible to disentangle.

All other experiments reported in this thesis are single-authored works.

The experiments in this dissertation were run under approval of the Behavioural Research Ethics Board, certificate no. H95-80023 (issued to Janet Werker as Principal Investigator).
Table of Contents

Abstract .............................................................. ii
Preface ................................................................. iv
Table of Contents ...................................................... v
List of Tables .......................................................... ix
List of Figures .......................................................... x
Acknowledgements ..................................................... xii
Preamble ................................................................. xiv

1 Introduction ........................................................... 1
  1.1 Overview of the Chapter ......................................... 5
  1.2 A Historical Overview of (Western) Thought on Speech Imagery . 6
  1.3 Two Types of Speech Imagery .................................... 9
  1.4 Speech Imagery Originates in the Motor System ................. 10
  1.5 Corollary Discharge .............................................. 14
    1.5.1 Forward Models ............................................... 15
    1.5.2 Inverse Models ............................................... 21
    1.5.3 The Two Functions of Corollary Discharge ............... 22
    1.5.4 Distinguishing Self-Produced from External Sensations . 23
    1.5.5 Sensory Attenuation ......................................... 24
    1.5.6 How Sensory is Corollary Discharge? .................... 34
List of Tables

Table 1.1  A Car Manual as a Toy Example of a Forward Model . . . . . 17
Table 2.1  Structure of Experiment 1-1 . . . . . . . . . . . . . . . . . . . . . . 49
Table 2.2  Structure of Experiment 1-2 . . . . . . . . . . . . . . . . . . . . . . 59
Table 3.1  Actions and Target Sounds of Experiment 2-1 . . . . . . . . . . 74
Table 3.2  Number of Repetitions of Each Step along the /a’ba/∼/a’ba/
Continuum for Experiment 2-1 Pre-Test . . . . . . . . . . . . . . . . 76
Table 3.3  Number of Repetitions of Each Step along the /a’ba/∼/a’ba/
Continuum for Experiment 2-2 Pre-Test . . . . . . . . . . . . . . . . 84
Table 3.4  Actions and Target Sounds of Experiment 2-3 . . . . . . . . . . 96
Table 4.1  Conditions and Context Sounds of Experiment 3-1 . . . . . . . 116
Table 4.2  Number of repetitions of Each Step along the /da/∼/ga/ Con-
tinuum for Experiment 3-1 Pre-Test . . . . . . . . . . . . . . . . . . . 117
Table 4.3  Number of repetitions of Each Step along the /da/∼/ga/ Con-
tinuum for Experiment 3-2 Pre-Test . . . . . . . . . . . . . . . . . . . 125
Table 4.4  Conditions and Context-Sounds of Experiment 3-3 . . . . . . . 131
## List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>An Open-Loop Model (i.e., no Feedback)</td>
<td>15</td>
</tr>
<tr>
<td>1.2</td>
<td>A Closed-Loop Model (i.e., with Feedback)</td>
<td>16</td>
</tr>
<tr>
<td>1.3</td>
<td>Efference Copy and Corollary Discharge</td>
<td>19</td>
</tr>
<tr>
<td>2.1</td>
<td>Timeline of Stimulus Presentation (Experiment 1-1)</td>
<td>52</td>
</tr>
<tr>
<td>2.2</td>
<td>Experiment 1-2 Results</td>
<td>56</td>
</tr>
<tr>
<td>2.3</td>
<td>Experiment 1-2 Results</td>
<td>62</td>
</tr>
<tr>
<td>3.1</td>
<td>Schematic Outline of Experiment 2-1</td>
<td>75</td>
</tr>
<tr>
<td>3.2</td>
<td>Results of Experiment 2-1</td>
<td>79</td>
</tr>
<tr>
<td>3.3</td>
<td>Strength of Disambiguation across Three Conditions in Recalibration Experiment 2-1</td>
<td>80</td>
</tr>
<tr>
<td>3.4</td>
<td>Schematic Outline of Experiment 2-2</td>
<td>85</td>
</tr>
<tr>
<td>3.5</td>
<td>Results of Experiment 2-2</td>
<td>87</td>
</tr>
<tr>
<td>3.6</td>
<td>Strength of Disambiguation across Three Conditions in Recalibration Experiment 2-2</td>
<td>88</td>
</tr>
<tr>
<td>3.7</td>
<td>Schematic Outline of Experiment 2-3</td>
<td>98</td>
</tr>
<tr>
<td>3.8</td>
<td>Degree of Adaptation across Three Conditions in Experiment 2-3</td>
<td>100</td>
</tr>
<tr>
<td>4.1</td>
<td>Corollary Discharge Function</td>
<td>105</td>
</tr>
<tr>
<td>4.2</td>
<td>Schematic of Contrast Explanation of Mann effect (/æl/ Influencing /dæ/<del>/?a/</del>/? to Sound More Like /?a/)</td>
<td>108</td>
</tr>
<tr>
<td>4.3</td>
<td>Schematic of Contrast Explanation of Mann effect (/æl/ Influencing /dæ/<del>/?a/</del>/? to Sound More Like /dæ/)</td>
<td>108</td>
</tr>
<tr>
<td>Figure 4.4</td>
<td>An Example of Orthographic Context Dependence</td>
<td>110</td>
</tr>
<tr>
<td>Figure 4.5</td>
<td>Timeline of Stimulus Presentation (Experiment 3-1)</td>
<td>118</td>
</tr>
<tr>
<td>Figure 4.6</td>
<td>Schematic of the Three Conditions of Experiment 3-1</td>
<td>119</td>
</tr>
<tr>
<td>Figure 4.7</td>
<td>Results of Experiment 3-1</td>
<td>122</td>
</tr>
<tr>
<td>Figure 4.8</td>
<td>Results of Experiment 3-2</td>
<td>127</td>
</tr>
<tr>
<td>Figure 4.9</td>
<td>Schematic of the Two Conditions of Experiment 3-3</td>
<td>133</td>
</tr>
<tr>
<td>Figure 4.10</td>
<td>Picture of the Stimulus Recording/Playback Set-Up for Experiment 3-3</td>
<td>134</td>
</tr>
<tr>
<td>Figure 4.11</td>
<td>Results of Experiment 3-3</td>
<td>139</td>
</tr>
</tbody>
</table>
Acknowledgements

I would like to thank my committee for their supervision: Bryan Gick, Eric Vatikiotis-Bateson, Janet Werker and Joe Stemberger. These are four very different people with different theoretical perspectives, but despite those differing viewpoints they allowed me a very free hand in the type of research I wanted to pursue and gave me the support I needed to pursue it. Bryan and Eric were my co-supervisors and were the first people I went to when I needed to discuss (or argue over) an issue during development. My work was all done in their labs — and I hope to replicate in my own lab (some day) some of the dynamics they have succeeded in achieving in theirs (thanks!). I’d also like to thank Bryan for essentially nursing me through the last day of my thesis submission, down to the last minutes. Joe provided much needed perspective — thanks for your time and patience. While I am a student in the Linguistics Department, this thesis simply could not have been done without support in myriad ways from Janet Werker in the Psychology Department — thank-you for letting me be a part of your lab for the past several years and thank-you for all your guidance and kindness.

I was supported by an NSERC fellowship during most of my dissertation. Funding for my research was provided by NSERC Discovery Grants to Bryan Gick, Eric Vatikiotis-Bateson and Janet Werker, as well as a James S. McDonnell Foundation grant to Janet Werker.

I would also like to express deep thanks to everyone else who helped me with (through) this thesis. My family and friends supported me through times of stress, doubt and worry. All children owe their parents more than they can ever repay and this is particularly true in my case. My parents have had years of practice in patience with me and this put them in a good position to be patient during this de-
gree (thanks and much love). I would like to thank Anita Szakay and Beth Rogers who made the Linguistics Department a fun place to be and are great friends and idea springboards (you too, Solveiga and Amelia, I miss Nerd Nights!). Thanks also to Nicole Ong who managed to be an academic friend with whom I didn’t talk academics. I would also like to thank Henny Yeung who was my partner in several experiments during my doctoral work, and was not just a partner but a good friend, and, though he is much younger than me, very much a mentor and teacher — my partnership with Henny was the foundation of this dissertation. Thanks also to all of the members of ISRL who tolerated my grouchy old man attitude (especially Donald Derrick, Murray Schellenberg, Chenhao Chiu, and honorary member Martin Oberg, who was the first person I ran to when in technical difficulty) and the members of the Werker Lab who were always there to provide suggestions when I was in trouble.

A quick shout also to some non-UBC friends: Frank Squires, Dale Jarvis, Kelly Jones, Danielle Irvine, Curtis Budden, Katrin Dohlus, Nimia Herrera, 中村織江, and all the others, thanks.
I am saying that to think is to speak and thought is speech - not to someone else, but to oneself in silence.

(Plato - Theaetetus)
Of course it is happening inside your head, Harry, but why on earth should that mean it is not real? — Albus Dumbledore
(Harry Potter and the Deathly Hallows - J.K. Rowling)

N.B. The hieroglyph on the previous page is the Egyptian symbol for “thought”, which was also the symbol for “speech”.
Chapter 1

Introduction

...la parole intérieure, qui, dans l'intelligence humaine, joue un rôle tout aussi important que la parole extérieure, s'il ne l'est davantage.

...inner speech, which, for human intelligence, plays a role just as important as external speech, if not a bigger one.

— de Cardaillac (1830)

Activation of the speech motor system is typically accompanied by an auditory experience. This is trivially true in the case of speaking out loud, but it is also true in cases where there is no externally audible sound created, such as when we silently rehearse a telephone number or rerun conversations in our head. This dissertation will examine the nature of this form of auditory imagery; hypothesizing that it is produced by a forward model in the motor system which generates a ‘mock’ sensory signal (termed corollary discharge) and that the representation of this sensory detail is similar enough to the representation of external sounds to alter their perception. Inner speech is defined in this dissertation as any introspective (not externally audible) experience of speech which has an auditory phenomenal quality.

Inner speech is one of the most common conscious acts we perform — occupying perhaps a quarter of our mental lives (Heavey and Hurlburt, 2008) and possibly more (Klinger and Cox, 1988). Inner speech is a fundamental part of the component of working memory known as the phonological loop (Baddeley, 1983; Baddeley and Hitch, 1974). It also constitutes a central component of the
experience of reading (Abramson and Goldinger, 1997; Baddeley et al., 1981). Inner speech is presumed to be the source of the voices heard by hallucinating schizophrenics (Fernyhough, 2004; McGuire et al., 1995), a particularly important area of research since schizophrenia is among the most damaging of psychological disorders — personally, socially and economically. Inner speech has been linked to language acquisition (de Guerrero, 2005), sports performance (Hardy, 2006), general skill acquisition (Clark, 2008), and childhood development (Fernyhough, 2008). Lastly, and most philosophically, inner speech has been proposed as the origin of consciousness (Dennett, 1984; see also Robinson, 2004; Steels, 2003 and Jackendoff, 2007 for related views). Yet, despite the central role played by inner speech in our mental lives, it has not been a particularly ‘hot’ topic (Morin, 2009).

While the question of how we generate the semantic and syntactic content of inner speech is interesting, it is likely to be very similar to the way we generate these aspects of normal, external speech. This dissertation is only concerned with the auditory experience of inner speech — the ‘sound’ of the voice in your head.

The main points argued in this dissertation are as follows:

- The sound of inner speech is represented similarly to the sound of external speech and, because of this, can influence the perception of external speech (chapter 2).

- Inner speech can contain auditory sensory detail and so can be more than a bare string of phonemes (chapter 2).

- The auditory sensory aspects of inner speech are constituted by corollary discharge and so are generated by a forward model tied to the motor system (chapter 4).

These claims are explored experimentally in subsequent chapters. In chapter 2, two experiments demonstrate that engaging in speech imagery alters the perception

\[ \text{There is also the (in)famous claim of Julian Jaynes, 1976; that consciousness is historically recent and is due to a reinterpretation of inner speech as being self-generated rather than as being the voices of the gods. Jaynes believes this transition can be detected in how the mental lives of characters are portrayed in the later \textit{Odyssey} vs. the earlier \textit{Iliad}. Jaynes' claim is (almost) universally rejected as pseudo-science but is still a popular topic of discussion in philosophy and cognitive science.} \]
of external speech and the details of this influence suggest that inner speech contains information below the level of the phoneme. The temporal properties of this influence are examined in chapter 3. The claim that corollary discharge constitutes the sound of inner speech is tested in chapter 4, in which sensory attenuation (a hallmark of corollary discharge) is demonstrated to occur during speech imagery.

The use of the term sensory is a potential source of debate. In fact this is a point of contention in the literature on visual imagery (e.g., Kosslyn, 1995) and to an extent in psychology as a whole (e.g., Barsalou, 1999). I do not deny that any neural representation is, by definition, an abstract representation; however, by claiming that inner speech has sensory content I merely mean that it contains detail below the level of the phoneme in the abstract processing hierarchy and that the neural representation of this detail is similar to the neural representation of ‘genuine’ sensory events and thus can reasonably be termed ‘sensory’. This issue is discussed further in subsection 1.5.6.

The details of corollary discharge will be discussed in section 1.5, but the basic idea is that in order for an animal not to make the potentially fatal mistake of confusing a self-caused sensation with a sensation caused by something in the outside world (imagine the disastrous consequences if we experienced the world as spinning every time we moved our eyes), the motor system sends a ‘mock’ sensory signal (= ‘corollary discharge’) to the sensory system as a warning of what is about to happen. This corollary discharge is a prediction of the sensory consequences of the action. The sensory system uses this prediction to segregate out self-caused sensations from the incoming flow of sensory stimulation so that its perception of external signals is not dulled or confused when it performs an action.2

This dissertation presents evidence that the ‘sound’ of inner speech simply is corollary discharge. That is, when we talk to ourselves in our head, what we are doing is running our speech motor system silently (either with or without actually

2Here and elsewhere I fall into using ‘teleological’ descriptions of corollary discharge (e.g. “the sensory system uses this prediction to segregate . . . ”). I believe this to be the clearest way to explain the concept of corollary discharge, which is, after all, a functionally defined term. However, I should make it clear that ‘function’ in a biological sense is not the same as ‘function’ in an everyday human sense. I am, of course, not claiming that the motor system or the sensory system has intentions or purposes. I do claim that these systems can be profitably analyzed as having functions in a strictly evolutionary sense.
moving our articulators) and thus generating a ‘mock’ sensory signal that constitutes a prediction of the sound of our own voice.

There is a debate in the literature as to whether inner speech is a single entity or whether it should be divided into two types, one ‘phonological’, which does not require any engagement of the speech articulators, and one ‘sensory’, which derives its sensory content from the engagement of the speech articulators (see section 1.3). Several of the experiments reported in this dissertation compare the effects of enacted (with speech movements) and non-enacted (with no speech movements) forms of speech imagery. However, this dissertation does not provide a crucial test of this distinction and so I will not take a position on the issue.

I would also add that the theory presented in this dissertation is relevant to models of speech perception that involve the motor system, e.g., the Motor Theory of Speech Perception (Liberman and Mattingly, 1985; Liberman and Whalen, 2000) and its variants. While this dissertation is not about such motor-based theories, it does offer an alternative explanation of some of the evidence used to support these theories, e.g., evidence that triggering the motor system alters speech perception (as discussed in Experiments 1-1 to 1-3 in chapter 2). Such motor-based perceptual effects have been interpreted as supporting a motor theoretic approach to perception; however, I believe a better interpretation is that these experiments are demonstrating the effects of corollary discharge, which certainly can affect perception if triggered, but would not usually be triggered in the normal course of speech perception (but see section 5.1). While a corollary discharge explanation of such effects does not rule out motor theories, by providing an alternative explanation for some of their supporting evidence it does ‘steal some of their thunder’.

A final comment is that the theory presented in this dissertation is ambiguously part of the recent turn towards embodiment in cognitive science. Embodiment is the view that much of cognition can profitably be seen as based in the body and in the body’s relationship with the environment. This allows much of the processing of cognition to be offloaded onto the body and the world: Instead of computing a detailed representation of the world, embodiment follows the motto that ‘the world is its own best representation’. Such a view contrasts with a more computational view of cognition in which cognition centres around the development of rich internal models of the world.
The theory presented in this dissertation falls both within and outside of this embodiment approach. The theory is certainly body-based in that it claims that a major component of our cognition is based in the motor system and will be dependent on the quirks of our body (what makes each vocal tract unique); however it is also heavily computational in that I will argue that the phenomenal experience of our inner voice is based on a rich computational model of our body, rather than being based directly on the body itself. I did not aim to straddle both camps in this dissertation, but I am not disappointed because I find value in both views.

1.1 Overview of the Chapter

The following is a brief ‘roadmap’ of this chapter. The chapter starts with a brief overview of the history of thought on inner speech followed by a review of the evidence that the production of inner speech is tied to the motor system.

Following this introductory material, the chapter moves on to the conceptual heart of the dissertation: corollary discharge. The central proposal of this dissertation is that corollary discharge constitutes the auditory, sensory, experience of inner speech, and thus corollary discharge is, of necessity, discussed in detail.

Corollary discharge is a functionally-defined term that was initially proposed as a theoretical solution to certain problems in theories of motor control. This chapter gives a detailed discussion of the a priori reasons for postulating corollary discharge, as well as the related functional concepts of ‘forward model’, ‘inverse model’, and ‘efference copy’. This discussion is followed by a review of the overwhelming experimental evidence (behavioural, brain-imaging, and neuroanatomical) demonstrating that these postulated mechanisms are indeed instantiated in animals, including humans. The experimental evidence reviewed in this chapter covers multiple sensory modalities and a wide array of animal models, but this is still only a small subset of the available evidence in the literature, driving home the point that corollary discharge is not just a theoretical construct but is an extremely well-supported fact of biology.

Inner speech is, of course, at the intersection of a vast number of important issues: consciousness, mental health, literacy, childhood development, language acquisition . . . ; enough to provide several careers worth of exploration. The subset
of these issues discussed in this dissertation only scratches the surface of a wealth of interesting and important avenues of future research.

1.2 A Historical Overview of (Western) Thought on Speech Imagery

What has been will be again, what has been done will be done again; there is nothing new under the sun. — (Ecclesiastes 1:9 NIV)

The following section gives a brief historical sketch of research on inner speech. The purpose of this section is to provide historical context and by doing so to emphasize the centrality of inner speech in cognition — inner speech is not a new or a peripheral aspect of our mental lives, but something that has been at centre-stage as far back as we can tell.\(^3\)

It is clear that the phenomenon of silently talking to oneself is not at all new. Plutarch, writing in the first century C.E. comments that the fact that there are two types of speech, one internal and the other external, is so well known that it is a ‘threadbare’ fact:

“The statement that there are two kinds of speech, one residing in the mind, the gift of Hermes\(^4\) the Leader, and the other residing in the utterance, merely an attendant and instrument, is threadbare”

Writing about four and a half centuries earlier, Plato makes several references to inner speech and identifies internal dialogue with thought. In The Sophists he remarks:

“Well, then, thought and speech are the same; only the former, which is a silent inner conversation of the soul with itself, has been given the special name of thought.”

This idea is also found, most famously, in the Theaetetus (as quoted in the preamble to this chapter). The idea that thought is an internalized conversation is similar to the Vygotskian notion of childhood development discussed below.

\(^3\)Much of this historical information is discussed in Panaccio (1999).

\(^4\)Several ancient authors mention that Hermes was considered the ‘patron god’ of inner speech.
Plato’s most famous student, Aristotle, frequently discusses inner speech, though in Aristotle it is unclear whether he is talking about an internalized human language or an abstract ‘language of thought’ in the style of Fodor (1975).

There are numerous other authors in the classical world who discuss inner speech, including Quintilian, who discusses the benefit of using inner monologue to memorize speeches. Heraclitus and Ptolemy (the Ptolemy who developed the Ptolemaic model of the solar system) both discuss inner speech (Panaccio, 1999), as does Philon of Alexandria, who wrote a dialogue in which he discusses the intelligence of animals and whether they have inner speech (some even have external speech, according to Philon, such as parrots). This point about animals is of particular interest since it now seems that some birds do in fact use subvocal articulation (perhaps equivalent to inner speech in humans) to practice their songs in silence (Cooper et al., 2006).

Interest in inner speech was maintained in the Stoic school of philosophy from which it spread into the doctrines of the early Christian church. The relationship between inner and external speech led many early Christian theologians to claim that Jesus’ theological relationship to the Christian god could be considered analogous to the relationship of external speech to inner speech — the external being the embodiment of the internal. The idea was so common that it was condemned as a heresy by the early church (Panaccio, 1999, p.95).

From Christian theology, the discussion of inner speech became an important thread in the scholastic tradition. Inner speech was a central theme in the work of several of the most important figures in this tradition. Albertus Magnus and his student Thomas Aquinas both discussed the concept of inner speech in detail. In fact, Thomas Aquinas developed an elaborate psychological model that included inner speech as a component (and a language of thought as another component). William of Occam followed on Aquinas’ work, discussing the relation of speech to thought.

In the 19th century, interest in inner speech was revived in France by several philosophers/psychologists, de Cardaillac, Egger, and Ballet, who wrote extensive works on the topic. De Cardaillac and Ballet viewed inner speech as based in the

---

5 The Aristotelian-based scholarly tradition of the medieval university system.
motor system, whereas Egger thought that it was essentially a product of auditory memory.

The advent of modern psychology is usually tied to the work of William James\(^6\) in America and Alfred Binet in France. Both of whom discussed inner speech, though neither committed himself to whether inner speech was necessarily motoric in origin (Binet 1886; James 1890)\(^7\).

The concept of inner speech as originating in the motor system was the basis of the behavioural approach. Indeed, Watson (1913, 1914, 1920), one of the founders of behaviourism, claimed that inner speech was simply a matter of micro-movements of the articulators. This view predominated in the behaviourist approach, as discussed in section 1.4.

Around the same time as the behaviourist movement took hold in the west, inner speech was being investigated under a very different framework in the Soviet Union. The leading figure was Vygotsky, who developed a theory of psychological development which viewed the social interaction between a child and others (particularly parents) as being crucial (Akhutina 2003). Under this model, the conversations between parents and children form the basis of the self-talk children engage in when alone. This self-talk is useful in solving problems — children learn that, by imitating the dialogues they have had with their parents, they can find solutions to problems that were not accessible without the dialogue. Eventually the dialogue is completely internalized and this, according to Vygotsky, is how inner speech develops.

Vygotsky’s ideas were expanded by Luria and Sokolov (whose work on inner

---

\(^6\)Who is reported to have quipped “the first lecture in psychology I ever heard was the first I ever gave” (Pajares 2003).

\(^7\)While James does not unambiguously support the idea that inner voice is dependent on the motor system, he does say that “Most persons, on being asked in what sort of terms they imagine words, will say ‘in terms of hearing.’ It is not until their attention is expressly drawn to the point that they find it difficult to say whether auditory images or motor images connected with the organs of articulation predominate. A good way of bringing the difficulty to consciousness is that proposed by Stricker: Partly open your mouth and then imagine any word with labials or dentals in it, such as ‘bubble’, ‘toddle’. Is your image under these conditions distinct? To most people the image is at first ‘thick’, as the sound of the word would be if they tried to pronounce it with the lips parted. Many can never imagine the words clearly with the mouth open; others succeed after a few preliminary trials. The experiment proves how dependent our verbal imagination is on actual feelings in lips, tongue, throat, larynx, etc.” (James 1890, p.62).
speech is discussed in section 1.4).

In recent years, work on inner speech has primarily been done as a component of research in other areas, such as research on schizophrenia, reading or the phonological loop.

### 1.3 Two Types of Speech Imagery

Oppenheim and Dell (2011) have argued for a ‘flexible abstractness of inner speech’ which is the claim that inner speech is primarily an abstract phonological code (i.e., a string of abstract category-labels = bare phonemes) that can be supplemented with some degree of motor engagement to provide a more detailed ‘phonetic’ experience. They provide evidence from speech error experiments that the types of errors reported in non-enacted inner speech (without engagement of the articulators) are those that can be accounted for via purely ‘category-level’ processes, while the errors reported in enacted inner speech (with engagement of the articulators) contains some errors that are only attributable to motor processing. Oppenheim and Dell performed an earlier experiment which also found that speech errors in non-enacted inner speech are primarily categorical in nature (Oppenheim and Dell, 2008). When inner speech involves some degree of articulator movement (however slight) it is often called *subvocal articulation*.

The results of Reisberg (1989) support the ‘flexible abstractness’ view of inner speech — He showed that potential ambiguities in speech imagery are better detected when people are allowed to engage their motor system (enacted inner speech), suggesting that enacted inner speech is more like a real-world sensory experience than non-enacted inner speech.

Of course, if there are these two kinds of inner speech, the enacted kind en-

---

8 By *phonetic* they mean including articulatory information.

9 This is reminiscent of the discussion of reinterpretability in the visual imagery literature. The Necker cube is ambiguous between two different percepts (the cube can be seen as being in either of two orientations: near face pointing down and to the right, or near face pointing up and to the left), and people often have the experience of the orientation flipping in their perception.

However, this ambiguity does not occur in imagery — when one visualizes the necker cube in the mind’s eye it is always visualized in a particular orientation and does not flip (Chambers and Reisberg, 1985).
tails the presence of the phonological kind. That is, if a person is engaged in enacted speech imagery (moving their mouth as they engage in inner speech), this can only happen if the string of phonemes to be mouthed has been decided on. Thus, mouthing sounds implies the presence of a string of phonemes but a string of phonemes does not have to be mouthed. This means that pure inner speech, under the ‘flexible abstractness’ hypothesis, would essentially involve experiencing a string of phonemes while enacted inner speech would involve a string of phonemes, but also more detailed sensory information provided by engagement of the motor system.

My dissertation claims that corollary discharge provides the sensory content we experience when we talk to ourselves. If we adopt the flexible abstractness theory proposed by Oppenheim and Dell, then the presence of corollary discharge (and hence sensory content) in inner speech would exist on a continuum depending on the degree of motor-system engagement. On one end of the continuum, the motor system would not be engaged at all and inner speech would be a purely symbolic phonemic code with no sensory content. At the other end of the continuum, inner speech would be generated with the full execution of motor commands and a great deal of sensory detail experienced (provided by corollary discharge).

This dissertation makes no claim about non-speech forms of auditory imagery (such as imagining the sound of a door slamming). Such non-speech imagery may be related to speech imagery, but that question is beyond the scope of this dissertation.

1.4 Speech Imagery Originates in the Motor System

This section reviews the evidence for the claim that inner speech is generated in the motor system (similarly to external speech). Of course, there is a large background of semantic and syntactic processing that would have to be done before either inner or external forms of speech are generated. When I say that speech (inner or external) originates in the motor system, I am referring to the sound of speech, whether that sound is made external or simply ‘heard in the head’. The choice of words and their syntactic and semantic integration would have, of necessity, occurred before this.
The view that inner speech originates in the motor system is not new or particularly controversial. The idea was well developed in the behaviourist movement, but even before that there were several physiologists and philosophers who argued for it. The French philosopher de Cardaillac, writing in 1830, argued that memory of sounds was triggered by subtle movements of the articulators that would normally produce them and this is what generates our auditory experience in inner speech. This idea was picked up by others including the physiologist Stricker (1885) who observed that it is difficult to hear a speech sound in inner voice if the mouth is held in the configuration for a different speech sound, an observation echoed by William James and Alfred Binet.10 The Russian psychologist Sechenov believed that private thoughts were constituted by speech reflexes that were interrupted before actual motor execution occurred (Daniels et al., 2007, p.37).

De Cardaillac made another important point that has been repeated by others, namely that inner speech is unlike other forms of imagery in that it is under our immediate and detailed control.11 When we try to imagine an object visually, it can be difficult to generate the image instantaneously; it takes time to materialize. This is not the case with inner speech, which we are able to generate on demand just as if it were external speech. This is reminiscent of the type of control we have over motor acts. This ‘isochrony’ of inner and external speech has been confirmed experimentally (e.g., Weber and Bach, 1969).

There are three streams of experimental evidence supporting the motor origin of inner speech: First, the fact that inner speech is often accompanied by micro-movements of the speech articulators; second, brain imaging studies showing activation of motor areas during inner speech production; and third, evidence from Transcranial Magnetic Stimulation (TMS) showing that interrupting activity in motor areas of the brain interrupts inner speech.

The most famous statement of the idea of a motor origin to inner speech was,

---

10 Subsection 5.2.2 contains some similar self-experiments of this kind that demonstrate inner speech’s dependence on the motor system.

11 “Comment se fait-il, qu’en opposition avec les souvenirs de toutes nos autres sensations, il soit toujours clair, exact, précis et déterminé, et surtout, qu’il soit, autant à notre disposition que la parole extérieure, effet de l’empire absolu que nous exerçons sur nos organes locomoteurs?” (sec 353 de Cardaillac, 1830) How is it that contrary to memories of all our other senses, it [inner speech] is always clear, exact, precise and determined and, above all, that it is always as much at our disposition as external speech, an effect of the absolute control we exercise over our organs of movement?
perhaps, Watson’s claim that verbal thoughts were generated by micro-movements of the speech articulators (Watson, 1913, 1914, 1920). The implication of this idea was that thought would be impossible if the speech articulators were paralyzed. Smith et al. (1947) disproved this — the lead author paralyzed his own articulators with curare and reported that he could still think clearly even though his articulators could not move (to the point that he needed intubation to allow him to breathe).

Despite the fact that actual execution of movements did not seem necessary (as Smith et al. had shown), the idea that inner speech was a motor act was standardly assumed within behaviourism and there was a great deal of research done demonstrating that inner speech is typically accompanied by subtle movements of the speech articulators.

Rounds and Poffenberger (1931) used a face-mask to measure breathing patterns and found similar patterns in inner and external speech (though they did not perform statistical analyses). Jacobson (1931) used electrical sensors to detect muscular activity in the speech articulators and found similar patterns of activation in both covert and overt speech. Locke and Fehr (1970) performed a similar experiment using electromyography (EMG) and found more lip activity when participants were asked to memorize lists of words with labial elements than in lists without labials. These are just a few examples from a vast literature of similar experiments.

McGuigan (1970, 1978) reviews a large number of such studies from the 1890’s to the 1970’s. He reports on a variety of measurement techniques (from mechanical sensing devices to electrical sensing of muscle activity) showing that small movements of the speech articulators can be detected when people think verbally or read. McGuigan himself contributed to this literature, using EMG to measure muscle activity in the lips and tongue as participants silently thought of syllables starting with either labial or alveolar consonants; the EMG showed increased activ-

---

12 This is in line with the behaviourists approach to other forms of imagery — visual imagery was seen as a reconstruction of a visual scene via movements of the eyes.

13 Dodge performed a similar experiment in 1896 — anaesthetizing his articulators with a 20% solution of cocaine and finding no change to his experience of inner speech; thus showing that inner speech is not dependent on tactile feedback from the articulators (Dodge 1896; cited in Jacobson 1931).
ity in the articulators that matched the imagined syllables (McGuigan et al., 1982).

In the Soviet Union, Sokolov was pursuing similar work, amassing a significant collection of data using both mechanical movement-sensors and EMG. All of this data (surveyed in Sokolov 1972) supports the claim that inner speech is often accompanied by minor movements of the speech articulators, implicating the motor system in the generation of inner speech.

The evidence cited above is strong support for the claim that inner speech originates in the motor system because it demonstrates that movements of the articulators can actually be detected when people are engaged in inner speech; however I would like to make it clear that the theory that I am proposing is not dependent on any overt execution of motor actions. Corollary discharge can presumably be generated without triggering any overt movement of the muscles involved.

Modern brain-imaging techniques have confirmed the evidence from the earlier EMG and mechanical studies. Wildgruber et al. (1996) found, using fMRI, that primary motor areas and Broca’s area were activated by silent speech. Friedman et al. (1998) produced very similar results. Bacić et al. (1999) found that the areas of activation in overt and covert (inner) speech were largely the same, and that these areas included the premotor areas (again, including Broca’s area). Bullmore et al. (2000)’s results agreed with the other fMRI studies in finding motor-area activation (including Broca’s area) in inner speech. These results were also confirmed by Shergill et al. (2001).

Brain areas can be temporarily incapacitated by repeated application of TMS. When this is done to Broca’s area or more primary motor areas, participants report a disruption of their ability to engage in inner speech — the same disruption is found if they are asked to speak aloud (Aziz-Zadeh et al., 2005).

The motor basis of inner speech has even been turned to advantage by NASA, who are developing a communication system for noisy environments which allows micro-movements of the speaker’s articulators to be picked up by sensors and translated into speech by computer software (Braukus and Bluck 2004).
1.5 Corollary Discharge

The central proposal of this dissertation is that the sensory content of speech imagery is constituted by corollary discharge which is a component of the motor-sensory loop.

I should make it clear that while I argue that corollary discharge has been ‘co-opted’ in an evolutionary sense to provide the sensory content of speech imagery in humans, I am not at all arguing that the presence of corollary discharge in an animal implies that the animal has some form of imagery. There is overwhelming evidence that the motor systems of such animals as crickets and tadpoles rely on the function of corollary discharge (as discussed below), but I would certainly not argue that these simple organisms have anything like mental imagery.

There are two related but distinct functions that are associated with the term corollary discharge:

1. Providing the motor system with faster sensory feedback than could be delivered by the sensory systems (discussed as part of the explanation of forward models in subsection 1.5.1).

2. Distinguishing self-caused from externally caused sensations (discussed in subsection 1.5.4).

Both of these functions require information about what the sensory consequences of an action will be, and thus are based on the concept of a forward model, which is a system that takes a motor command as input and provides a prediction of sensory consequences as output. I will discuss both of these corollary-discharge functions below, as they are both relevant to an understanding of forward models; though I should emphasize that the theory I am proposing identifies the sensory content of inner speech only with the second of these functions, and only for the auditory modality. The motor system must, of course, predict sensory consequences not just for hearing, but also for sight, kinaesthesis etc.; however this dissertation is only dealing with the claim that auditory corollary discharge has been ‘co-opted’ in humans to provide the sensory content of inner speech.
1.5.1 **Forward Models**

There are two basic approaches to controlling a system: open-loop and closed-loop. Open-loop control is exemplified by a toaster — once you have set the duration of the toasting (longer for darker toast), the machine follows a set routine and does not take into account any sort of feedback about how hot the elements are getting or how dark the toast has become. In fact, there is no *feedback* of any kind to alter the performance of the machine. It is this absence of feedback that makes a system open-loop (see Figure 1.1). Open-loop systems follow a pre-set behaviour no matter what happens. They do not receive feedback, so are incapable of altering their behaviour when circumstances change.

Closed-loop systems are those which use some sort of feedback to alter their performance (and thus *close the loop*). The canonical example of a closed-loop system is a thermostat which measures the temperature of a room and uses that feedback to alter its behaviour (see Figure 1.2) — turning the furnace on when the heat falls below a set level and turning the furnace off when the temperature rises above a set level.

![Figure 1.1: An Open-Loop Model (i.e., no Feedback)](image)

Closed-loop control allows for much more elaborate and context-dependent behaviour and so is ubiquitous in biological control systems (such as our motor system). There is a problem facing a closed-loop system, though, and that is the time delay between initiating a behaviour and receiving the feedback. We have all
experienced the frustrations that this delay can cause — think of the difficulties of getting the shower temperature just right when there is a long lag between turning the tap and feeling a difference in the water temperature; we often fall into an oscillating pattern of over-compensating one way and then the other.

This sort of time delay is an unavoidable problem in most control systems and is a particular problem in the motor systems of animals. In addition to whatever time it takes an animal to decide on an action to perform, there are two additional sources of delay in biological motor control. First, muscles cannot respond to a command instantaneously; they have inertia and so there is a delay between the motor command and the response of the body. A second source of delay is sensory. Our senses do not operate instantaneously; it takes time for a change in the environment (or in our body) to be transduced by our end-organs, then transmitted to and processed by the central nervous system (CNS) and for a motor correction on the basis of this information to be issued. This delay can be quite considerable; in the case of vision it is typically on the order of 100-130 ms, for proprioception

---

14Technically, this source of delay is called a lag in the motor control literature since the inertia of muscles will likely alter the shape of the motor command (e.g., a punctate command to move a body-part will be smoothed into a gradual acceleration of the body-part). A delay that does not cause a change in shape of the commanded action is simply called a time-delay. While this distinction is important for control-system engineering, it is not relevant to this dissertation, so will be ignored from here on.
it is estimated at around 70-100 ms\textsuperscript{15} (Desmurget and Grafton, 2003), for speech perception it has been estimated at around 130 ms (Jones and Munhall, 2002). This means that feedback is not available (or minimally available) for many actions and we are back to an open-loop (and thus much less sensitive) form of action control.

A commonly accepted solution to the problem of delayed feedback is to use a forward model.

A forward model is an internal model (internal to the CNS) of consequences that will occur when an action is performed. Given a starting state and an action, a forward model predicts what the effects will be. A very simple example of a forward model is the instruction manual for a car. The manual will tell you that if the car is in a particular state (e.g., first gear), then pressing on the accelerator will have a certain effect (move you forward); however, if the car is in a different state (e.g., reverse), then pressing on the accelerator will have a different effect (move you backward). The manual constitutes a simple forward model of the car. It allows you to input a current state (the gear) and a planned action (pressing the accelerator) and it outputs the result (whether the motion is forward or backward).

<table>
<thead>
<tr>
<th>Initial State</th>
<th>Command</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>first gear</td>
<td>push accelerator</td>
<td>move forward</td>
</tr>
<tr>
<td>reverse gear</td>
<td>push accelerator</td>
<td>move backward</td>
</tr>
</tbody>
</table>

This simple kind of forward model is called a \textit{look-up table}. With thousands of motor units to control and with the effects of activating any one motor unit being dependent on the state of other units, such a look-up table faces a combinatorial explosion in which the number of entries in the table would exceed the number of particles in the universe (Wolpert et al., 2001). So, in the CNS, the forward model could not actually be implemented as a simple look-up table and would have to be a more generative process.

One benefit of a forward model is that it allows the sensorimotor system to

\textsuperscript{15}This estimate for proprioceptive feedback is for centrally-controlled alterations of an action based on proprioceptive feedback, rather than spinal reflex-adjustments. This is the distinction between M3 (central), and M1 and M2 (spinal) responses to proprioception (Desmurget and Grafton, 2003, p.300)
check’ what the consequences of an action will be before performing it. In the case of forward models in moving animals, if the forward model receives information about what actions are about to be performed by the animal, it can generate a prediction of what the consequences are going to be and this prediction can be available to guide action before actual sensory feedback is available. In this way, the time-lag problem inherent in sensory feedback can be avoided. The system does not use actual feedback, but uses the predicted feedback generated by the forward model. This allows for a much faster detection of errors and the system becomes almost closed loop (sometimes called pseudo closed-loop).

The discussion above hinges around the difference between predicted sensory feedback and actual sensory feedback. Another term for actual sensory feedback is reafference. Reafference is simply a term of convenience for stimuli in the environment that were produced by the person perceiving them. So, for example, if you speak to me we both hear the sound of your voice, but this sound is afference for me, and reafference for you, because you produced it.

The signal that is sent from the motor system to the forward model, informing the forward model about an action that is about to happen, is called efference copy. Efference copy is thus a motor signal — a ‘copy’ of the motor commands that are about to be executed. The prediction of sensory consequences that is output by the forward model is called corollary discharge and is, by definition, a sensory signal (a prediction of upcoming sensations). So, efference copy is the motor-signal input to the forward model and corollary discharge is the sensory-signal output of the forward model.

Because efference copy and corollary discharge are so closely related, and are parts of a functional whole, the two terms are often used interchangeably, though there is an important distinction to be drawn.

The relationship between efference copy, forward models, and corollary discharge is represented in Figure 1.3.

The visualization in Figure 1.3 is, of course, very simplified. There are multiple modalities of corollary discharge (tactile, kinaesthetic, visual, auditory, vestibular) generated by the forward model and it is assumed that there are also multiple forward models working in conjunction and at multiple levels (Kawato 1999; Wolpert 1997; Wolpert and Kawato, 1998).
Of course, predicted feedback is only useful if the prediction is very accurate. The forward model is helped to make good predictions by being corrected by error-signals. Actual sensory feedback (reafference) is compared with the sensory prediction (corollary discharge); if the reafference does not match the corollary discharge, then the forward model can be updated to account for the mismatch and hopefully make a better prediction next time.

Forward models were initially developed in the field of engineering control to solve computational problems; however it quickly became apparent that the same problems were faced in biological control and experimental evidence has since been found for forward models in biological systems. One form of experimental evidence for the existence of forward models in biological systems is found in measurements of grip force. When people are asked to move an object with which they are unfamiliar, their grip force changes reactively. That is, the changes are based on sensory feedback about whether the object has started to slip in their hands (and so more grip force is required). However, when moving a familiar object, where prediction of the needed forces can be provided through a forward model, grip
forces change simultaneously with movements, showing that people are not reacting to reafference about their movements, but are anticipating the requirements of their movements. This is strong evidence for the use of a forward model (Kawato, 1999).  

In the speech domain, there is also experimental evidence supporting the existence of forward models. Motley et al. (1982) showed that experimentally induced spoonerisms that were likely to induce people to produce a taboo word accidentally were often caught before they were actually produced out loud, however an increase in galvanic skin-response (a simple measure of arousal) showed that these unspoken ‘dirty’ words had been detected — strong evidence for an internal monitoring system that monitors the output of a forward model.

Further evidence for forward models in the speech domain is adduced from experiments showing that we compensate very quickly in our speech production (within 135ms) for perturbations of our auditory feedback. For example, when the pitch of our voice is artificially shifted up or down, we quickly shift our pitch in the opposite direction to compensate (Jones and Munhall, 2002). A similar compensation happens for artificial changes in formant values (Houde and Jordan, 1998; Tourville et al., 2008). Without a forward model to provide a point of comparison, such immediate compensation would be hard to explain. Similarly, our ability to continue speaking accurately when normal auditory feedback has been blocked (Pittman and Wiley, 2001) is evidence of a forward model supplying substitute feedback to guide production. In the case of people who have lost hearing in adulthood, their speech production can continue to be quite normal for a significant length of time in the absence of auditory feedback (Guenther and Perkell, 2007), suggesting that other forms of feedback are substituting for the missing auditory

---

16In a perfect world, with no errors and no noise, inverse and forward models would be perfect inverses of each other (an inverse model gives the motor command needed to produce a particular output and a forward model gives the output of a particular motor command — for further discussion, see subsection 1.5.3). This means that it is often difficult to tease apart whether it is an inverse or forward model that is responsible for an observed effect. For example, if the inverse model is good enough, it could predict the motor commands (including the necessary changes in grip force) needed to manipulate a familiar object, and so the grip-force evidence described above could be the result of an inverse, not a forward, model. However, the fact that predictive changes in grip force can be shown to occur when subjects are still in the process of learning how to manipulate an unfamiliar object (and so before an accurate enough inverse model could have developed) demonstrates that this predictive grip force is not the result of an inverse model (Flanagan et al., 2003).
information.

Forward models are a common component of motor-control theories generally (Desmurget and Grafton 2000) and have often been invoked in theories of speech motor-control specifically (e.g., Gracco, 1995; Honda, 1996; Kent, 1997). Perhaps the best elaborated model of speech production that incorporates forward models is Guenther et al.’s 2006 DIVA model (Directions into Velocities of Articulators), which not only models the computational structure of speech motor-control, but also tries to map these computational functions onto appropriate neural regions. A similar proposal has been made by Hickok et al. (2011) who provide a detailed model of speech production that relies heavily on forward models and is, incidentally, a good review of much of the material discussed in this chapter.

1.5.2 Inverse Models

The flip side of the forward model is the inverse model. While a forward model takes a command as input and outputs the predicted consequences, an inverse model takes a desired consequence as input and outputs the command necessary to achieve it. It can be difficult to tease apart which aspects of motor behaviour are due to a forward rather than an inverse model, since computationally they are often equivalent: If an inverse model is accurate enough, motor commands will behave as if they were following predictive forward models.

One computational problem in motor control that seems to require inverse models is error-coding. Errors in speech are picked up in the auditory domain but must be corrected in the motor domain and this must happen very quickly. This argues for a very quick mapping between auditory and motor codings of speech. Inverse models provide that mapping, by taking a sensation and outputting the motor command which would generate it. This is certainly not unique to speech perception. The motor system generates effects in multiple sensory modalities and for all of these modalities (auditory, tactile, visual, vestibular) an error in sensory coding must be quickly corrected in a motor coding.

As inverse models map from sensation to motor command, they are obviously a central component of imitation (Iacoboni 2005). This point is picked up in section 5.1.
Inverse models are an important component of motor-control theory. However, inverse models do not form any part of the theory proposed in this dissertation, and none of the experiments reported in subsequent chapters test any aspect of inverse model function, so I will not go into any further detail on inverse models.

1.5.3 The Two Functions of Corollary Discharge

Forward models convert a copy of motor commands (efference copy) into a prediction of sensory consequences (corollary discharge). As mentioned above, the two functions of corollary discharge are:

- To guide actions with feedback at a shorter latency than can be provided by genuine sensory feedback.
- To allow the sensory system to segregate self-caused sensations from externally-caused sensations.

These two functions must occur at different times. Sensations are time-varying signals and in order to ‘filter out’ the self-caused component of such signals, corollary discharge must match up with reafference temporally (at least at some point in the processing chain). However, the only value of the feedback function of corollary discharge is that it is available before feedback through normal sensory channels. So, while both of these functions are termed corollary discharge, they must take place at different times.

This dissertation proposes that the second function of corollary discharge (the sensory filtering function) constitutes the sensory content of speech imagery. This filtering function is discussed in subsection 1.5.4 below.

Forward models were first hypothesized because researchers realized the need for short-latency feedback. It is for this reason that this feedback function of corollary discharge is discussed in detail above. This feedback function is connected to issues of speech-monitoring and so is important to research on stuttering and speech errors. However, the theory proposed in this dissertation is that the sensory content of inner speech is provided by the filtering function of corollary discharge, so I will now turn to examine that aspect of corollary discharge. The short-latency

---

17 For a discussion of the perceptual effects of this temporal difference, see subsection 5.2.4.
feedback aspect of corollary discharge does not form part of the theory proposed in this dissertation, so is not discussed further.

1.5.4 Distinguishing Self-Produced from External Sensations

Imagine the life of a tadpole. It is in constant danger of being eaten and so has evolved instincts to protect itself, such as the avoidance reflex: Whenever it is touched it immediately moves away from the source of the touch. This creates a problem, though. If the tadpole registers a touch on its right side, it will arc its body to the left. However, doing so will compress its left side against the surrounding water, triggering the avoidance reflex on its left side. So, the tadpole will arc its body to the right, which will compress the right side of its body against the surrounding water and so trigger the avoidance reflex on its right side. So, the tadpole will arc its body to the left... this obviously leads to a vicious circle. The way out of this is for the tadpole to suppress the ability of sensory stimulation (on the side in the direction of movement) to trigger the avoidance reflex, which is exactly what it does (Sillar and Roberts 1988). A similar mechanism is found in the nematode C. Elegans and the neural circuitry for it has been mapped out in detail (Chalfie et al., 1985).

This suppression of stimulation (and hence of the avoidance reflex in the direction of movement) requires that the motor system inform the sensory system of its actions — not just that a movement is about to happen, but the details of the movement (in this case, which direction). The signal that carries this information from motor to sensory areas is corollary discharge.

The original formulation of the idea of corollary discharge is usually attributed to Helmholtz (1866, cited in Bridgeman 2007). Helmholtz argued that the reason we do not experience the world as spinning when our eyes dart back and forth (or 'saccade'), is that there is a signal sent from the motor system that alerts the visual areas about the upcoming movement and so the visual flow caused by the movement is not misperceived as a spinning world. However, Grüsser (1986, 1995) has reviewed the history of the corollary discharge concept and found that a similar idea was well established among the philosophers of ancient Greece (e.g., Plato, Chrysippos, and Empedocles of Akragas) and among the Arab philosophers who
followed them (e.g., Avicenna, Alhazen). In 1613, Franciscus Aguilonius gave a remarkably full and detailed account of the concept of corollary discharge that is largely identical with the modern concept, but his work was quickly forgotten. In the modern era, the idea was again rediscovered and popularized independently by Sperry (1950) and von Holst and Mittelstaedt (1950) who used the terms ‘corollary discharge’ and ‘efference copy’ respectively for this motor-to-sensory signal.

Von Holst and Mittelstaedt rotated the head of a blowfly 180 degrees. The effect of this manipulation was that the fly fell into a constant circling pattern, seemingly unable to break out of a behavioural loop. This circling, however, did not occur when the fly was in the dark. Von Holst and Mittelstaedt argued that this behaviour was caused by a discrepancy between the fly’s visual feedback and its ‘efference copy’;\textsuperscript{18} The idea is that the fly is comparing the visual feedback it gets with a prediction of the feedback. When there is a discrepancy, it makes a corrective movement. However, when the visual information is inverted (because its head is upside down), the fly misperceives the direction of the discrepancy between prediction and visual feedback (left is perceived as right), and so the corrective motor-command is in the wrong direction which only makes the discrepancy between prediction and sensation worse, leading to another corrective motor command, making the error worse . . . leading to a vicious circle, both figuratively and literally.

Sperry published his paper in the same year that von Holst and Mittelstaedt published theirs, and it was very similar in structure. Instead of the head of a fly, Sperry inverted the eye of a fish. A similar pattern of circling behaviour (again, except in the dark) led Sperry to the same conclusion, though he termed the internal comparison signal corollary discharge. The two terms (efference copy and corollary discharge) are now used for different components of the forward model system (as discussed in subsection 1.5.1).

1.5.5 Sensory Attenuation

The primary diagnostic effect of corollary discharge (in its filtering function) is that it attenuates an organism’s response to self-produced sensations. This sensory

\textsuperscript{18}This was the first use of the term efference copy. Today, the signal compared with reafference would normally be termed corollary discharge.
dampening is seen in several sensory modalities, in behavioural and brain imaging measures, and in humans as well as other animals (indeed throughout the animal kingdom). A good review of the sensory attenuation literature is found in Cullen (2004). The following section reviews some of the evidence for sensory attenuation found in the different modalities using both behavioural and brain imaging measures.

**Touch**

There is a simple demonstration that can make you aware of the effects of corollary discharge in yourself. Compare the sensation of running your hand over an object with the sensation of keeping your hand still while a friend moves the same object under your hand, these two experiences will feel different even though the pattern of stimulation on the skin of your hand is the same. The difference is that when you are the one producing the movement corollary discharge is being generated and so the sensory signal coming from your hands is being modified.

Touch is perhaps the most thoroughly researched modality for sensory attenuation in humans. Weiskrantz et al. (1971) demonstrated that self-produced touches are perceived as less ticklish. Blakemore et al. (1999) extended this result, finding that though people rated self-produced touches as not ticklish, but when a time delay was introduced between their action and their receiving the touch, the sensations became more ticklish. Blakemore et al. argue that this is due to corollary discharge which attenuates the effects of touch when action and sensation are in synch but fails to attenuate the sensation (thus making it more ticklish) when a delay causes action and sensation to be out of synch. The case of tickle is discussed further below.

Brain imaging studies have been done to corroborate the behavioural experiments on tickle. Blakemore et al. (1998) found an attenuated response to touch in

---

19 It is interesting to note that, as with so much in biology, Darwin and Aristotle were there first. Darwin noticed that self-produced touch was not ticklish and proposed a role for knowledge in the explanation (Darwin, 1872). Darwin was himself ‘scooped’ by Aristotle who made a very similar point about the role of knowledge in attenuating tickle: “Why can no one tickle himself? Is it for the same reason that one feels another’s tickling less if one anticipates it, and more if one does not see it coming? So that one will be least ticklish when one is aware that it is happening.” (Aristotle)

20 By means of a robotic hand that delivered a touch either in synch or at a delay from the participants’ movements.
the somatosensory cortex when the touch was self-caused.

Shergill et al. (2003a) showed that when people are asked to push back with the same force with which they are pushed, they consistently apply more force, suggesting that they are underestimating the degree of force they are applying (because of sensory attenuation). Shergill et al. suggest that this may be the origin of escalating shoving-wars, where each person thinks (incorrectly) that they are shoving back with the same force with which they were shoved. When people are asked to equate forces via an intermediary (like a joystick), the sensory attenuation does not occur and so people are far more accurate in their estimation of self-produced forces (ibid.).

Voss et al. (2006) showed that sensory attenuation of touch stimuli could occur even in the absence of overt movement. By using TMS over primary motor areas, they delayed participants’ execution of a finger movement. This procedure, however, did not prevent the motor command from being issued and, in line with the predictions of efference copy and corollary discharge, they found that a touch delivered to the finger when the movement would have occurred was still attenuated as if the movement had actually occurred.

Vision

Some of the earliest evidence for corollary discharge was from the visual modality. Indeed, as discussed in section 1.5, Helmholtz’s original reason for proposing corollary discharge was to explain why the world does not seem to spin when we move our eyes, even though the visual flow induced by such a movement should be indistinguishable from a spinning world. von Holst and Mittelstaedt as well as Sperry did their seminal work on corollary-discharge mechanisms in vision.

Schafer and Marcus (1973) performed an electroencephalography (EEG) study showing that brain responses to flashes of light caused by the participant pressing a button were attenuated in comparison with responses to flashes presented at random.

There has been a lot of work done in the visual domain; I will just mention two more examples: saccadic suppression and blink suppression.

Saccadic suppression — Despite the fact that our visual experience seems
smooth and flowing, our eyes are in fact constantly darting back, making short ‘jumps’ (= saccades) several times per second. Our eyes are moving quite rapidly during these saccades, however we do not experience a rapidly shifting visual field (as would happen if our bodies moved through space at the speed of our saccades). Sylvest er et al. (2005) show that corollary discharge, sent from motor areas controlling the eye to visual perceptual areas, suppresses our experience of this distracting visual flow.

Blink suppression — A similar phenomenon occurs during blinking; we do not notice the sudden darkening of our visual experience when we blink our eyes, and studies have shown that visual acuity is lower for a short period before a blink actually occurs, suggesting an active mechanism of sensory attenuation (Bristow, 2006).

Vestibular sense

The vestibular sense (the sense which tells you up from down and whether you are accelerating) also relies on a corollary-discharge signal. When we turn our heads, our vestibular organs send information about the movement to the CNS, and it is crucial that a voluntary head movement is not mistaken for a jarring of the head by an external event. Roy and Cullen (2001) recorded individual neurons in monkeys and found that brain nuclei that encode vestibular information do not respond to self-generated head movements. Roy and Cullen (2004) found similar results.

I would argue that the attenuation of vestibular signals by corollary discharge is behind the fact that few people get motion sick when they are driving even though their passengers, experiencing the same accelerations, often do.

Electroreception

Electroreception is the sensing of electrical fields either passively (picking up on fields generated by other living creatures) or actively (by generating an electrical field and detecting the distortions in the field caused by the conductivity of surrounding matter — analogous to a bat’s sonar). While this is not a sensory modality found in humans, it is quite common among vertebrates: many taxa of fish have it, as do some amphibian species and it is even found in one subclass of mammals
Electroreception is a useful animal model for examining corollary discharge since single-cell recordings can be simultaneously made of both the cells generating the electrical field and those detecting it; allowing for a fine-grained analysis of the correlation between action and perception. Bell (1981) did just this (see also Bell (2001)). He used curare to paralyse the cells which generate the electrical field in mormyrid fish, preventing the cells from discharging. However, since the command for an electrical discharge was still issued by the fish’s motor system, efference copy and corollary discharge were still generated. By measuring the activity of the electrical-field sensor cells in the same fish, they found that these cells responded as if they expected an electrical discharge; furthermore, these cells were attempting to counteract the expected discharge — firing negatively in anticipation of a positive discharge and firing positively in anticipation of a negative discharge. This highlights the sensory attenuation function of corollary discharge, which is often described as a photographic negative of sensation which combines with the reafferent sensation to cancel it out.

Hearing

I have just reviewed the evidence for corollary discharge in other modalities, but the claim of this dissertation is that it is the auditory modality of corollary discharge that has been co-opted to provide the sensory content of speech imagery. Thus, for the purposes of this dissertation it is the auditory modality, and specifically audition for speech, that is of greatest importance.

There is a large body of evidence, both behavioural and brain imaging, that self-produced sounds show sensory attenuation. When the sounds are vocalizations (which, of course, is the normal case), the attenuation is called speaking-induced suppression, though attenuation can also occur to arbitrary sounds that are paired with actions, e.g., attenuation to a click (Schafer and Marcus, 1973) or to a tone (Bäss et al., 2008) caused by pressing a button.

Unlike tickle for the touch modality, it can be difficult to find a behavioural

---

21 Monotreme mammals are the egg-laying group of mammals, e.g., platypus and echidna.
22 In general, throughout the rest of the dissertation, “corollary discharge” will refer to “auditory corollary-discharge”.
measure of sensory attenuation in the auditory modality (but see chapter 4), so the majority of the studies described below are brain imaging studies (using a variety of techniques).

There are a number of animal models available for studying corollary discharge in the auditory modality. Using insects has even allowed the mapping of components of the neural pathways involved. For example, Poulet and Hedwig (2002) found that crickets are able to prevent being temporarily deafened by their own singing (which is VERY loud - c. 100 dB) by the use of corollary discharge. This is of high evolutionary importance to the (male) cricket, since it needs to sing to attract a mate, but producing such high intensity sounds could deafen it and make it vulnerable. By doing single-cell recordings on singing crickets, Poulet and Hedwig were able to identify the specific neurons where the corollary discharge attenuation was effected (Omega 1 interneurons). The corollary discharge was even found to occur during ‘silent singing’ (where the motor command is generated by the cricket but, since a forewing has been removed, no sound is produced).

One mammal with an extraordinary reliance on the interaction of vocalizations and hearing is the bat, which uses the echoes from its vocalizations to detect prey and navigate in the dark. It is therefore not surprising that bats rely on corollary discharge to help distinguish a call from its echo. Suga et al. (1972) found that the response of neurons in the lateral lemniscus was attenuated by about 25 dB when the bats vocalized in comparison to hearing a recording of those same vocalizations. Suga and Shimozawa (1974) found a similar attenuation in lateral lemniscus neurons (though only 15 dB this time) and found that the timing of the attenuation was closely synchronized with the bat’s vocalizations.

Evidence from non-human primate models can be extended more confidently to models of human behaviour. There is a great deal of brain imaging work done on sensory attenuation of vocalizations in primates. Eliades and Wang (2004) performed single-unit recordings from marmoset auditory cortices. They found that recorded units suppressed their firing by about 71% when the animal was vocalizing and that the suppression preceded the onset of vocalization by about 220ms. They found (as did Eliades and Wang 2008) that

\[23\text{N.B. A cricket sings by rubbing its forewings together. A related point is that its ears are on its forelegs so they are completely exposed to the sound of the singing.}\]
some units showed enhancement while others showed suppression.

In addition to the many animal studies discussed above, there is a substantial brain-imaging literature on auditory sensory-attenuation in humans.

Several brain imaging studies have demonstrated that the auditory cortex is less responsive during vocalizations. One of the earliest was Schafer and Marcus (1973) who showed that EEG responses to clicks that were self-initiated (by a button press) were lower than randomly presented clicks.

Most of the brain-imaging studies have been fairly indirect because the non-invasive fMRI and EEG methods only show patterns of activity for large assemblies of neurons. However, Greenlee et al. (2011) was able to provide more detailed information by implanting intracranial electrodes in 10 patients about to undergo brain-surgery for intractable epilepsy. They found that auditory cortex response was attenuated when a patient spoke a word in comparison to when the patient heard a recording of their own voice speaking the same word.

As mentioned above, the attenuation of sounds can even be found in cases where the sound-action mapping is arbitrary and newly learned. Aliu et al. (2009) had participants repeatedly press a button to trigger a tone. When there was no delay between button-press and tone, MEG results showed a left-hemisphere reduction in response to the tone. This reduced response did not occur when a 300 ms delay was introduced between button press and tone. However, when participants were instead trained with a button-press that triggered a tone 300 ms later, the response-reduction was present in both hemispheres and did occur when the 300 ms delay was removed. The authors argue that this left-hemisphere preference for zero-delay sounds is due to the left hemisphere’s specialization for speech (which has an essentially zero-delay between motor execution and sensory result — though, crucially, not a zero-delay between the motor command and sensory result).

Numminen et al. (1999) found a similar attenuation to tones using magnetoencephalography (MEG). In this study, participants either read silently or aloud while being presented with tones — auditory cortex responses were attenuated by 44-71% in the reading-aloud condition.

Bäss et al. (2008) also measured auditory cortex response to tones, examining the effect of predictability on attenuation. In this study, participants generated tones
via a button-press. The effects of tone predictability (in pitch and onset time) were examined using EEG (examining N1). The greatest N1 suppression was found when tone pitch and onset were most predictable. Martikainen et al. (2005) performed a very similar study using MEG and again showed that there is a reduced response in the auditory cortex to tones initiated by the participant (via a button press).

Greater Suppression for Closer Matches between Sensation and Corollary Discharge

The degree of sensory attenuation seems to be dependent on the degree of match between the corollary discharge signal and the incoming sensory signal (reafference). The greater the match the greater the attenuation. For example, in an MEG study, Houde et al. (2002) found a reduction in the M100 auditory cortex response equivalent to a 13 dB drop in sensory signal intensity in the left hemisphere and equivalent to a 7 dB reduction in the right, when comparing self-produced sounds to playback of those sounds. When participants spoke while hearing tones, the reduction in M100 was equivalent to only about a 3 dB reduction in signal intensity, suggesting that there is a general reduction in responsiveness but the reduction is less severe as the match between incoming sensory signal and corollary discharge signal becomes weaker. In a third experiment, Houde et al. (2002) eliminated the attenuation completely by masking the sound of a speaker’s voice with white noise (a case of extreme divergence between sensory and corollary discharge signals).

This match-dependent level of sensory suppression was also shown by Heinks-Maldonado et al. (2005) who used EEG to measure the N100m response of the auditory cortex to self-produced vocalizations (repetitions of the vowel /a/). The feedback that participants heard was altered via pitch shifting or substituting an alien voice. The degree of N100m attenuation was greatest when there was no alteration of the feedback. Heinks-Maldonado et al. (2006) performed a similar analysis using MEG and also found a greater sensory attenuation when auditory feedback of speech matched the production.
My proposals for alternative interpretations of sensory attenuation

The sensory attenuation of corollary discharge is usually explained as a simple lowering of the reafference signal intensity. This is often described as analogous to laying a photographic negative over its positive. While such a cancellation explanation is certainly possible, I would like to propose two other interpretations.

First, the inhibited brain responses found for self-caused sensations may not be simple attenuation of the sensation, but may be due to the fact that corollary discharge provides information to sensory areas allowing these areas to ‘skip’ much of the usual processing. The sensory systems are being told what the incoming signal is, so they do not need to work as hard to come to their conclusions.

A useful analogy for this idea is to view the body as a multinational corporation that ships material back and forth between various warehouses and factories. The corporation will be much more efficient if, when a shipment is made to a warehouse, an invoice is sent at the same time which informs the warehouse what is going to arrive and when. If a shipment arrives at the warehouse without any warning there would be a lot of panicked last-minute preparation, which is less efficient. This is just good organization; and it is what the motor system does in our bodies. Corollary discharge is like the invoice sent to warn the warehouse of an arriving shipment. The attenuated brain response under corollary discharge may be due to a more efficient categorization of sensations. In perception, external events are, in a way, like processing a shipment without warning: all the processing has to be done when the sensation arrives. However, with corollary discharge, much of the processing is already done by ‘higher’ areas and so the sensory processing system has less work to do. In essence, the sensory system can simply give sensations that match corollary discharge less processing, since the system already ‘knows’ what is contained in the signal.

A second alternative to the idea that sensations are simply canceled out by corollary discharge is that they are channelled into different processing streams, and so interfere with external sensory-streams less. This is similar to the idea of auditory scene analysis (Bregman, 1990), in which the confusing mess of sound fragments that arrives at our ears from multiple sources is segregated and grouped into streams so that sound components that were caused by the same event are treated
as belonging together (in the same stream) and are distinguished from sound components caused by different events. It is through such auditory scene analysis that we are able to follow a conversation in a noisy environment. I am suggesting that corollary discharge would simply provide the auditory system with information that allows it to channel self-produced sounds into their own stream, thus segregating them from the streams for externally-caused sounds. This would isolate the effects of self-produced sounds from external sounds since the perceptual impact of sounds on each other is limited if they belong to different sensory streams (Bregman, 1990, p.514). Thus, the lowered impact of self-produced sounds would not be because the sound was somehow attenuated, just segregated.

For example, in the work on tickle (Blakemore et al., 1999) the reduced ticklishness of self-touch is described as due to a lessening of the intensity of the touch. While this is a useful analogy, it is probably not the whole truth. In the case of tickle, if lessening of intensity were the explanation, that would imply that more force = more ticklish, which seems obviously false. In fact just the reverse seems true — one tickles with a feather not a club. It does not seem possible to tickle oneself by simply ramping up the force used, so the lack of tickle-response seems to be about something more than intensity. What is essential is who is doing the action. It is the fact that it is someone else doing the tickling that makes something ticklish, which is in fact what Graziano (2009) has proposed in terms of the evolutionary origin of tickle:

“[…] tickle-evoked laughter evolved from play fighting in which a strong defensive reaction broadcasts that one animal has succeeded in penetrating the defenses of another animal and has contacted a vulnerable body part. The signal is not all-or-nothing. It is a graded signal, in which a stronger or more intense signal is evoked by a greater degree of violation of personal space” (Graziano, 2009, p.187)

If this argument about tickle is correct, then this suggests that the sensory attenuation we see in corollary discharge is, at least in part, due to attributing the source of a sensation to a different cause rather than a simple ‘lowering of intensity’.

---

24E.g., when two sounds that are normally dissonant are tricked into being processed as belonging to different sources, the sounds are no longer heard as dissonant.
I am arguing here that corollary discharge is analogous (and perhaps can be identified with) the top-down schemata of auditory scene analysis. There is also a strong similarity between my theory, Bregman’s schemata, and Ulric Neisser’s claim that imagery is constituted by perceptual expectations (also called schemata). Neisser argued that schemata are ‘top-down’ templates that are predictions of upcoming sensation and used to shape perception; he further claimed that these schemata can be triggered in the absence of external stimulation and then constitute imagery (Neisser, 1976, 1978). This is discussed further in subsection 1.5.7.

Of course, these three explanations of sensory attenuation (cancellation, lightened processing-load, streaming) are not mutually exclusive. Perhaps all three processes contribute in some degree to the overall effect of sensory attenuation.

1.5.6 How Sensory is Corollary Discharge?

I have regularly referred to corollary discharge as a sensory signal. In what sense is it sensory?

Corollary discharge is considered to be sensory because of its function. Corollary discharge anticipates incoming sensations and so must carry the same kind of information as carried by the coding of incoming sensations. Since we consider an incoming sensation as sensory (by definition), corollary discharge is defined as sensory by extension. However, we should be careful about viewing corollary discharge as being like “sound in the brain” or “light in the brain”. It is instantiated by neural coding and thus is an extremely abstract signal that is possibly quite different from the raw detail that is transduced at the cochlea (or at the retina for vision). Thus, though corollary discharge is sensory, it is quite abstract.

At some level of abstraction, this sensory content may be more accurately referred to as perceptual. Where the dividing line occurs is a matter of debate and is not at issue in this dissertation. Throughout this dissertation I will use “sensory” to refer to processing that occurs below the level of the category perceived. So, for example, all of the processing that leads to the perception of the phoneme /v/ will be referred to as sensory.

When I refer to auditory sensory content, I am leaving unspecified exactly what

---

25 Even at the level of transduction, the coding is already abstract.
this content is. Does it include a representation of formant frequencies? Burst frequencies? Fundamental frequency? My opinion is that it probably does implicitly encode such information, but not in a direct way. The encoding will be abstract and will probably conflate various raw acoustic components into higher-level properties. Trading relationships and context-dependence will mean that the encoding of auditory sensations will specify information that is based on the acoustics but not in a way that will allow a simple one-to-one mapping between the neural coding and aspects of the acoustics. This means that specifying exactly which components of the acoustics are present in the encoding will be difficult.

An example of this issue would be the following ‘toy’ neural coding scheme for conveying information about formant frequencies. Imagine that instead of conveying the actual formant frequencies, the auditory system encodes the difference between neighbouring formants.

This would mean that if a vowel with formants at 600 Hz 1700 Hz and 2600 Hz is presented to the ear, the three formants values would be coded as just two pieces of information: (1100, 900). Such a coding conflates information, but can be quite useful since the relative position of formants can be more important than their absolute value.

In this toy example, is the first formant frequency encoded in the neural signal? Well, yes, but not in a direct way.

The same issue may be true, in general terms, of the auditory representation of sounds in the brain — the coding can be said to encode acoustic details, but not in a simple or direct way.

I would like to emphasize that this is just a mock-up of the issues involved to demonstrate the problem. I am not claiming that the toy coding scheme described above reflects what actually happens in the human auditory system.

When I talk about the sensory information carried by corollary discharge, I do not mean that the representation of this information will correspond to our normal idea of sound — it will presumably be very abstract. The crucial point is that it will be abstract in the same way that normal sensory perception is abstract, and thus can safely be labelled “sensory”. This issue is discussed further in subsection 5.1.2.

It is well beyond the scope of this dissertation to determine exactly how abstract the coding of corollary discharge is. The answer to that question would require an
understanding of the neural coding of auditory sensations in general and that is a vast and ongoing research program.

1.5.7 Is Corollary Discharge Simply a Form of Attention?

Corollary discharge is an *anticipation* of sensory features. In future chapters, I argue that this anticipation influences perception by causing ambiguous aspects of an external stimulus to be perceived in line with the anticipation. This is not a new claim, the fact that people tend to see or hear what they expect to see or hear is a mainstay of psychological research. One of the more compelling examples of this is the “White Christmas” test (Merckelbach and Ven, 2001) in which some people can be induced to hear the song “White Christmas” when presented with white noise, simply by telling them that the song *might* be buried under the noise (the song is, in fact, not present at all). Anticipations influence perception and thus corollary discharge, as an anticipation, would be expected to influence perception. That is the claim tested in chapter 2.

This description of corollary discharge is essentially identical to a description of attentional effects on perception. This is not coincidental. The similarity between corollary discharge and attention is striking and I believe them to be related concepts. Both are functionally defined concepts and in many instances their functions overlap.

A similar point was made clearly by Ulric Neisser (Neisser, 1976, 1978) who developed a theory of visual imagery that is quite similar to the theory of speech imagery proposed in this dissertation. In Neisser’s theory, imagery consists of *schemata* which are structured anticipations. These anticipations are a functional aspect of normal perception where they structure the act of seeing, allowing the perceiver to pick out the relevant information in a visual scene. Thus, in normal perception, these schemata are functionally equivalent to attention. Neisser claims that these schemata can be divorced from their use in normal perception to serve the function of visual imagery. He also argues that since these schemata are sensory anticipations, imagery should influence perception, a claim with which I wholeheartedly agree and which the experiments reported in this dissertation sup-

---

26 This description could also be identified with Bregman’s *schemata* in his theory of auditory scene analysis.
Thus, under Neisser’s theory, schemata are just another description of the function of attention. I would make a similar statement about corollary discharge. The function of corollary discharge is to anticipate sensations, and thus to highlight certain aspects of incoming sensations for specialized processing. This much fits the functional definition of attention and so could equally be labelled attention. Thus, I would argue, one function of corollary discharge is equivalent to the function of attention.

However, attention and corollary discharge, while their functions overlap in this case, are not completely synonymous. In addition to anticipating sensations (which overlaps with the function of attention), corollary discharge is also defined as performing the function of attenuating an organism’s response to these sensations. This is not a function of attention. In fact, attention is usually considered to boost the impact of sensations.

So, while attention and corollary discharge are functionally equivalent in one respect (and so either label for this point of overlap would be appropriate and equivalent), they differ in another respect. The experiments reported in chapter 4 provide strong evidence for sensory attenuation and thus support the idea that imagery involves corollary discharge. For this reason, it is more consistent to use “corollary discharge” as the label for the totality of functions implicated in speech imagery; though it would not be an error to say that the anticipation aspect of corollary discharge is functionally equivalent to attention.

I should also explain the points of difference between the corollary-discharge theory that I am proposing and Neisser’s schemata theory. Schemata are structured sensory anticipations, but they are of a very general type. They can be of self-caused or externally-caused sensations and they are not tied to the motor system. Corollary discharge, on the other hand, is a very specific type of anticipation. It is an anticipation of self-caused sensations and it is generated by a forward model tied to the motor system. Thus, the claim I am making about speech imagery is more specific than that made by Neisser, though the two claims are not mutually exclusive.
1.6 Brain Imaging Evidence that Speech Imagery Involves Corollary Discharge

As discussed above (section 1.4), there is strong evidence that speech imagery originates in the motor system. This is hardly surprising, since external speech originates in the motor system as well. This dissertation argues that sensory content of speech imagery should be identified with a particular subcomponent of the motor system — the corollary discharge output of the forward model. This is not a particularly surprising claim. Speech imagery is, after all, an internal sensory signal that seems to be generated by the motor system. Corollary discharge is an internal sensory signal generated by the motor system, so it is not a giant leap to identify one with the other. This similarity was noticed by Grush (1995), who proposed that motor imagery should be identified with the forward model system. His proposal and the distinction between it and the proposal of this dissertation are discussed in subsection 1.6.1 below.

There have been some brain imaging studies that support the presence of corollary discharge in inner speech. Tian (2010) proposed that speech imagery involved corollary discharge mechanisms, and using MEG, found evidence of activation in both motor and auditory areas when participants engaged in inner speech (consistent with the theory). Numminen and Curio (1999) used MEG and found that auditory areas were less active in response to auditory stimulation if the subject was engaged in subvocal articulation. Such attenuation is a typical sign of the presence of corollary discharge. Kauramäki et al. (2010) performed a similar study showing that the N100M auditory-cortex response to tones was attenuated by c.20-25% when participants silently articulated, again showing the sensory attenuation that is the hallmark of corollary discharge.

1.6.1 Grush’s Theory - the Kalman Filter

The theory proposed in this dissertation is similar to that proposed by Grush (1995), who argued that motor imagery is generated by running a forward model offline (i.e., without actually executing the motor commands)\footnote{Grush’s theory is somewhat reminiscent of the view of imagery proposed in 1940 by Sartre. In a work published ten years before von Holst and Mittelstaedt or Sperry reintroduced the notion of the forward model.} When a forward model
is used in this way, it is termed an *emulator*, since it is emulating the response of the body. In Figure 1.3 this emulation would be equivalent to the *effference-copy-to-forward-model* path, without commands being sent to the body as well.

The basis of Grush’s theory is that the forward model uses a *Kalman filter* to constantly update the predicted state of the body. The *Kalman filter* is an algorithm developed in the field of control engineering which takes estimates of the state of a system (each estimate derived from a different source) and provides the statistically most reliable estimate of the state. The essential idea of a Kalman filter is that any source of information, no matter how unreliable, has something to offer and so a Kalman filter will accept any source of information about the state of a system, with each source’s estimate weighted by the (predicted) reliability of that source. The Kalman filter has been used in models of sensorimotor integration because it offers an optimal estimate of the state of any system. An organism never *knows* the state of its body (the position and movement of its limbs . . .); all an organism can do is to infer its body’s state from the available information. This is what the Kalman filter offers: Combining various sources of information to provide the best estimate of the current state of the system.

The important point of this discussion is that Grush identifies imagery with the Kalman filter update of the forward model and thus with the predicted *state* of the body. That is, when we are imagining an action we are experiencing a sequence of predicted changes in body-state. The theory proposed for inner speech in this dissertation is very similar to Grush’s, however I identify imagery with another component of the internal-model loop: corollary discharge. This means that I do not identify imagery with a predicted state of the body caused by an action but with the predicted sensory consequences of the action.

I do not want to draw too big a distinction here, since corollary discharge provides some of the information the Kalman filter uses to update its estimate of *effference copy/corollary discharge*, Sartre proposed a concept of motor-intention that was very similar, and suggested a motor-based theory of imagery (Sartre 2004).

28 Here are some of the technical details: A source’s estimate of a state is treated as a normally-distributed probability density-function, with the mean of the distribution representing that source’s prediction of the state and the variance of the distribution representing the reliability of the source. Thus, if a source is completely unreliable, its estimate is weighted to have no influence and, conversely, if a source is infallible, then its estimate is weighted to be a certainty.
the body’s state. However there is a defining difference between Grush’s proposal and mine in that I identify the sensory components of speech imagery with corollary discharge and corollary discharge is responsible for the attenuation of self-caused sensory signals (as discussed in section 1.5). So, by identifying inner speech with corollary discharge, I predict that when inner speech (and thus the corollary discharge signal) matches a simultaneously presented external sound the effects of that external sound will be attenuated (as shown in chapter 4). This is not necessarily what would be predicted from identifying imagery with a process of updating the forward model itself.

1.6.2 Brain Areas Involved in Internal Models and Speech Imagery

Attempting to localize functions to specific parts of the brain often misses the point that brain processing is massively parallel and a lot of processing will be decentralized. Furthermore, the theory presented in this dissertation is functional and so is not tied to any particular realization in the brain. With those caveats in mind it is important to note that one region that does crop up repeatedly in studies of corollary discharge: the cerebellum. Speech (including inner speech) will necessarily involve several brain regions. Most obviously the motor system is needed to generate sounds and the auditory system is needed to process those sounds. However, in terms of internal models and the comparing of reafference with corollary discharge the cerebellum seems to be central.

Several researchers have proposed a role for the cerebellum in forward model/corollary discharge function. These proposals were initially based on an analysis of the anatomy of the cerebellum and its connections to other brain areas, which suggested that it played a key modulatory role in motor control. Anatomically, the cerebellum, with modular layout and exclusively inhibitory outputs, seems the ideal structure for internal models. Indeed, in electroreceptive fish, it has been shown that it is the homologue of the cerebellum that deals with the forward/inverse models needed for electroreception (Bell, 2001). This anatomical structure has led to many theories of motor control which postulate a role for the cerebellum

29The input of corollary discharge into the update of the forward model is not represented in Figure 1.3. Corollary discharge is used in combination with reafference (when that is available) to provide one source of information about the state of the body.
in generating or evaluating internal models and corollary discharge.

For example, Ito (2008) argues that the anatomical connections into and out of the cerebellum are consistent with the cerebellum being the origin of either forward or inverse models. Similarly, Wolpert et al. (1998) argues that the cerebellum is the likely source of internal models, and suggests it contains pairs of inverse and forward models working together. Desmurget and Grafton (2003) make a similar claim, but argue that the cerebellum’s role is that of storing inverse models, not forward ones.

In addition to the anatomical evidence, there is significant brain-imaging support for the claim that the cerebellum is central to corollary discharge function. Blakemore et al. (1998) found, using fMRI, that the cerebellum was more active for self-produced, in comparison to externally produced, touch sensations, suggesting that the cerebellum is involved in generating the error signal associated with comparing reafference with corollary discharge.

Similarly, Blakemore et al. (2001) used positron emission tomography (PET) and found that the cerebellum showed greater activation in step with greater discrepancy between predicted sensation and actual sensation (the discrepancy was created by introducing a delay between movement and touch). This is consistent with it playing a crucial role in comparing reafference with corollary discharge.

In terms of speech processing, the DIVA model assigns part of the forward model role to the cerebellum (Guenther et al., 2006).

Lastly, there is significant brain imaging data suggesting that the cerebellum is a component of inner speech.

Ackermann et al. (1998) found cerebellar activation in inner speech using fMRI. Ackermann et al. (2004) found the same and Ryding et al. (1993) found cerebellar activation in silent counting. Furthermore, Katanoda et al. (2001) found cerebellar activation in silent naming of pictures, and Hubrich-Ungureanu et al. (2002) found cerebellar activation in silent speaking using fMRI.

1.7 Some Issues Left Unresolved by this Dissertation

There are a few important issues that this dissertation does not have the data to address. I would like to mention them here in the spirit of “full disclosure”.

41
First there is the intractable problem of constituting vs. accompanying. I am arguing that corollary discharge constitutes the sensory content of speech imagery (perhaps in conjunction with other elements, such as purely symbolic phonemic representations). However, it is difficult to prove a constitutive role as distinct from an accompanying role. I believe this dissertation demonstrates the existence of corollary discharge during speech imagery, but this could be explained by arguing that corollary discharge merely accompanies speech imagery rather than constituting it. After all, corollary discharge is assumed to accompany the sensory content of external speech without constituting it. This is a difficult distinction to tease apart (and is, in part, an issue of definition) and I leave it for further research.

I would also like to draw attention to the fact that this dissertation does not resolve the issue of whether there are two forms of inner speech or only one (see section 1.3). This is also a difficult question to resolve, as discussed in chapter 4, since it is difficult to rule out the possibility that when participants are asked to only imagine a speech sound without articulating it, they are in fact producing very small movements of their articulators (if even occasionally), which would mean that conditions involving such non-articulated speech imagery would in fact contain a certain amount of low-level articulation.

A final issue is the interconnection between the theory proposed here and the Motor Theory of Speech Perception (discussed in section 5.1). The Motor Theory of Speech Perception, as the name implies, proposes a role for the motor system in speech perception. The theory I am proposing in this dissertation assumes that the motor system generates the sensory content of speech imagery (via corollary discharge) and so speech imagery will necessarily mean that the motor system is active\(^{30}\). Thus, any experiment involving speech imagery is open to its results being explained in terms of the Motor Theory. While this is an unavoidable issue for certain experiments in this dissertation (those reported in chapter 2), I believe that the overall pattern of results across all experiments in this dissertation fit better with an imagery/corollary discharge explanation. This issue is discussed further in subsection 2.3.6.

---

\(^{30}\)With the possible exception of genuinely ‘pure’ speech imagery (as discussed above) if such a thing exists.
1.8 Organization of the Dissertation

To summarize the main claims of this dissertation: I argue that the sound of speech is generated by the motor system in inner speech just as it is in external speech. Furthermore, the neural representation of the sound of inner speech overlaps sufficiently with the auditory processing of the sound of external speech to influence the perception of external speech.

I also argue that the sensory content of inner speech is constituted by corollary discharge, which is a sensory-prediction signal generated by forward models. This corollary-discharge signal attenuates the impact of self-caused sounds, and thus the impact of speech sounds should be attenuated in the presence of speech imagery (under the hypothesis that corollary discharge is present in speech imagery).

In chapter 2, I first establish experimental evidence for the claim that triggering of inner speech can alter the perception of external speech sounds. I present the results of two experiments. The first experiment establishes that when played sounds ambiguous between /a'ba/ and /a'va/, participants’ perception of these sounds is altered by mouthing or imagining /a'ba/ and /a'va/ in synch with these external sounds. The second experiment demonstrates that this effect is not the result of category priming, and suggests that it is dependent on the sensory content of speech imagery.

Having established that inner speech (in both enacted and pure forms) can influence the perception of external sounds, chapter 3 expands this finding. The first two experiments in this chapter explore the duration of the perceptual influences of speech imagery (whether the effects reported in chapter 2 can be shown to linger). These experiments use a sensory-recalibration paradigm (Bertelson et al., 2003) to show that repeated exposure to the effects reported in chapter 2 has lingering consequences, continuing to alter perception after the imagery has stopped. The third experiment in this chapter assesses sensory attenuation using adaptation (Samuel, 1986). One of the uses of corollary discharge is to prevent self-produced sounds from interfering with our perception of external sounds (see section 1.5). Repeated exposure to a sound can ‘fatigue’ the auditory system; I tested the possibility that corollary discharge (proposed to be present in mouthing) would prevent this ‘fatiguing’ and thus inhibit selective adaptation. However, I was unable to find even
tentative support for a reduced level of adaptation in the presence of speech imagery.

In chapter 4, I present the results of three experiments which support the other main claim of this dissertation, that corollary discharge is present in speech imagery. These experiments demonstrate that when mouthing in time to an external sound, the impact of that external sound is attenuated. These experiments use the Mann context-effect as a method of assessing attenuation (Mann 1980). Sensory attenuation is the prime diagnostic for the presence of corollary discharge, as discussed in subsection 1.5.5, and so the experiments in this chapter are the strongest evidence that speech imagery does involve corollary discharge.

The experimental chapters are followed by a discussion of some issues that are related to the dissertation’s main claims, but were too far afield to discuss in the introduction or experimental chapters. This discussion includes several very speculative ideas that are presented merely as suggestions and possible topics for future research.

The final chapter is a brief conclusion which summarizes the work of the dissertation.
Chapter 2

Interaction of Speech Imagery with the Perception of External Speech

...le timbre; il appartient également à la parole, et il se retrouve aussi dans la parole intérieure.

...timbre; it belongs equally to speech, but is also found in inner speech.

— Victor Egger (1881)

2.1 Introduction

This chapter reports on two experiments which establish that speech imagery can influence the perception of external speech sounds. The properties of this effect support the claim that inner speech contains sensory information. Both enacted and non-enacted forms of imagery are shown to influence speech perception. As discussed in chapter 1, enacted speech imagery refers to inner speech produced by silently mouthing speech sounds; non-enacted speech imagery is inner speech without any such movement.

This dissertation argues that the sensory content of speech imagery is consti-
tuted by corollary discharge, which is a sensory-prediction signal. Thus, this dissertation claims that the perception of external speech is altered because corollary discharge prepares the auditory system to hear those sensory features which the corollary-discharge signal carries.

One of the earliest demonstrations of an interaction between imagery and perception (in the visual domain) was performed by Perky in 1910, and so is known as the “Perky effect”. In Perky’s experiment, participants were asked to look at a blank screen and to imagine specified objects (such as a banana). While they were imagining, a very faint picture of the imagined object was projected onto the screen. Participants were typically unable to distinguish their imagining from the projected picture and in many cases participants incorporated aspects of the projection into their imagining without being aware of it (e.g., imagining the banana in the orientation of the projection despite reporting that they were trying to imagine it in a different orientation).

The Perky effect demonstrates that visual imagery can interfere with perception, suggesting that these two activities overlap, to some degree, in terms of representation/processing. This suggestion has been confirmed with both behavioural and brain-imaging studies. It should be noted that the vast majority of imagery studies deal with visual imagery. While this dissertation is concerned with speech imagery, the evidence from visual imagery is useful in providing a point of comparison.

Kosslyn and Thompson (2000) discuss three streams of evidence for the claim of shared mechanisms underlying visual perception and visual imagery. First, there are behavioural studies showing that visual imagery can interfere with visual perception, depending on the degree to which the imaged object is visually similar or dissimilar to the object to be perceived (e.g., Kosslyn et al., 2006). Second, there are studies showing that neural damage that compromises visual perception can cause parallel problems with visual imagination (e.g., Shuttleworth et al., 1982). Third, there are brain-imaging studies showing that many of the areas activated during perception are also activated during imagination (e.g., Farah, 2000).

However, while perception and imagination may be supported by many of the same structures, the overlap is not complete. There are some neurological patients who have visual-processing damage without parallel damage to visual imagination
In the field of brain imaging, an fMRI study by Ganis et al. (2004, p.226) found that:

“visual imagery and visual perception draw on most of the same neural machinery. However [...] the spatial overlap was neither complete nor uniform; the overlap was much more pronounced in frontal and parietal regions than in temporal and occipital regions.”

There have been fewer studies on auditory imagery, but what evidence there is seems to support a situation similar to vision, with significant but not complete overlap between the processes involved in perception and imagery.

Zatorre et al. (1996) found that there was a great degree of overlap in brain regions activated by sound and by auditory imagery. Bunzeck et al. (2005) found that similar auditory regions were activated in both imagining and hearing common non-speech sounds (such as a hair-dryer, hands clapping ...). These regions included the secondary auditory cortex, but not the primary auditory cortex. This is a common finding in such studies — auditory imagery does not seem to involve the most basic of auditory cortical areas (primary auditory cortex). This was found for speech imagery by Shergill et al. (2001), and for musical imagery by Ohnishi et al. (2001). However it is dangerous to draw conclusions from null results since it is always possible that the sensitivity of the brain-imaging techniques is simply not great enough to detect activation (see, for example, King, 2006); and, in fact, some studies have found primary auditory cortex activation in auditory imagery (e.g., Yoo et al., 2001).

In addition to the brain-imaging studies discussed above, there are many behavioural studies demonstrating a link between auditory imagery and auditory perception. For example, Crowder (1989) investigated the effects of auditory imagery of timbre. In this experiment, participants were played a sine-wave tone and asked to imagine the tone as played by a given instrument. They were then played tones from different instruments. When the timbre of what they heard matched the timbre of what they were imagining, they were faster at determining whether the pitches matched.

Segal and Fusella (1970) replicated the Perky effect for auditory imagery, finding that in a detection task, auditory stimuli could be confused with auditory im-
agery of similar sounds. Farah and Smith's (1983) findings were different — they found that imagining a tone of a particular frequency improved detection of externally presented tones of the same frequency, rather than causing external and imagined tones to be confused. Okada and Matsuoka (1992) used a slightly different experimental design and confirmed Segal and Fusella's (1970) results, finding that auditory imagery of tones interfered with the detection of tones of the same pitch. Okada and Matsuoka (1992) argue that disagreement between their results and Farah and Smith's (1983) is that the structure of Farah and Smith's (1983) experiment made it a discrimination task, not a detection task. Okada and Matsuoka argue that imagery can aid discrimination/identification but may interfere with detection. This fits with the experiments presented below which show that speech-imagery influences discrimination.

Of most direct relevance to this dissertation is the work of Sams et al. (2005). This study was not about speech imagery but it did involve participants mouthing speech sounds. The results showed that when participants mouthed /ka/, this interfered with their perception of /pa/. Sams et al. interpreted their findings as the result of triggering efference copy (a claim with which this dissertation would agree), but did not discuss a relation to speech imagery.

The following experiments extend the research of Sams et al. (2005), replicating the effects of mouthing speech sounds (which I would classify as enacted speech imagery) and comparing these effects with those of non-enacted speech imagery, that is with silent, non-articulated, inner speech.

The experiments reported in this chapter do not specifically address the question of whether the sensory content of inner speech is constituted by corollary discharge; that issue is taken up later, in chapter 4.

2.2 Experiment 1-1

This experiment establishes the basic claim that both enacted and non-enacted imagery can influence the perception of speech sounds. This is both a replication and extension of Sams et al. (2005). The presence of sensory content in inner speech is

---

1In a survey of the literature, Hubbard (2010) agrees that the best conclusion is that imagery can aid discrimination/identification but may hinder detection.
addressed in Experiment 1-2. Evidence that this sensory content is constituted by corollary discharge is presented in chapter 4.

Participants were asked to mouth or imagine one of two sounds (/א"ו/ or /א"ב/) in synchrony with an external sound (a sound that was itself ambiguous between /א"ו/ and /א"ב/) and then to categorize the ambiguous sound. There was also a baseline condition in which participants simply categorized the ambiguous sounds without performing any mouthing or imagery.

This gives us the following structure for Experiment One: There are 5 types of block — 2 Conditions (Mouthing and Imagining) with two levels within each condition (/א"ו/ and /א"ב/) and a Baseline condition, as shown in Table 2.1.

<table>
<thead>
<tr>
<th>Action</th>
<th>Mouthed or Imagined Sound</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mouth</td>
<td>/א&quot;ב/</td>
</tr>
<tr>
<td></td>
<td>/א&quot;וו/</td>
</tr>
<tr>
<td>Baseline (just listen)</td>
<td></td>
</tr>
<tr>
<td>Imagine</td>
<td>/א&quot;ב/</td>
</tr>
<tr>
<td></td>
<td>/א&quot;וו/</td>
</tr>
</tbody>
</table>

Table 2.1: Structure of Experiment 1-1: 5 types of block

Corollary discharge is a sensory prediction and so if an external sound is presented in synchrony with corollary discharge, then ambiguous content in the external sound should be influenced by the sensory prediction, leading to the external sound being perceived as matching the prediction. Under the assumption that the sensory content of speech imagery is constituted by corollary discharge, this leads to the following predictions for this experimental design: in the Baseline condition sounds should be perceived roughly equally often as /א"ב/ and /א"וו/, while both Mouth /א"ב/ and Imagine /א"ב/ should induce more /א"ב/ percepts and both the Mouth /א"וו/ and Imagine /א"וו/ conditions should induce more /א"וו/ percepts.

These predictions are exactly the same as those for a priming model. Under a priming interpretation, people would be expected to hear more /א"וו/ when
mouthing /aˈva/ because, by mouthing the sound, they have primed their mental representation for /v/ making it easier for that category to be triggered by external sounds.

The distinction between priming and corollary discharge is not addressed in this experiment, but is taken up in Experiment 1-2.

The predictions of this experiment are also identical to those for an attentional explanation. That is, if a person is mouthing or imagining a sound, that may draw attentional resources to the sensory features of these sounds and thus increase the likelihood of an ambiguous sound being perceived in line with what is mouthed/imagined. This is not, I believe, a confound for my experiment. Rather, I would argue that this is simply restating my claim using different terminology. This issue is discussed in subsection 1.5.7.

2.2.1 Methods

The structure of each trial was identical. Participants were asked to Mouth (i.e., silently articulate) or Imagine (without articulation) a disyllable (either /aˈba/ or /aˈva/) in synchrony with a target sound (a sound ambiguous between /aˈba/ and /aˈva/). They were then asked to categorize the target sound in a forced choice between “aba” and “ava”.

All sounds were presented in free field. This is the case for all experiments reported in this dissertation except Experiment 3-3.

In pilot testing it was found that the impact of imagery was highly dependent on participants mouthing/imagining in close synchrony with the recorded sound. It can be hard to be in synch with a sound played in isolation, so prior to each target, timing was first established. This was done by playing a ‘murmured’ sound before the target was played. The ‘murmured’ sounds were simply low-pass filtered versions of the targets, filtered at 275 Hz so that no phonetic information other than pitch, timing and amplitude envelope were available in the murmur. One token of the sound was played prior to each target so that participants could know exactly

2 And similarly hear more /aˈba/ when mouthing or imagining /aˈba/.
3 Except in the baseline condition, in which participants merely listened.
4 In the main text of this dissertation I have used IPA transcriptions, however participants were presented with the English spelling of the sounds (“aba” and “ava”).
when the target would occur and match their mouthing/imagining to its timing.

To augment the sense of rhythm established by the murmur sounds, a video of a red dot (like in Karaoke) played in time to the sounds. This red dot got bigger and smaller in synchrony with the intensity of the sounds. The videos looked like a large red bouncing ball whose size matched the sound track exactly, flashing bigger and smaller in rhythm with the sound.

The participant began each trial by pressing on the spacebar. After a 138 ms delay, the trial began with one token of the murmur sound followed by the target sound, with 138 ms between murmur and target. Participants were asked to articulate or imagine synchronously with both the murmur and target sounds. After the target was presented, a prompt appeared and participants identified the target as either “aba” or “ava”. Participants indicated their answer by pressing either the left or right arrow key on the computer keypad. The side of the “aba” and “ava” responses was counterbalanced across participants.

This experiment consisted of only 5 types of block (as shown in Table 2.1). All 5 types of block were presented in random order. To prevent order or position effects, 14 cycles of these 5 types of block were presented with a new randomization of their order on each cycle. The experiment thus consisted of 70 blocks (14 cycles X 5 types of block).

Participants completed 6 categorizations in each block. Since each block was repeated 14 times, this means that there were 84 tokens categorized for each type of block. The experiment took about 24 minutes to complete.

Before the experiment itself began, participants performed a ‘pre-test’ to determine where along the /’a’b/~/’a’v/ continuum their perceptual boundary lay. In this pre-test participants were presented with multiple repetitions of 11 equally-spaced steps from the continuum (ranging from a clear /’a’ba/ to a clear /’a’va/). Each step was presented 16 times in random order. The data were submitted to a probit analysis to determine the point along the continuum corresponding to the perceptual boundary between /’a’ba/ and /’a’va/ for that participant (the point along the continuum where they were equally likely to hear the sound as /’a’ba/ or /’a’va/). The 44% and 56% points were also determined. These three ambiguous sounds (different for each participant) were then used as the ‘to-be-categorized’
target sounds for the experiment. The categorizations of these three target sounds were pooled for analysis. This pre-test took about 4 minutes to complete.

Using sounds near each participant’s perceptual boundary, where ambiguity is highest, would presumably make it easier for external influences (such as mouthing and imagining) to have an influence, thus increasing the power of the experiment.

Figure 2.1: This is a timeline of how stimuli were presented in each trial. The “wa...WA” in the audio track represents the low-pass filtered ‘murmur’ sounds. The target sound is ambiguous between /a va/ and /a ba/ (represented as “a...VA ~ a...BA” in the audio track). Each ‘frame’ of this timeline represents roughly 200ms. Notice how the red ball is smaller for the first (unstressed) syllable and larger for the second (stressed) syllable — the diameter of the red ball matches the intensity of the sound.

As a cover-story, participants were told that jaw movement induces a protective middle-ear reflex (the tensor tympani reflex), and that this experiment was investigating how triggering of that reflex effects auditory processing speed. Thus, participants were led to believe that this was a response-time experiment. Participants were also told that the reflex was particularly strongly activated in speech, so were asked to imagine hearing their voice in their head as they mouthed the sounds (in the Mouth condition), to ensure that the activity was properly ‘speech-like’; and to imagine the sounds vividly in the Imagine condition. They were warned not to whisper and a microphone in the booth monitored for any audible signs of whisper. Participants reported that they had no trouble keeping the mouthing silent and no measurable whispering was recorded. This cover story was used for most

---

This pre-test design is based on that found in Bertelson et al. (2003). It might be objected that a logit analysis would be more appropriate. However the choice of one test over the other is largely moot since the two analyses typically produce nearly identical results. For comparison purposes, I also performed a logit analysis on each participant’s data and the results of the two analyses were essentially identical.
Because this experiment was long and demanding, a closed-loop camera was set up in the sound-treated booth so that participants could be monitored from outside to ensure that they were performing the task correctly.

Participants were given an extended practice session to familiarize them with the experimental set-up and only moved on to the main experiment when both they and the experimenter felt they were ready. Before beginning each block, participants had to type in a two-letter code that represented the condition (Mouthing or Imagining) and the syllable (/a’ba/ or /a’va/) that they were about to perform. This was to ensure that participants were paying attention to the task. The experiment itself took about 24 minutes to complete. Participants were monitored via closed-circuit camera and microphone to ensure they were performing accurately.

Sounds were presented over speakers at c. 60dB for both murmur and target sounds.

This experiment (and all others in this dissertation) was conducted in a sound-treated booth at the Interdisciplinary Speech Research Laboratory in the Linguistics department at the University of British Columbia. All experiments reported in this dissertation were run on Psyscope (Macwhinney et al., 1997).

2.2.2 Stimuli

A continuum from /a’ba/ to /a’va/ was created using STRAIGHT (Kawahara et al., 1999). This software uses a synthesis technique to generate intermediates between two recorded endpoints. The software produces an accurate filter function of the spectrum of the recorded endpoints and then generates an acoustic compromise between the two sounds at any specified intermediate (e.g., 10% similar to sound one, 90% similar to sound two). The results of this synthesis sound surprisingly natural, not at all like synthesized speech. STRAIGHT also allows for the spectral information to be altered independently of the pitch and intensity contours, thus allowing two sounds to be morphed along a spectral continuum, while keeping

---

6 I would like to thank Dr. Kawahara for providing this software (for free!) and for the considerable amount of time he spent showing me how to use it.

7 The percentage of morphing represents a percentage of the log-distance in frequency between the spectra of the two sounds.
pitch, duration and intensity constant. Using this system allows for very natural-sounding continua and for a very fine-grained division of the continuum space. One issue with using STRAIGHT, though, is that it does not morph individual aspects of the acoustics in isolation from others. So, all aspects of the spectrum are altered together.

A female native English speaker was recorded saying /æˈvæ/ and /æˈbæ/. Tokens that were similar in length and intonation and free of artefacts were chosen as the basis of morphing. A 168-step continuum was created with STRAIGHT. The sounds were 481ms in duration (the duration of all tokens was identical). As described above, only three tokens from this continuum were used for each participant. The three target tokens were different for each participant and corresponded to the steps on the continuum closest to each participant’s perceptual boundary ± 6% (i.e., 44%, 50%, 56%), as determined by a probit analysis of pre-test data (described above). Tokens were taken from near the category boundary so that they would be maximally ambiguous. Three tokens were selected so that participants would hear at least some variation in the tokens presented to them. In the analysis, categorizations of all three tokens were pooled.

The murmured versions of the sounds were created by low-pass filtering the target sounds at 275 Hz, using Praat (Boersma and Weenink, 2001).8

2.2.3 Participants

There were 20 participants (average age = 21.7 years; SD = 3.7 years); all of whom received course credit or payment for their participation.

A potential problem with this experiment (and all others involving corollary discharge) is the possibility of a mismatch between the sex of the participant and the sex of the voice used for the stimuli. Corollary discharge is a sensory prediction and if there is too large a distance between the content of the prediction and the content of the external sounds used as stimuli, there may be little influence of one on the other (see subsubsection 1.5.5). Sex differences in voice may be one such source of mismatch. Thus, in order to eliminate this potential source of conflict only female participants were run in this experiment (to match the stimuli which

---

8This is a free phonetics software package that has become a standard tool in phonetics research. It includes a scripting language.
were of a female voice).

This was kept consistent for the whole dissertation: experiments that used female voices for stimuli used female participants and those that used male voices used male participants.\footnote{Except for the first experiment reported in chapter 3, which was run before this issue was considered.} This also means that most of the experiments in this dissertation used female voices and female participants (this was a pragmatic necessity because more women than men volunteered for these experiments).

\subsection*{2.2.4 Results}

A repeated-measures ANOVA was performed with the 5 types of experimental block (\textit{Mouth /$\text{a'ba}$/}, \textit{Mouth /$\text{a'va}$/}, \textit{Baseline}, \textit{Imagine /$\text{a'ba}$/}, \textit{Imagine /$\text{a'va}$/}) as 5 levels of a single factor. This ANOVA was highly significant $[F(4,76) = 39.489, p < 0.001]$. Pairwise comparisons (using a Holm-Bonferroni correction) confirmed that each of the 5 experimental manipulations was significantly different from all the others ($p < 0.05$). These results are shown in Figure 2.2.
Figure 2.2: Experiment 1-1 Results: The data are scored as % of times the target sounds were categorized as /aˈbɑ/. A lower score means more /aˈvɑ/-perceptions and a higher score means more /aˈbɑ/-perceptions. Standard-error bars are shown.

This demonstrates that both enacted speech imagery (mouthing) and ‘pure’ imagining of a speech sound can influence the perception of an external speech sound.

2.2.5 Discussion

When mouthing or imagining /aˈbɑ/ participants were more likely to perceive an ambiguous sound as /aˈbɑ/. The same pattern was seen for mouthing or imagining /aˈvɑ/.

The two-way directionality of the effect found in this experiment (mouthing/imaging /aˈbɑ/ pulling perception in one direction but mouthing/imaging /aˈvɑ/ pulling
in the opposite direction) demonstrates that it is the sound being mouthed/imagined that is responsible for the effect and not some extraneous factor like cognitive load.

This experiment also showed that the difference between Mouthing and Imagining was significant. For both /aˈba/ and /aˈva/, the Mouth conditions were significantly further from the Baseline than the Imagine conditions (as shown in the pair-wise comparisons). That is, there is significantly more of an influence on perception when a person mouths a sound in comparison to when they imagine it.

Under the corollary discharge account I am proposing, these results are explained as follows: speech imagery (realized as either mouthing or pure imagery) produces corollary discharge, which is a prediction of what is about to be heard. This prediction has sensory detail and influences the perception of the external sounds, causing the external sound to be heard in line with the prediction. Such an interpretation is equivalent to an explanation in terms of attention as discussed in subsection 1.5.7.

For example, when the sound /aˈva/ is mouthed or imagined, a corollary-discharge signal is produced that carries an expectation of hearing a consonant that is a fricative and is labiodental. When an external sound is presented in synchrony with this corollary-discharge signal, those components of the external sound which are compatible with the sensory prediction are perceived in line with the prediction and so the acoustics of the consonant tend to be heard as being fricative and labiodental, hence the sound is perceived as /aˈva/.

I should note that in the example above I referred to “fricative” and “labiodental”. These are traditional phonetic descriptions, however I merely use them for ease of exposition and I am not taking a stance on whether the corollary-discharge signal carries information that is coded in this way (for a discussion of the coding of the corollary-discharge signal, see subsection 1.5.6 and subsection 5.1.2).

An alternative explanation is that the effects reported above are due to category-priming. Perhaps the phoneme /b/ was simply primed by mouthing/imagining and so people were more likely to hear /aˈba/. This issue is taken up in Experiment 1-2.

It is important to note that the degree of perceptual interference was not the same between articulate and imagine blocks. Silent articulation (mouthing) had a stronger influence than simply imagining. The difference in the strength of the effect, under the theory presented in this dissertation, would be accounted for by
assuming that forward models (and thus corollary discharge) are more highly activated the more overt the activity becomes. This could mean either a higher level of activation of a set of forward models, or alternatively more forward models being engaged. Thus, non-enacted imagery would result in the lowest level of forward-model activation while overt speech would involve the highest level of forward-model activation, with silent mouthing somewhere in between. This issue (of different levels of corollary discharge) is taken up in chapter 4.

A final point is that the results of this experiment do not completely tally with the experience that my research partner (Henny Yeung) and I have when we present this effect to others. When we demonstrate this effect and ask people to check for themselves whether they experience the shift in perception, most people report that the effect is extremely compelling and occurs almost every time they mouth or imagine in synchrony with the target. While the experiment above did find that mouthing/imagining had a significant impact, the effect was weaker than we would have predicted. I suspect the reason for this is that participants were not told about the purpose of the experiment (they were told the experiment was testing the impact of the tensor tympani reflex on their auditory processing speed); thus they presumably were trying to distinguish the presented target sound from the distraction of what they were mouthing/imagining. This probably led participants to fight against confusing the two, thus reducing the impact of mouthing/imagining. In a set-up where participants are told of the effect of mouthing/imagining and do not try to pull apart the mouthed/imagined sound from the heard sound, there may be a much stronger effect. This is taken up in chapter 3.

2.3 Experiment 1-2

The results of Experiment 1-1 established the basic result that both enacted (mouthing) and non-enacted (pure) imagery can influence speech perception. Experiment 1-2 examines whether this influence can be found when the phonemic category of the mouthed/imagined sound is different from that of the external sound which participants categorize. That is, will mouthing or imagining /a'pa/ influence people’s perception of a sound ambiguous between /a'ba/ and /a'va/, causing them to hear more /a'ba/ percepts? And similarly, will mouthing or imagining /a'fa/ induce more
/a va/ percepts? If so, this would suggest that speech imagery includes lower-level sensory information. Furthermore, such an influence would suggest that the effects reported in Experiment 1-1 are unlikely to be due solely to category priming (i.e., due to priming of the phonemic categories).

This experiment was essentially identical in design to Experiment 1-1 — the only difference being the sounds that participants were instructed to mouth/imagine. As with Experiment 1-1, participants mouthed or imagined a speech sound in synchrony with an external sound and then categorized that external sound. The sounds to be mouthed/imagined were always /a'fa/ or /a'pa/ and the external sounds were always ambiguous between /a'ba/ and /a'va/ (taken from a computer-generated continuum). As with Experiment 1-1, a baseline was included in which participants simply listened to and categorized the target tokens, without mouthing/imagining. This gives us the following structure:

Table 2.2: Structure of Experiment 1-2: 5 types of block

<table>
<thead>
<tr>
<th>Action</th>
<th>Mouthed or Imagined Sound</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mouth</td>
<td>/a'pa/</td>
</tr>
<tr>
<td></td>
<td>/a'fa/</td>
</tr>
<tr>
<td>Baseline (just listen)</td>
<td></td>
</tr>
<tr>
<td>Imagine</td>
<td>/a'pa/</td>
</tr>
<tr>
<td></td>
<td>/a'fa/</td>
</tr>
</tbody>
</table>

The theory that this experiment is testing is that speech imagery contains sensory detail, not just category-level information (phoneme-level information). If this is the case then speech imagery which shares sensory detail with an external sound (but belongs to a different phonological category) may still influence the perception of the external sound. For example /f/ is a different phonological category from /v/, but shares the properties of labiodentality and frication. If imagining/mouthing

---

10 Keeping in mind the discussion from subsection 1.5.6, in which I identify sensory information as information below the level of the phoneme.

11 I am using traditional feature specifications for convenience, but without making any theoretical
/f/ induces people to perceive a /v/-/b/ ambiguous sound as belonging to the /v/ category, that would suggest that it is the presence of the sensory information in the /f/ that is responsible. A parallel situation holds for /p/, which shares the features of being a stop and being bilabial with /b/; and thus mouthing/imagining /p/ in synchrony with a /v/-/b/ ambiguous sound should induce people to perceive that ambiguous sound more often as /b/. If these influences are demonstrated, that would be strong evidence that speech imagery does indeed contain information below the level of the phoneme. Furthermore, if such an influence is found it would make the phoneme category-priming interpretation of Experiment 1-1 less tenable.

Previewing the discussion in subsection 2.3.4 below, the result of this experiment is that, as predicted, mouthing /a'fə/ and imagining /a'fə/ both influenced speech perception of the /ə'ba~/~/ə'va/ ambiguous target sounds, making them sound more like /ə'va/. Similarly, mouthing and imagining /ə'pa/ caused the ambiguous targets to sound more like /ə'ba/.

### 2.3.1 Methods

The procedures were identical to Experiment 1-1 with the exception of the content of what was mouthed/imagined. The sounds which participants mouthed/imagined were always /ə'pa/ or /ə'fə/. Participants were asked to mouth or imagine these disyllables in synchrony with an external sound and then to categorize the external sound in a forced choice between “aba” and “ava”. To facilitate rhythm, ‘murmur’ movies were again used. As in Experiment 1-1, the main experiment was preceded by a pre-test to determine participants’ perceptual boundaries.

The predictions of this Experiment are straightforward: In the Baseline condition sounds should be perceived roughly equally often as /ə'ba/ and /ə'va/, while both Mouth /ə'pa/ and Imagine /a'pa/ should induce more /ə'ba/ percepts and both the Mouth /ə'fə/ and Imagine /ə'fə/ conditions should induce more /ə'va/ percepts.

As with Experiment 1-1, participants completed 6 categorizations in each block; since each block was repeated 14 times, this means that there were 84 tokens categorized for each type of block. The experiment took about 24 minutes to complete.
The same tensor-tympani cover story was used, leading to participants believing that this was a response-time experiment.

As with Experiment 1-1, sounds were presented over speakers at c. 60 dB for both murmur and target sounds.

2.3.2 Stimuli

Exactly the same stimuli were used as in Experiment 1-1.

2.3.3 Participants

There were 20 female participants (average age = 21.5 years; SD = 3.5 years). All received course credit or payment for their participation.

2.3.4 Results

A repeated-measures ANOVA was performed with the 5 types of experimental block (Mouth /ɑˈpɑː/, Mouth /ɑˈfɑː/, Baseline, Imagine /ɑˈpɑː/, /Imagine /ɑˈfɑː/) as 5 levels of a single factor. This ANOVA was highly significant [F(4,76) = 52.215, p < 0.001]. Pairwise comparisons (using a Holm-Bonferroni correction) confirmed that each of the 5 experimental manipulations was significantly different from all the others (p < 0.05). These results are shown in Figure 2.3
Figure 2.3: Experiment 1-2 Results: The data are scored as % of times the target sounds were categorized as /aˈbɑl/. A lower score means more /aˈvɑl/-perceptions and a higher score means more /aˈbɑl/-perceptions. Standard-error bars are shown.

2.3.5 Discussion

As can be seen in Figure 2.3, Mouthing /aˈfɑl/ caused people to perceive the /aˈbɑl~/~aˈvɑl/ targets more often as /aˈvɑl/, and conversely, Mouthing /aˈpɑl/ had the opposite effect, causing people to hear the targets as /aˈbɑl/.

This experiment also showed that the difference between Mouthing and Imagining was significant. For both /aˈfɑl/ and /aˈpɑl/, the Mouth conditions were significantly further from the Baseline than the Imagine conditions (as shown in the pair-wise comparisons). That is, there is significantly more of an influence on perception when a person mouths a sound in comparison to when they imagine it. This
is the same pattern seen in Experiment 1-1.

The greater effectiveness of mouthing could be attributed to many factors, but the most plausible seem to be that either more forward models are engaged in mouthing than in imagining, or that the same forward models are engaged more strongly in mouthing than in imagining.

The two-way directionality of the effect found in this Experiment (mouthing/imagining /a’fəl/ pulling perception in one direction but mouthing/imagining /a’pəl/ pulling in the opposite direction) is the same pattern shown in Experiment 1-1 and again demonstrates that it is the sound being mouthed/imagined that is responsible for the effect.

As discussed in subsection 2.2.4 under a corollary-discharge account, the effect of mouthing/imagining is to send a detailed sensory estimate (corollary discharge) to auditory areas, which channels incoming sounds into the sensory estimate, thus causing incoming sounds to be heard as more similar to the mouthed/imagined sound. In this experiment that means that Mouth /a’fəl/ or Imagine /a’fəl/ caused people to hear an /a’ba’/∼/a’va’/ ambiguous sound as containing a labiodental fricative, i.e., /a’va’/. Similarly, Mouth /a’pəl/ or Imagine /a’pəl/ caused people to hear an /a’ba’/∼/a’va’/ ambiguous sound as containing a bilabial stop, i.e., /a’ba’/.

A possible alternative explanation for the results in both this experiment and Experiment 1-1 is that we are seeing priming effects. That is, triggering the representation of /a’fəl/ or /a’pəl/ is simply sensitizing the participant to respond /a’ba’/ and vice versa for triggering /a’va’/ or /a’fəl/.

While this priming theory is certainly plausible for the results of Experiment 1-1, it is less tenable for the /a’fəl/ and /a’pəl/ cases reported in the current experiment, in which the mouthed/imagined sound was a different category from both the presented target-sound and from the resulting percept. In the /a’fəl/ case, participants perceived a target /a’ba’/∼/a’va’/ as being /a’va’/ even though the sound that they were mouthing/imagining was /a’fəl/. This could not be a case of simple category priming (since the percept is a different category from the prime).

Perhaps one could argue for a spreading-activation form of priming in which phonemes that share sensory features are neighbours in a multi-dimensional ‘phoneme space’ and so activating the representation of one phoneme will cause its neigh-
bours to also receive some raised level of activation. This spreading-activation model is certainly possible, but it becomes unclear how such a model is different from the claim that mouthing and imagery both contain sensory content that can influence the perception of external speech. Under both a corollary-discharge account and a spreading-activation model, this claim remains true.

Another issue with a priming model is timing. This was not pursued experimentally in this dissertation, however, while piloting these experiments it became clear that the effect of mouthing/imagining on perception is strongly dependent on timing — if the mouthing/imagining is not in synch with the audio (timed so that the audio could be the mouter/imaginer’s own voice), then the effect is much weaker. This is why the experimental design emphasized rhythm so much. The effect would not work without synchrony. This is suggestive of the sound being treated as feedback of the participant’s own production, in line with the corollary discharge hypothesis. However, as this issue was not tested experimentally, it can only be taken as anecdotal evidence.

A further argument against priming are the results reported in chapter 4, which would be hard to account for under a priming model.

While these arguments provide evidence that priming is not the cause of the results reported here, they do not conclusively exclude a priming explanation.

It is interesting to note that the effects found in this experiment are somewhat similar to the McGurk effect (McGurk and MacDonald, 1976). In the McGurk effect, video of a face pronouncing one speech sound is synchronized with the acoustics of a different speech sound. When people are presented with this mixed audiovisual signal, they typically perceive the whole as belonging to a third category, intermediate between the auditory and visual information. This is somewhat similar to the results found in this experiment, in which the acoustics of an /aˈbɑ:/~/aˈva/ ambiguous sound were combined with mouthing of /aˈpa/, causing people to hear /aˈbɑ:/ (and a parallel influence was found for mouthing/imagining /aˈfa:/).

12This is my own theoretical counter-argument, not a position widely held in the field.

13Though in some McGurk studies, a simple case of dominance is shown, in which the visual information dominates the auditory. For example, overlaying video of “ava” on audio of “aba” causes people to perceive the sound specified in the video, “ava”.

64
Comparison of Experiments 1-1 & 1-2

It is possible, though not explored in this dissertation, that the effects of category priming and corollary discharge are additive. If this were the case, one would expect to find that when the mouthed/imagined sound is of the same category as the resulting percept, both the sensory anticipation of corollary discharge and the category priming of the phonemes contribute to the perceptual impact and the shift in perception would be greater.\textsuperscript{14}

The experiments reported above were not intended to address the possibility of additive effects of phoneme priming and corollary discharge. However, a post-hoc comparison of the strength of the perceptual shift in Experiment 1-1 vs. Experiment 1-2 was performed. If category-priming were contributing to the perceptual shifts seen in Experiment 1-1, then we should expect the perceptual shifts in Experiment 1-1 to be larger than those in Experiment 1-2.

A post-hoc between-participants ANOVA was performed, with factors “Experiment” and “Treatment” (with “Treatment” being coded to compare mouthing vs. imagining across the two experiments). The effect of “Experiment” was not significant ($F(1,38) = 2.586, p = 0.28539$). This means that no statistically significant support for a role of phoneme priming was found; however as this is a post-hoc analysis with a small number of participants, the failure to find a significant effect is not very meaningful.

2.3.6 Motor-Theory Interpretation

The experiments reported in this chapter were conducted with co-authors: Henny Yeung, Bryan Gick and Janet Werker. These experiments grew out of a research program looking into the Motor Theory of Speech perception (Liberman and Mattingly, 1985). The Motor Theory claims that speech perception proceeds by using the acoustic signal to recover the articulatory gestures that created the sound. This theory and its relationship to corollary discharge is discussed in section 5.1. The current experiments are not couched in this framework, and I believe that the re-

\textsuperscript{14}Though, of course, when there is agreement between mouthed/imagined sound and percept, there will also be a greater degree of similarity between the sensory anticipation provided by corollary discharge and the sensory content of the external sound, which would also likely increase the impact of mouthing/imagining.
sults are better interpreted as due to corollary discharge. However, I should discuss the relationship between these experiments and the Motor Theory.

In both experiments reported in this chapter, it is shown that engagement of the speech articulators during perception influences the perception of speech. Such a prediction is *prima facie* compatible with the Motor Theory of Speech Perception. Given that this theory argues for a role of the speech motor-system in speech perception, a plausible prediction of the theory is that engaging the speech motor-system during speech perception may cause interference. It is for this reason that both of the experiments reported below include a ‘pure imagery’ condition. This condition is to determine whether the impact of mouthing is really due to the motor system or whether it is due to the speech imagery that accompanies mouthing.

In both experiments above the results show that the impact of ‘pure imagery’, while weaker than mouthing, has qualitatively the same effect on perception. This supports the claim that the effects reported in these experiments are due to imagery rather than to the Motor Theory of Speech Perception. I do, however, have to mention a couple of issues with this interpretation.

First, it is always possible, as discussed in section 1.7, that participants are engaging in occasional micro-movements of their articulators in the ‘pure imagery’ conditions and thus that these conditions are not genuinely free of motor engagement.

Second, my theory of corollary discharge presumes that the motor system is engaged during imagery in order to generate the corollary discharge. Thus, even under my proposal, the motor system is still predicted to be involved in both the mouthing and pure imagery conditions of the experiment.

These issues are difficult to disentangle and, since my theory presupposes motor-system involvement, there may be no way to pull apart a corollary discharge explanation from a Motor Theory explanation for the effects reported in this chapter. However, the experiments reported in chapter 4 provide strong evidence for the presence of auditory corollary discharge in speech imagery, evidence which is not plausibly accounted for under the Motor Theory. Thus, explanatory simplicity would support the claim that both the effects reported in this chapter and those reported in chapter 4 are due to corollary discharge. The alternative possibility, (that the effects in this chapter are due to the Motor Theory and those in chapter 4 are
due to corollary discharge) is less parsimonious. These issues are explored in more detail in section 5.1.

2.4 Conclusion

These experiments have established the first claim of this dissertation: that inner speech can influence the perception of external speech and that this influence involves not just phonemic, but sensory information. The influence is found whether inner speech is enacted (with movement of the speech articulators) or non-enacted (with no articulator movement).

The corollary-discharge account presented in this dissertation argues that the auditory expectation carried by corollary discharge channels the perception of external sounds into categories as similar to the corollary-discharge signal as possible. This means that ambiguity in the acoustic signal will be resolved in favour of the sensory content in the corollary-discharge signal. Obviously, if the external sound is not at all ambiguous and not at all compatible with the corollary-discharge signal, there can be little influence.

The effect of inner speech on external-speech perception is more fine-grained than can be accounted for under a simple category-priming model. In Experiment 1-2 it was found that mouthing or imagining /a′fəu/ induced a greater number of /a′vo/ perceptions (and more /a′ba/ percepts were induced by mouthing/imagining /a′pəu/). This cannot be due to priming the categorical representations, since the mouthed/imagined sounds and the resulting percepts belong to different categories. One could argue that this is a case of priming activity ‘leaking’ to similar neighbours (so that priming of /a′fu/ would leak some priming to similar /a′vo/), but such an explanation already admits the primary claim of these experiments, namely that inner speech has more content than mere phonemes and that this content can influence the perception of external speech.

A related interpretation of the data reported above is that they are due to multiple additive causes. Under this hypothesis, category-priming would contribute to, and so intensify, the shift in perception caused by corollary discharge. This is certainly possible and does not alter the central claims of this dissertation.

One issue that I should address here is the problem of micro-movements of the
articulators. As discussed in section 1.3, it is possible that participants were making tiny movements of their articulators in the *Imagine* tasks in both Experiment 1-1 and 1-2. Kleinschmidt and Toni (2005) discuss a similar issue with motor imagery. It has been found that it is difficult to rule out the effect of micro-movements, because people may not be able to prevent themselves from slightly executing an imagined action.

One cannot rule out such a possibility and if it is the case, then the experiments reported above are really comparing the effect of enacted inner speech with *less-enacted* inner speech. This is an interesting point to pursue, but it does not undermine the purpose of these experiments, which is to explore the effect and sensory content of inner speech. If inner speech exists on a continuum of motor engagement, then it is appropriate that these experiments reflect that.

Examining completely non-enacted inner speech may turn out to be impossible. Even with the administration of curare (see section 1.4) it is possible that motor commands still reach the articulators but cannot be executed.

Given this issue of possible motor engagement in the ‘pure’ imagery conditions, the results of these experiments cannot be taken as ruling out a Motor Theory explanation.

A final issue relates to corollary discharge. As discussed in subsection 1.5.5, the hallmark of corollary discharge is that it attenuates an organism’s response to self-produced sensations. If the effects found in Experiment 1-1 and Experiment 1-2 above are at least partly due to corollary discharge, it might be asked: ‘why does inner speech make external speech seem *more* like the mouthed/imagined sound rather than *less* like it?’ This issue relates to discrimination vs. detection. The experiments reported above tested discrimination of sounds, *not* detection (where sensory attenuation might interfere). Even in detection tasks, though, it should be remembered that, as discussed in subsection 1.5.5, corollary discharge does *not* necessarily attenuate the intensity of a sensation under the current view, but is perhaps better described as channelling perception and, by doing so, attenuating the *impact* of that sensation. For example, when you touch yourself, the perceptual features of that touch are not attenuated by corollary discharge, but the *impact* of the touch is attenuated which is why you cannot tickle yourself. This issue is dealt with in a lot more detail in chapter 4 where experiments are reported that
demonstrate that speech imagery attenuates the *impact* of perceived speech.

In summary, the experiments reported in this chapter support the claim that inner speech has sensory content and that the processing/representation of inner speech is similar enough to that of external speech perception for inner speech to influence external speech perception. This effect is present in both *enacted* inner speech and *non-enacted* inner speech.
Chapter 3

Interaction of Speech Imagery with Recalibration and Adaptation

Imagination comes to the rescue of sensation, and thus we hear better than our interlocutor speaks.
— Victor Egger (1881)

3.1 Introduction

This chapter reports on three experiments which examine the interaction of imagery with two other speech-perception phenomena: recalibration and selective adaptation. These two phenomena are superficially very similar, but they have been used to examine different questions in this dissertation; they are discussed within the relevant sections below.

The purpose of this chapter is to extend the findings reported in chapter 2 and to begin the work of establishing the parameters of imagery influences on speech perception.
The first two experiments reported below establish that the effect of imagery, described in [chapter 2], can continue to influence the perception of external speech even after the speech imagery has ended. This is a phenomenon known as **recalibration**. The third experiment examines whether speech imagery can attenuate the impact of selective adaptation (in line with the sensory-attenuation function of corollary discharge discussed in subsection 1.5.5). While the effect of selective adaptation was replicated, no evidence of an interaction with speech imagery was found.

### 3.2 Experiment 2-1

This experiment examines the duration of the effects established in [chapter 2]. In [chapter 2] speech imagery was shown to influence the perception of external speech sounds when the imagery and the external sound were in synchrony. The experiments reported in the current chapter establish that the impact of speech imagery can linger even after the speech imagery has ceased.

#### 3.2.1 Recalibration

There is enormous variation between people in how speech sounds are produced. This variation is what gives each voice its unique identity and it presents a problem for speech perception: the greater the variability the more uncertain the processing system is about the identity of a sound. The speech-perception system accommodates this variability, in part, by being capable of very rapid adjustment to the speech characteristics of a particular speaker. This rapid adjustment to speech characteristics is known as **recalibration**. It shows that people are able to dynamically adjust the boundaries of their phoneme categories.

In a typical recalibration study ([Baart and Vroomen, 2010], [Sjerps and McQueen, 2010], [Vroomen et al., 2007]), participants are repeatedly exposed to a highly ambiguous speech sound which is ‘pushed’ into one category or another by means of other sources of information. This can be visual information, such as in the McGurk effect, or lexical information, in which only one interpretation of the ambiguous sound results in a real word. For example, given a sound ambiguous between /s/ and /ʃ/, it will be heard as /ʃ/ when placed at the end of “Christma?".
Similarly, a sound ambiguous between /v/ and /b/ will be heard as /v/ if accompanied by video of a facing pronouncing /v/ (and as /b/ when paired with a face pronouncing /b/). After repeated disambiguation, the auditory signal loses some of its ambiguity and is perceived as a relatively clear member of the category into which it has repeatedly been pushed.

This experiment tests whether the imagery effects demonstrated in chapter 2 are able to induce such recalibration. The prediction of this experiment is that after repeatedly being induced to hear an ambiguous sound as /əˈba/ — because of the influence of mouthing or imagining — participants will recalibrate their perceptual boundaries and thus continue to hear the ambiguous sound as /əˈba/ even in the absence of external influences. The equivalent pattern should hold for /əˈva/ (hearing more /əˈva/ after repeated recalibration of an ambiguous sound to /əˈva/).\(^1\)

Recalibration from speech-read visual information was first reported by Bertelson et al. (2003). In that experiment they showed that when video of a face pronouncing /aba/ or /ada/ was matched with audio that was ambiguous between these two sounds, participants heard the sound as belonging to the category indicated by the video. That much is simply the McGurk effect, but interestingly, after repeated exposure to one of these ‘McGurk’ stimuli, participants experienced an after-effect — when the video was removed, participants continued to categorize the ambiguous sound as they had when the video was present. Their phoneme category had recalibrated so that the ambiguous sound became more firmly a member of a particular category.

Van Linden and Vroomen (2007) showed that recalibration from lexical information (also called the Ganong effect) and visual information (the McGurk effect) were very similar. The impact of visual information was stronger, but the general pattern was similar for both. This similarity did not extend to the duration of the after-effect, though. The recalibration due to lexical information appears to last longer (Eisner and McQueen, 2006; Kraljic and Samuel, 2005) than that induced

---

\(^1\)I should point out that repeated exposure to an ambiguous sound does not cause selective adaptation (the effect discussed in Experiment 2-3 below). This was first demonstrated by Sawusch and Pisoni (1976).

\(^2\)speech-reading is often also called “lip-reading.”
by visual information (Vroomen and Baart, 2009).

The similarity between visual influence on speech perception (the McGurk effect) and the speech-imagery effects reported in chapter 2 would argue that imagery, like the McGurk effect, should induce recalibration. However, in the McGurk effect — and in fact in all other recalibration studies — the auditory stimulus and the stimulus inducing the recalibration are plausibly from the same source. In the mouthing/pure imagery paradigm established in chapter 2, the auditory stimulus and the mouthing/imagining are inherently from different sources, which may prevent the effect from causing recalibration.

This is related to the findings of Kraljic et al. (2008). In that experiment people did not show recalibration of an ambiguous /s~/-/ʃ/ sound if the person pronouncing the ambiguous sound had a pen in their mouth at the time and thus the unusual pronunciation could be attributed to that perturbation. Thus, recalibration is quite subtle and seems to take into account the appropriateness of the inducer. The current experiment examines whether speech imagery is an appropriate inducer of sensory recalibration.

3.2.2 Methods

This experiment was roughly modelled on Bertelson et al. (2003).

Participants were repeatedly exposed to an ‘illusory’ percept[^3], after which they were tested on ambiguous sounds to see whether their perceptual boundary was shifted because of exposure to the illusion. There were three illusions used: Video (the McGurk effect), enacted speech imagery (mouthing), and non-enacted (pure) speech imagery.

Thus there were three conditions: **Mouth**, **Imagine**, and **Watch**; and two categories to which participants were exposed in these conditions: /aˈba:/ and /aˈva:/, producing 6 types of block in the experiment, presented in Table 3.1.

[^3]: A sound ambiguous between /aˈba:/ and /aˈva:/ which was disambiguated to one of these categories by visual information (the McGurk effect) or imagery.
Table 3.1: Actions and Target Sounds of Experiment 2-1

<table>
<thead>
<tr>
<th>Action</th>
<th>Mouthed / Imagined / Watched Sound</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mouth</td>
<td>/a'ba/</td>
</tr>
<tr>
<td></td>
<td>/a'va/</td>
</tr>
<tr>
<td>Imagine</td>
<td>/a'ba/</td>
</tr>
<tr>
<td></td>
<td>/a'va/</td>
</tr>
<tr>
<td>Watch</td>
<td>/a'ba/</td>
</tr>
<tr>
<td></td>
<td>/a'va/</td>
</tr>
</tbody>
</table>

In each of these blocks, participants were exposed to a sound ambiguous between /a'ba/ and /a'va/. This ambiguous sound corresponded to each participant’s boundary between /a'ba/ and /a'va/ — the point at which they heard the sound as /a'ba/ 50% of the time — as determined by the pre-test discussed below.

On each exposure, the ambiguous sound was disambiguated to one of these categories (always to the same category during a given block) because of the accompanying mouthing, pure imagery, or video. Participants were exposed to 14 repetitions of this disambiguation, with 150 ms between tokens. The intensity of the sounds ramped up over the first three tokens, reaching c. 60 dB on the third token and staying at that level for the remainder of the exposure phase.

Following the exposure phase there was a 3.5-second pause after which participants categorized 9 audio-only tokens of ambiguous /a'ba/~a'va/ sounds (also presented at c. 60 dB). The sounds presented in the test phase corresponded to each participant’s 50% point in perceptual space between /a'ba/ and /a'va/ ± 8%; meaning the points along the continuum where the participant heard the token as /a'ba/ 42% of the time, 50% of the time and 58% of the time, as determined by the pre-test discussed below. The categorizations of these three sounds were pooled in the data analysis.

As a check that the three illusions (disambiguations) were functioning properly, during the exposure phase participants were asked to hit the spacebar any time any of the illusions (disambiguations) failed to occur. That is, if they were
mouthing/imagining/watching /aˈva/, but still perceived the ambiguous sound as /aˈba/, they were to hit the spacebar (similarly, if they were mouthing/imagining/watching /aˈba/ and perceived the ambiguous sound as /aˈva/, they were to hit the spacebar). A schematic outline of Experiment 2-1 is shown in Figure 3.1.

Figure 3.1: Schematic Outline of Experiment 2-1
Each of the 6 types of block was presented once per cycle of the experiment and there were 9 cycles in the experiment. Potential order and position effects were dealt with by randomizing the order of blocks on each of the 9 cycles (a different randomization on each cycle). The position of response keys was counterbalanced across participants.

As with the experiments reported in chapter 2, a continuum between two clear endpoints (/aˈba/ and /ɑˈva/) was created and each participant’s perceptual boundary between these sounds was determined by a pre-test. Since the middle of the continuum is where the greatest variation in responses is likely to occur (and thus the most information about the boundary), the central steps of the continuum were over-represented in the pre-test. Eleven equally-spaced steps[^4] were used and the number of repetitions of each of these steps is presented in Table 3.2. The perceptual boundary (50% point) between /aˈba/ and /ɑˈva/ as well as the 42% and 58% points were determined by probit analysis.

<table>
<thead>
<tr>
<th>Step on Continuum</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td># of Repetitions</td>
<td>8</td>
<td>8</td>
<td>16</td>
<td>16</td>
<td>16</td>
</tr>
<tr>
<td><strong>Continued …</strong></td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>16</td>
<td>16</td>
<td>16</td>
<td>16</td>
<td>8</td>
</tr>
</tbody>
</table>

[^4]: Equally-spaced steps between the spectra of /aˈba/ and /ɑˈva/ in log-frequency.

Table 3.2: Number of Repetitions of Each Step along the /aˈba/~/aˈba/ Continuum for Experiment 2-1 Pre-Test

This continuum turned out to be strongly skewed. Participants tended to hear far more /aˈba/ than /ɑˈva/. Participants’ perceptual boundaries were often quite near the end of the continuum (so that even at the extreme /ɑˈva/ end of the continuum participants were still perceiving the sounds as /aˈba/). This meant that 5 participants could not be run in the main experiment because their pre-test results showed that their 50% point was beyond the end of the available continuum steps. These participants were paid/given credit for their time but did not take part in the rest of the experiment.

A potential issue with the pre-test was that the stimuli were not presented in
completely random order (as with all other experiments reported in this dissertation), but in pseudo-random order. This was intended to prevent multiple repetitions of exactly the same token, but it means that that participants heard several cycles of the same pseudo-random order during the pre-test rather than a different random order on each cycle. While this pre-test is a good estimate of participants’ boundaries, full randomization may have provided an even better estimate.\footnote{While this is not the first experiment reported in this dissertation, it was the first run. After this experiment I decided that full randomization was more appropriate and thus all other experiments used fully random rather than repetitive pseudo-random presentation in the pre-test.}

3.2.3 Stimuli

A male native speaker of English was recorded saying /a’ba/ and /a’va/. This experiment was the first run, before the issue of sex mismatch between participant and stimulus became apparent. That is why this experiment used a male voice with female participants, while all other experiments were careful to match the sex of stimulus with the sex of participant.

Two of the recorded tokens were selected (one of each category) that were similar in duration, intonation and intensity profiles and which were free of artefacts. A 226-step continuum between these sounds was created using STRAIGHT (Kawahara et al., 1999), the same software used to create the continua for the experiments in chapter 2.

The same male speaker was video-recorded saying /a’ba/ and /a’va/. Two appropriate clips (one for each sound) were selected and used to synchronize with the auditory stimuli, creating McGurk-effect inducing movies in which the audio was typically heard as corresponding to the disyllable visible in the video.

3.2.4 Participants

There were thirty-two female participants (average age = 21 years; SD = 2.4 years), all were either paid or given course credit for their participation.
3.2.5 Results

The dependent measure was the percentage of /a ba/-categorizations across the various conditions. The categorizations from all three target sounds were pooled for analysis.

A three by two repeated-measures ANOVA was performed with “Action” factors: Mouth, Imagine and Watch, and “Sound” levels: /a ba/ and /a va/. A significant main effect of Sound was found [F(1, 31) = 7.920, p = 0.00841] as well as a significant interaction of Action and Sound [F(2, 62) = 3.164, p = 0.04917].

The impact of visual information, in the form of the McGurk effect, has already been established (Baart and Vroomen 2010; Vroomen et al. 2007), and so the Watch condition was used as a filter to exclude from analysis participants who failed to show recalibration. This led to the exclusion of eight participants who did not show a recalibration effect in the Watch condition. A second ANOVA was conducted on this subsetted data, only examining the Mouth and Imagine conditions (Watch being excluded because it was used as a subsetting control-condition). This second ANOVA had “Action” factors: Mouth and Imagine and “Sound” levels: /a ba/ and /a va/. Again, a significant main effect of Sound was found [F(1, 23) = 9.4527, p = 0.00536].
Levels of Recalibration Across Two Experimental Conditions

% of /aba/ categorizations

Mouthed or Imagined disyllable
- /aba/
- /ava/

Imagine   Mouth

Experimental Conditions

Figure 3.2: Results of Experiment 2-1. Standard-error bars are shown.

Planned t-tests were performed comparing the means of the /aba/ and /ava/ levels for both the Mouth and Imagine factors. Both of these tests were significant (p < 0.05).

As a check that the illusions were functioning properly, participants were asked, in the exposure phase, to hit the spacebar whenever one of the illusions failed (that is, whenever one of the actions failed to disambiguate the ambiguous sound properly). A repeated-measures ANOVA was performed on this measurement of the strength of disambiguation across the three conditions. The dependent measure was the % of exposures in which the illusion (disambiguation) occurred. This ANOVA found a main effect of Sound [F(1, 31) = 18.17594, p < 0.001], a main effect of Action [F(2, 62) = 18.89557, p < 0.001] and a significant interaction be-

---

6 Participants actually responded when the illusion failed, but for clarity in the graphs that has been recoded as proportion of exposures in which the illusion succeeded.
between Sound and Action \[F(2, 62) = 17.84403, p < 0.001\]. These results are shown in Figure 3.3.

These results show that the illusions were functioning properly for all three conditions of the experiment. Furthermore all contrasts were significant in follow-up pairwise comparisons (using the Holm-Bonferroni correction), showing that the illusions were stronger (for both /aʼba/ and /aʼva/) in the Watch condition than the Mouth condition, and stronger in the Mouth condition than in the Imagine condition.
3.2.6 Discussion

These results show that imagery, in both enacted and non-enacted forms, induces recalibration. Thus, the perceptual effect demonstrated in chapter 2 does not just influence simultaneous perception but induces a perceptual shift that lingers after the imagery has stopped.

It is important to point out that the number of /a’ba/ categorizations in the test-phase, while significantly different between /a’ba/ and /a’va/ contexts, in all conditions always remained above 50%. This indicates that even when people were being induced to hear /a’va/ in the exposure phase they heard more than half of the sounds as /a’ba/ in the test phase afterwards. This suggests that there was little recalibration induced by imagining or mouthing /a’va/ while there was quite a lot of recalibration after imagining or mouthing /a’ba/. This is perhaps also an indication that the target sounds were too skewed — being heard overwhelmingly as /a’ba/, despite the use of a pre-test to calibrate the sounds to each participant’s perceptual boundary. This may be related to the type of randomization of stimuli used in the pre-test problem (this is discussed above).

On a different note, a greater impact on perception of mouthing over pure imagery was clearly shown in chapter 2 and so it is reassuring to see that finding replicated here (as shown by the analysis of participants’ experience of the illusion during the exposure phase of the experiment — displayed in Figure 3.3). This examination of the strength of imagining vs. mouthing (vs. watching) in disambiguating a sound is very similar to what was reported in chapter 2, including the stronger disambiguating influence of mouthing over imagining. In chapter 2, I suggested that if participants were told of the perceptual impact of mouthing/imaging, and thus did not consciously attempt to segregate the heard sound from what they were mouthing/imaging, the impact of mouthing/imaging might be larger. This seems to be the case here, where participants were told of the effects and the proportion of tokens in which the ambiguous sound was disambiguated was c. 91% for imagining and c. 94.5% for mouthing (higher than in the experiments reported in chapter 2).

7 The same pattern was seen for the Watch condition in the pre-subsetted data.
3.3 Experiment 2-2

Experiment 2-1 established that recalibration can indeed occur when a sound is disambiguated by either enacted or pure speech imagery. There were two issues with this experiment, though. First, the voice used for the stimuli was male, while the participants were female. It is unclear whether a sex disagreement between the participant and the stimuli would be problematic, but as all other experiments in this dissertation matched sex of participant with sex of stimuli, this experiment was replicated below using a female voice for the stimuli (to match the sex of participants, who were all female). A second issue is the skew in the continuum used for the stimuli. Participants heard these sounds overwhelmingly as /a’ba/ in Experiment 2-1. This issue was also addressed in Experiment 2-2.

3.3.1 Methods

The structure of this experiment was identical to Experiment 2-1, except the following changes:

- Stimuli were of a female voice.
- The pre-test calibration of participants’ perceptual boundaries was extended for greater precision, and full randomization of tokens was used.
- The number of exposure tokens was increased to 17 (from 14 in Experiment 2-1).
- The timing between tokens in the exposure phase was reduced to 120 ms.
- More tokens were presented in each test phase (15 tokens in each test phase vs. 9 in Experiment 2-1).
- Rather than randomize the order of blocks, there was tight counterbalancing of block order across participants.
- There were 6 cycles (vs. 9 in Experiment 2-1).

On each block, participants were exposed to 17 tokens of a sound ambiguous between /a’ba/ and /a’va/, with 120 ms between exposures. The first token in
this exposure phase was played at half volume; all other sounds were played at c. 62 dB. On each token the ambiguous sound was disambiguated by means of exposure to one of three ‘illusory’ influences: Mouthing, pure speech imagery and the McGurk effect (Mouth, Imagine, Watch). These three influences all have the same effect, causing the ambiguous sound to be heard as matching the content of what is mouthed/imagined/seen in the video.

After these 17 exposures there was a 2-second pause following which participants categorized 15 tokens of sounds ambiguous between /aˈba/ and /aˈva/ (presented at c. 62 dB), in a forced choice between “aba” and “ava”. These target sounds were three maximally ambiguous sounds between /aˈba/ and /aˈva/, corresponding to each participant’s 42.5%, 50%, 57.5% points along the stimulus continuum as discussed in subsection 3.3.2.

The order of blocks was strictly counterbalanced across participants, as was the position of the response keys. One half of participants started with disambiguation to the /aˈba/ category and the other half with disambiguation to the /aˈva/ category. The category to which participants disambiguated alternated back and forth in strict succession throughout the experiment. Each participant repeated the same order of conditions six times (six cycles).

As with the other experiments presented in this dissertation, participants’ perceptual boundaries were determined via pre-test and probit analysis. The pre-test presented participants with more tokens from the centre of the continuum than from the ends. The distribution of number of tokens presented for each step of the continuum is presented in Table 3.3.
<table>
<thead>
<tr>
<th>Step on Continuum</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td># of Repetitions</td>
<td>5</td>
<td>5</td>
<td>10</td>
<td>10</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Continued ...</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Continued ...</td>
<td>12</td>
<td>13</td>
<td>14</td>
<td>15</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>15</td>
<td>10</td>
<td>10</td>
<td>5</td>
</tr>
</tbody>
</table>

**Table 3.3:** Number of Repetitions of Each Step along the /a'ba/~/a'ba/ Continuum for Experiment 2-2 Pre-Test
3.3.2 Stimuli

A continuum from /a'va/ to /a'ba/ was created using STRAIGHT. Shortly before this experiment was run, a new version of STRAIGHT was released which allowed for finely-grained continua to be generated automatically. Thus, a 2001-step con-
A female native English speaker was recorded saying /aˈvə/ and /aˈba/. Tokens that were similar in length and intonation and free of artefacts were chosen as the basis of morphing. The sounds were 609 ms long (the duration of all tokens was identical). As described above, only three tokens from this continuum were used for each participant. The three target tokens were different for each participant and corresponded to the steps on the continuum closest to each participant’s perceptual boundary ± 7.5% (i.e., 42.5%, 50%, 57.5%), as determined by a probit analysis of pre-test data (described above). Tokens were taken from near the category boundary so that they would be maximally ambiguous.

A different female native English speaker was video-recorded saying the target sounds /aˈvə/ and /aˈba/. A different speaker was used so that there would be a slight mismatch in identity between audio and video for the Watch condition, to parallel the mismatch in identity in the Mouth and Imagine conditions.

Video tokens that closely matched the timing of the /aˈvə~/~aˈba/ continuum were selected and used to create the ‘McGurk’ stimuli in the Watch condition.

A different female speaker was used in this experiment from the person used in chapter 2. This was because the boundary between /aˈvə/ and /aˈba/ in the continuum used in chapter 2 was very variable (across participants).

### 3.3.3 Participants

There were 24 female participants (average age = 20.7 years; SD = 2 years). All participants were either paid or given course credit for their participation.

### 3.3.4 Results

A three by two repeated-measures ANOVA was performed with “Action” factors: Mouth, Imagine and Watch, and “Sound” levels: /aˈba/ and /aˈvə/. There was a significant interaction of Action and Sound \(F(2, 46) = 5.477, p = 0.00735\).

As with Experiment 2-1, the Watch condition was used as a filter to exclude from analysis participants who failed to show recalibration. This led to the exclusion of six participants who did not show a recalibration effect in the Watch condition. A second ANOVA was conducted on this subsetted data, only exam-
ining the Mouth and Imagine conditions (Watch being excluded because it was used as a subsetting control condition). This second ANOVA had “Action” factors: Mouth and Imagine and “Sound” levels: /a'ba/ and /a'va/. A significant main effect of Sound was found \[F(1, 17) = 5.5444, p = 0.03082\].

**Levels of Recalibration Across Two Experimental Conditions**

![Image](image.png)

**Experimental Conditions**

Figure 3.5: Results of Experiment 2-2. Standard-error bars are shown.

Planned t-tests were performed comparing the means of the /a'ba/ and /a'va/ levels for both the Mouth and Imagine factors. These tests showed significant recalibration (p < 0.05) in the Imagine condition, but not in the Mouth condition (p = 0.1605).

These results are similar to those found in Experiment 2-1, despite the fact that the stimuli were completely different (a female voice in this experiment, a male voice in Experiment 2-1), and the procedures were slightly altered as well. While the Mouth condition did not reach significance in this experiment, in both
this experiment and Experiment 2-1, the effect of pure imagery induced significant recalibration.

As with Experiment 2-1, a check was performed on the ability of video versus mouthing versus pure imagery to disambiguate the ambiguous /a'ba/~/a'va/ sounds.

A repeated-measures ANOVA was performed on this measurement of the strength of disambiguation across the three conditions. The dependent measure was the % of exposures in which the illusion (disambiguation) occurred. This ANOVA found a main effect of Action \[F(2, 46) = 8.897, p = 0.00054\]. These results are shown in Figure 3.6.

---

**Figure 3.6:** Strength of Disambiguation across Three Conditions in Recalibration Experiment 2-2. Standard-error bars are shown.

---

8 As with Experiment 2-1, participants actually responded when the illusion failed, but for clarity in the graphs that has been recoded as proportion of exposures in which the illusion succeeded.
These results are again very similar to those reported in Experiment 2-1 (and to the experiments reported in chapter 2). They show that the degree of disambiguation is weakest in the Imagine condition, stronger in the Mouthing condition and strongest in the Watch conditions. This again replicates the findings of chapter 2 and of Experiment 2-1, in showing that mouthing has a stronger effect on the perception of external speech than pure imagery does. This shows that the failure of mouthing to induce recalibration in this experiment (while pure imagery does) cannot be attributed to the strength of disambiguation in mouthing vs. pure imagery.

3.3.5 Discussion

This second experiment replicates (for one condition) the recalibration found in Experiment 2-1; showing that the effects reported in chapter 2 can continue even after the imagery has ceased. Unfortunately, this experiment failed to replicate an effect for the Mouth condition.

In contrast to Experiment 2-1, in this experiment, the sex of the stimuli matched the sex of the participants. The results were very similar between these experiments and so it seems that, at least for recalibration, a sex-mismatch is not critical.

The failure to replicate the recalibration effect for the Mouth condition in this experiment was disappointing. It may simply be due to the vagaries of sampling or a lack of experimental power, and it is always dangerous to draw conclusions from a null result. With those caveats in mind, I offer the following possibility merely as speculation.

The weaker effect of enacted speech imagery (mouthing) in this experiment\(^9\) may be due to a greater degree of corollary discharge in the mouthing condition. This suggestion is purely post-hoc speculation. As discussed in section 1.3, it has been claimed that enacted speech imagery is more like an external speech sound than non-enacted speech imagery. In this dissertation I have argued that enacted speech imagery may show a greater degree of forward model engagement (and thus produce a stronger corollary discharge signal — either a stronger signal or

\(^9\)Which was not statistically significant, and so the rest of this discussion is speculation based on the contingent possibility that future research will show that the failure to replicate an effect of mouthing in this experiment is due to mouthing being less efficient than pure imagery in inducing recalibration.
more signals from more forward models). If this is correct, then we would predict that the effects of corollary discharge would be stronger for enacted than for pure speech imagery.

Perhaps the presence of corollary discharge interferes with the process of recalibration. This interference could be explained as being due to the *sensory attenuation* function of corollary discharge (as discussed in subsection 1.5.5). If the repeatedly presented sound in this experimental paradigm has its impact attenuated by corollary discharge on each exposure, perhaps the remapping of the sound to a phonemic category is also attenuated. This is a very vague idea and would be dependent on the details of how the remapping of boundaries is accomplished in recalibration (which is unknown) and exactly how the sensory attenuation of corollary discharge is accomplished (which is unknown). This proposal should be treated as speculation. Furthermore, the difference between enacted and non-enacted speech imagery was not significant.

If this interpretation is correct, that would mean that corollary discharge is both inducing the ambiguous sound to be heard as a particular percept and at the same time attenuating the impact of that percept. This may seem odd, but it is not at all a contradiction. In fact, I believe that this description, in general terms, accurately captures how corollary discharge functions in speech imagery. If the channelling description of corollary discharge that I proposed in subsection 1.5.5 is correct, we would predict such a scenario. Corollary discharge is an anticipation of hearing a particular sound, and so when a sound occurs that is ambiguous but still compatible with expectation, that sound is treated as corresponding to expectation. Furthermore, since the purpose of corollary discharge (under the present analysis) is to channel the expected sound into a ‘self-caused’ perceptual stream, it is unsurprising that the impact of this sound is different from the impact of sounds that are treated as ‘externally caused’. Under this analysis, the segregation of the self-caused stream would lead to an attenuation of recalibration; though I have no theory as to why segregation should have that effect in this case. Again, the proposal that corollary discharge is responsible for the failure to replicate recalibration in the *Mouthing* condition of this experiment is simply speculation.

A final point is that both this experiment and Experiment 2-1 found that recalibration seems to be carried purely by a shift in the perceptual boundary for
In Experiment 2-1, /aˈba/ was heard in the test phase more than 50% of the time in all conditions (whether participants had been caused to disambiguate the ambiguous sounds to /aˈba/ or /aˈva/ in the exposure phase). Nearly the same thing was found in this experiment, where test tokens were categorized as /aˈba/ more than 50% of the time in the Mouth condition after disambiguation to either /aˈba/ or /aˈva/ in the exposure phase. Nearly the same pattern was seen in the Imagine condition, though recalibration to /aˈva/ did cause the number of /aˈba/ categorizations to fall slightly below 50% in the test phase. This strongly suggests that recalibration is being carried almost entirely by disambiguation to /aˈba/ in the exposure phases and little or no recalibration from disambiguation to /aˈva/ is occurring in any condition.

While these results suggest that the continuum used in this experiment was less skewed than that used in Experiment 2-1, there was still a strong bias towards hearing /aˈba/.

On a different note, looking at the data on how often participants experienced the illusion during the exposure phase we can see that, as discussed above, these results again confirm the suggestion that the impact of mouthing/imagery is stronger when participants are told of the potential impact and so do not attempt to segregate the heard sound from what they are mouthing/imagining. c. 95.5% of tokens were disambiguated by imagining and c. 97.5% were disambiguated by mouthing; much higher rates than in the experiments reported in chapter 2, in which participants were not warned of the impact of mouthing/imagining and so presumably tried to segregate the content of mouthing/imagining from the presented sound and in doing so, presumably ‘fought against’ the effect.

### 3.4 Experiment 2-3

This experiment examines the interaction of selective adaptation with enacted speech imagery (mouthing).

Selective adaptation refers to the observation that after repeated exposure to a speech category, participants are less likely to perceive subsequent speech sounds as belonging to that category. For example, repeated exposure to the sound /ba/ will cause people to categorize a sound ambiguous between /ba/ and /da/ as be-
longing to the /da/ category. This effect has often (though not without controversy) been attributed to exposure causing ‘fatigue’ in the auditory system (Eimas and Corbit 1973).

The current experiment repeatedly exposed participants to clear instances of either /aʼba/ or /aʼva/ (adapting sounds) and then tested them on ambiguous tokens from a /aʼba/∼/aʼva/ continuum (test sounds). In two of the experiment’s three conditions participants either mouthed along with the adapting sounds or saw a face pronouncing the adapting sound[10] In the third condition, participants merely listened to the adapting sound with no other manipulation. Thus, in this experiment, there are conditions examining visual and enacted speech-imagery influences but there is no ‘pure’ speech-imagery condition.[11]

Another reason for examining the intersection of speech imagery and selective adaptation is that selective adaptation, as discussed below, is potentially the result of fatiguing part of the auditory perceptual system.

For example, as discussed in subsection 1.5.5, it has been shown that crickets use corollary discharge to prevent their auditory system from being overwhelmed (and temporarily deafened) by their own singing (Crapse and Sommer, 2008). Without corollary discharge, the ability of a cricket’s ears to detect sounds out in the environment would be severely compromised. If we view the auditory ‘fatigue’ caused by sensory adaptation as a less severe case of the auditory system being overwhelmed, then it is possible that corollary discharge serves a similar function in humans. Perhaps fatigue from self-caused sounds is attenuated by corollary discharge. This would serve the function of preventing distortions in the perception of externally-caused sounds which may be produced when the auditory system is fatigued by self-caused sounds.

If this is correct, then corollary discharge should attenuate selective adaptation. Thus, the tentative prediction of this experiment is that in the presence of mouthing (and thus in the presence of corollary discharge), participants should show less selective adaptation than in other conditions. To preview the results, while this

---

[10] In both of these conditions, what participants mouthed/saw matched what they heard. Thus if they were listening to /aʼba/, they mouthed/saw /aʼba/ and similarly with /aʼva/.

[11] This experiment required the addition of a ‘listening only’ condition, and as the experiment would be too long with four experimental conditions, for pragmatic reasons only one form of imagery could be tested, and so mouthing (enacted speech imagery) was chosen.
experiment did replicate the usual selective adaptation effects, no impact of speech imagery was found.

3.4.1 Selective Adaptation

Selective adaptation is in some ways the mirror image of the recalibration effect discussed above. In a selective adaptation paradigm, participants are repeatedly exposed to a sound but (unlike the recalibration paradigm) the sound is a clear, unambiguous representative of its category. After these repeated exposures, participants typically show a shift in their category boundary. Ambiguous sounds are less likely to be perceived as belonging to the category to which participants were just exposed. For example, Eimas and Corbit (1973) showed that after repeated exposure to /ba/, participants categorized fewer sounds from a /ba/-/pa/ continuum as being /ba/ (and vice versa after exposure to /pa/). This is the inverse of the recalibration paradigm in which repeated exposure to a category induces more reports of hearing that category in subsequent testing.

As mentioned above, selective adaptation was initially believed to be the result of fatiguing some sort of linguistic ‘feature-detector’. Later work suggested that there is more to the effect than simple fatigue. Diehl et al. (1978) and Diehl (1981) have proposed that the shift in category boundaries seen in selective adaptation is not a matter of fatiguing but merely the result of ‘response contrast’. That is, participants use the adapting stimulus as a reference point for the category. When they hear a subsequent, less canonical, token they assume that the contrast between the canonical (adapting) token and the ambiguous (test) token is large enough to represent a shift in category. Thus, they report that the test token is of a different category from the adapting token.

However, more recent work has shown that the impact of adaptation can be dissociated from the perceived category of the sound, and thus is less reasonably the result of response contrast (since there is no contrast in response in these cases). For example, Saldana and Rosenblum (1994) repeatedly exposed participants with

---

12 The first demonstration of selective adaptation.
13 They used an aspirated [pʰ] as the realization of /p/.
14 Though the sound is ambiguous and is perceived as belonging to the category through some non-auditory source of extra information, like the McGurk effect.
a ‘McGurk’-style stimulus in which the audio and video disagreed — the audio was
of /ba/ while the video was of a face pronouncing /va/, causing participants to per-
ceive the sound as /va/. Under this set-up, even though participants perceived the
adapting sound as /va/, they showed levels of adaptation equivalent to when they
were exposed to an audio /ba/ unaccompanied by video. Thus the video had no
effect on the pattern of adaptation: The presence of the video changed how partic-
ipants perceived the exposure sounds (in the presence of the video, they perceived
an audio /ba/ as /va/), but this change in the perceived category of the adaptor did
not change the pattern of adaptation.

This result shows that adaptation is not dependent on the perceived category
of the adapting stimulus, only on its lower-level acoustic or auditory aspects. This
argues against the ‘response contrast’ theory since, in Šaldana and Rosenblum’s
 experiment, participants are reporting fewer percepts of a category that they were
not aware they had been exposed to. Furthermore Samuel and Kat (1998) showed
that selective adaptation is not dependent on attention, since levels of adaptation
are not affected by distractor tasks. This suggests that selective adaptation is a
fairly ‘low-level’ phenomenon, and cannot be attributed to contrast, at least at the
conscious level. The possibility remains that the contrast effect is not operating at
a conscious, but at a much lower level.

As further evidence for the lack of influence of conscious factors on selective
adaptation, Vroomen et al. (2004) showed that selective adaptation cannot be due
to strategic responding. They performed an ABX test in which subjects had to
distinguish an ambiguous auditory token from a clear auditory token of the same
category, with both tokens accompanied by video of a face pronouncing this cate-
gory. Subjects were very poor at this task (52% correct with chance being 50%).
This suggests that they could not consciously distinguish the audio in the “ambigu-
ous audio” blocks from that in the “clear audio” blocks, and yet these different
sounds drove different effects. The ambiguous audio causes recalibration while
the clear audio causes selective adaptation. This is strong evidence that conscious
factors are not at play in either recalibration or selective adaptation.

Others have argued that selective adaptation is not specific to speech but is
the result of domain-general auditory processing. This claim has been supported
by experiments showing that selective adaptation is dependent on acoustic overlap
between the adapting and test sounds (e.g., Ades, 1977; Cooper, 1974). The debate about whether selective adaptation is domain general or particular to speech sounds is not directly relevant to this experiment.

This experiment examines whether corollary discharge (assumed to be present when people mouth sounds) will attenuate selective adaptation. There has already been some work done on the interaction between self-caused sounds and selective adaptation. For example, Cooper et al. (1976, 1975) tried to induce sensory adaptation to mouthed and whispered speech (in the whisper condition white noise was used to block normal sensory feedback). They were marginally successful; however, subsequent attempts at replication have not found clear evidence of an effect (e.g., Summerfield et al., 1980).15

The lack of consensus in these experiments as to whether self-produced speech can induce selective adaptation suggests a possibility: Perhaps self-produced speech is accompanied by corollary discharge and this corollary discharge weakens selective adaptation (as outlined above), making it difficult for experimenters to find selective adaptation to self-produced speech. It is this possibility that led to the current experiment, which tests the impact of corollary discharge (which I hypothesize to be a constitutive part of speech imagery) on selective adaptation.

### 3.4.2 Methods

This experiment is, in most respects, identical in design to Experiment 2-2 above.

The primary design difference between this experiment and Experiment 2-2 above, is that in this experiment participants were not presented with ambiguous sounds in the exposure phase, but instead heard clear sounds from the ends of the continuum. Furthermore, instead of an Imagine condition, there was a Hear condition which acted as a baseline, replicating the basic effect of selective adaptation.

As with Experiments 2-1 and 2-2, a video condition was included. However, in this instance the video cannot be considered a ‘McGurk’ stimulus: the category seen in the video matched that of the audio (both audio and video were of /aˈba/ or both were of /aˈva/). A McGurk stimulus would need the video to alter the perception of the audio in some way, whereas here the audio and video reinforced

---

15 Andrew Lotto (personal communication) is pursuing a research program that hypothesizes that auditory imagery should induce such effects and is currently testing that prediction.
each other. There were two reasons for including this video condition. First, it allowed comparison between the effects of mouthing and watching a face, and second it allowed for a check that recalibration (as discussed above) was not taking place. If the auditory stimuli in the exposure phase were not properly unambiguous, then the presence of the video would induce recalibration (as shown in both experiments above) — which would push the results in the opposite direction from selective adaptation. As shown in subsection 3.4.5, there was no significant difference between the video (Watch) condition and the control (Hear) condition, showing that there was no significant recalibration occurring in this experiment.

Thus, there were three conditions in this experiment: **Watch**, **Mouth**, and **Hear**, and each condition had two levels: /a'ba/ and /a'va/; that is, participants were adapted to both /a'ba/ and /a'va/ in each condition. This structure is shown in Table 3.4.

<table>
<thead>
<tr>
<th>Action</th>
<th>Mouthed / Watched / Heard Sound</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mouth</td>
<td>/a'ba/</td>
</tr>
<tr>
<td></td>
<td>/a'va/</td>
</tr>
<tr>
<td>Watch</td>
<td>/a'ba/</td>
</tr>
<tr>
<td></td>
<td>/a'va/</td>
</tr>
<tr>
<td>Hear</td>
<td>/a'ba/</td>
</tr>
<tr>
<td></td>
<td>/a'va/</td>
</tr>
</tbody>
</table>

**Table 3.4:** Actions and Target Sounds of Experiment 2-3

On each block, participants were exposed to 17 clear examples of /a'ba/ or /a'va/, with 120 ms between exposures (always the same category within a block, alternating back and forth between /a'ba/ and /a'va/ between blocks). After these 17 exposures, there was a two-second pause after which participants categorized 15 tokens of sounds ambiguous between /a'ba/ and /a'va/, in a forced choice between “aba” and “ava”. In the **Mouth** condition, participants mouthed along with the exposure tokens, mouthing /a'ba/ when the computer presented /a'ba/ and mouthing /a'va/ when the computer presented /a'va/. In the **Watch** condition,
a video of a face pronouncing the same sound as the audio appeared in synchrony with the audio exposures. In the **Hear** condition, participants merely heard the unambiguous adapting sounds with no other activity.

It is important to point out that this was not a “McGurk”-style set-up in which the sound presented to participants in the exposure phase differed from what they were seeing in the video or differed from the content of their mouthing. In all instances the content of the action (what was mouthed or watched) matched the audio. The outline of the experiment is presented in Figure 3.7.

As with Experiment 2-2, the order of conditions and levels was strictly counter-balanced across participants, as was the side of the response-buttons. Each participant started with being adapted to either /aˈba/ or /aˈva/ (the starting sound was counterbalanced across participants) and they alternated between adapting sounds from block to block in strict succession throughout the rest of the experiment.

As with Experiment 2-2, there was a pre-test to determine participants’ perceptual boundaries between /aˈba/ and /aˈva/. Exactly the same procedure was used as in Experiment 2-2. Based on the results of this pre-test, three points were selected from the continuum, corresponding to each participant’s perceptual boundary ± 7.5% (i.e., 42.5%, 50%, 57.5%). These are the maximally ambiguous sounds for each participant and were used in the test phase to detect boundary shifts after adaptation. These are the same points used in Experiment 2-2 above. The results for these three sounds were pooled for analysis.

In addition to these three ‘test’ sounds, two unambiguous sounds, corresponding to each participant’s 0.5% and 99.5% points along the continuum (the points on the continuum where the participants would hear the sound as /aˈba/ 0.5% of the time and 99.5% of the time) were also determined via probit analysis. These two clear sounds were used to induce adaptation in the exposure phase.

3.4.3 **Stimuli**

The continuum from /aˈba/ to /aˈva/ was the same as used in Experiment 2-2 above.

In addition to the three ambiguous tokens (from the centre of the continuum) used in Experiment 2-2, two clear tokens (from the ends of the continuum) were
selected as well. Thus, there were 5 sounds selected from the continuum for each participant, corresponding to that participant’s 0.05%, 42.5%, 50%, 57.5%, and 99.5% points on the continuum. The three ambiguous test-tokens (42.5%, 50%, 57.5%) were pooled for data analysis.

The video used for the Watch condition was the same as used in Experiment
2-2.

3.4.4 Participants

As with Experiment 2-2, there were 24 female participants (average age = 20.7 years; SD = 2.26 years). All participants were either paid or given course credit for their participation.

3.4.5 Results

A three by two repeated-measures ANOVA was performed with “Action” factors: Mouth, Watch and Hear, and “Sound” levels: /a’ba/ and /a’va/. A main effect of Sound was found [F(1, 23) 84.292, p < 0.001]. No interaction was found. Unlike Experiments 2-1 and 2-2, the Watch condition was not a simple replication here, and thus was not used to filter the data. The crucial comparison was between the /a’ba/ and /a’va/ sounds for each of the three Action factors and so three planned t-tests were conducted. These t-tests were all significant (p < 0.001), showing that adaptation occurred in all conditions.
These results replicate the typical selective-adaptation effect for all three conditions. That is, participants heard less of a category after repeated exposure to that category in all conditions. Crucially, however, there was no interaction between this selective adaptation and the Action being performed by the participant. The level of adaptation was not at all affected by mouthing or by the presence of video.

3.4.6 Discussion
The lack of difference between the Watch and Hear conditions validates the choice of stimuli — if the stimuli had been at all ambiguous, there should be some level of recalibration mitigating the selective adaptation and thus weakening the effect in the Watch condition.

The lack of difference between the Mouth condition and the other conditions
could be explained in several different ways.

First, since this experiment is based on the claim that corollary discharge is present when people mouth speech sounds, the lack of impact of mouthing may indicate that this claim is wrong: perhaps corollary discharge is simply not present in mouthing. This issue is taken up in chapter 4 where strong evidence is shown that corollary discharge is present when people mouth speech sounds.

Second, it is possible that the characterization of selective adaptation as due to ‘fatigue’ is simply incorrect. Perhaps selective adaptation is, as Diehl et al. (1978) and Diehl (1981) have argued, a matter of (subconscious) contrast effects. If this is the case, the lack of interaction with corollary discharge is unsurprising.

Another possibility is that corollary discharge and selective adaptation operate at different levels of processing. If adaptation is a comparatively lower processing-level phenomenon, then the type of corollary discharge that occurs in mouthing may not occur at a low enough level to alter selective adaptation. As discussed in subsection 5.1.2, the point in the chain of perceptual processing at which corollary discharge has its effects is unknown. Perhaps corollary discharge interacts with auditory perception at a stage of processing after adaptation has occurred, in which case a lack of interaction is to be expected.

Alternatively, the two effects (adaptation and corollary discharge) may not be occurring at different levels but may be independent of each other. Remember, corollary discharge is, I argue, probably not a lowering of signal intensity, but more a channelling of stimuli into separate processing streams based on whether the stimulus is considered ‘self’ or ‘other’ (see subsection 1.5.5). Perhaps adaptation is not sensitive to the distinction between self-caused and externally caused, and thus is not affected by the presence of corollary discharge.

3.5 Conclusion

The experiments reported in this chapter extend the findings of chapter 2.

The first two experiments showed that the speech-imagery effects reported in chapter 2 can linger even after participants have stopped engaging in speech imagery. These experiments demonstrated that speech imagery induces recalibration, which is a shift in perceptual boundaries caused by repeated exposure to an illu-
sion. Recalibration causes the illusion to persist even after the cause of the illusion has been removed. In Experiments 2-1 and 2-2 it was shown that after repeated exposure to a speech imagery induced illusion, participants continued to categorize sounds as if the speech imagery were still present. This is very similar to the impact of visual information in the McGurk effect.

However, while pure speech-imagery induced significant recalibration in both experiments, the effect of enacted speech imagery (mouthing) failed to reach significance in Experiment 2-2.

The third experiment was an examination of the interaction between speech imagery and selective adaptation. The purpose was to test whether the presence of corollary discharge (postulated to be present when people mouth speech sounds) would attenuate selective adaptation. While this experiment did succeed in replicating selective adaptation, it found no interaction between speech imagery and selective adaptation.
Chapter 4

Attenuation of a Context Effect by Corollary Discharge

Perception is basically a controlled hallucination process — Ramesh Jain (as quoted in Grush, 1995)

4.1 Introduction

This chapter reports on three experiments that address a central claim of this dissertation, that the sensory content of speech imagery is constituted by corollary discharge. In chapter 2 it was demonstrated that speech imagery includes sensory content and that the representation/processing of speech imagery is sufficiently similar to that of external speech that speech imagery can alter the perception of external speech. This chapter demonstrates that speech imagery induces the hallmark feature of corollary discharge: sensory attenuation.

Sensory attenuation is discussed in detail in subsection 1.5.5 but I will give a brief summary of the topic here.

Self-caused sensations are constantly impinging on our sense organs. We walk and create audible footsteps, we breathe and can hear the passage of air (particularly when we have a cold), and most pertinently, we talk and hear the sound of our own voice. This constant stream of self-produced sounds is unavoidable if we are
going to act in the world, but it is also a source of potential confusion: Was that my footstep or that of a predator? Did she say 'tall' when I said 'top' or was she saying all and the /t/ was mine? If we are to survive and be evolutionarily successful, making the distinction between self-caused and externally caused sensations is critical.

Unfortunately, externally caused sensations do not come with a ‘tag’ proclaiming their provenance; so instead we have to tag the self-caused sensations ourselves, which is what we do. Throughout the animal kingdom (Crapse and Sommer, 2008) animals use a component of their motor system called a forward model (discussed in subsection 1.5.1) to predict the sensory consequences of their actions and send this prediction to sensory areas to act as a tag of self-caused sensations, thus preventing the confusion between self-caused and externally caused that would otherwise result. The signal that is sent from the motor system to sensory areas is called corollary discharge.

The tagging of self-caused sensations by corollary discharge is typically demonstrated experimentally by showing that an animal’s response to self-caused sensations (either behaviourally or in terms of brain activity) is attenuated in comparison to its response to equivalent sensations that are externally caused. This is called sensory attenuation and is represented graphically in Figure 4.1.

Sensory attenuation has been found in multiple modalities (hearing, touch, seeing, balance, electoreception) and in a wide range of animals, from those with a simple nervous system, such as a flatworm, to those with an elaborate nervous system such as humans (and other primates).

For the sake of simplicity, sensory attenuation is often described as corollary discharge cancelling out incoming sensations, much like laying a photographic negative on its positive. It is this simplistic ‘cancellation’ model of corollary discharge that is shown in Figure 4.1 but this is an over-simplification. It is not the case that we cannot perceive self-caused sensations; they are not simply “cancelled out”. I suggest that it is likely that sensory attenuation is a matter of channelling incoming sensations into different processing streams based on the source of the sensation (self vs. other) in a manner analogous to auditory scene analysis. Auditory scene analysis is the process by which the auditory system tries to group components of the acoustic signal into different streams based on whether they were
caused by the same event (Bregman, 1990). Under this interpretation, corollary discharge does not simply attenuate a self-caused sensation, but attenuates the impact that a self-caused sensation has, by isolating the self-caused sensation into a segregated processing stream. This kind of perceptual streaming has been demonstrated several times. Bregman (1990) showed that normally discordant pitch intervals are not perceived as discordant when the participants are ‘tricked’ into processing the pitches as belonging to separate streams. Similarly, the impact of a tone on speech perception is weakened when the tone is segregated more strongly into a separate perceptual stream (Ciocca and Bregman, 1989).

This chapter reports on three experiments that show that speech imagery in-
duces sensory attenuation and thus that speech imagery very likely involves corollary discharge.

Sensory attenuation is demonstrated in these experiments by means of a well-known context effect: when a target sound which is ambiguous between /dA/ and /gA/ is preceded by the context /Aô/, people tend to hear the target as /dA/, but if the target is preceded by /al/ people tend to hear it as /gA/. This is the ‘Mann effect’\(^1\) and is discussed in detail in subsection 4.1.1. For now it is enough to know that the perception of a target sound is influenced by the immediately preceding context.

In the first two experiments reported below I demonstrate that this context effect is attenuated when participants engage in enacted speech imagery (mouthing) in synchrony with the context sound. It might be argued that the attenuation is due to the distraction of mouthing, but crucially the attenuation is strongest when the sound that participants are mouthing matches the context sound they are hearing (for example mouthing /al/ while hearing /al/), but the attenuation is weaker when participants mouth a sound that does not match the context (for example, mouthing /al/ while hearing /Aô/). The condition in which participants mouth a different sound from what they are hearing is presumably more distracting than when they mouth the same thing, but it is the condition in which they mouth the same sound that shows a significantly attenuated Mann effect.

The third experiment in this paper demonstrates that this attenuation of the Mann effect is found in external speech as well.

In the experiments reported in this chapter, only enacted speech imagery is tested. In other words, participants only “mouthed” sounds in these experiments and did not engage in “pure” imagery. This was a pragmatic choice, since effects are predicted to be stronger (and so easier to detect) in enacted speech imagery. This issue is discussed in subsection 4.3.6.

\(^1\)Named for the researcher who first demonstrated it, Mann (1980).
4.1.1 The Mann Effect

The Mann effect is one of several ‘context’ effects, in which the perception of a target sound is influenced by the surrounding context. Mann (1980) showed that an ambiguous /da/∼/ga/ syllable is more often heard as /da/ after /a/ and more often as /ga/ after /a/. This effect has been replicated several times (Fowler, 2006; Fowler et al., 2000; Holt, 1999, 2006; Holt and Lotto, 2002; Lotto et al., 1997, 2003) and is the subject of some controversy as to its cause. There are two main proposals as to why the context has such an effect: Spectral Contrast and Compensation for Coarticulation. I will discuss each of these in turn. I should emphasize I do not take a stance on which of these alternative explanations is correct (perhaps both are). I use the Mann effect as a behavioural measure in the experiments reported below and the exact cause of the effect does not alter the claims made in this chapter.

4.1.2 The Spectral Contrast Explanation of the Mann Effect

Holt (1999) and Lotto and Holt (2006) claim that the Mann effect is the result of spectral contrast. Spectral contrast is a ubiquitous aspect of auditory perception, it refers to the fact that differences between adjacent sounds are exaggerated. For example a sound at 100 Hz may be perceived as being lower than normal if it follows a 110 Hz tone. This is because the drop in pitch between the sounds is exaggerated by the auditory system. Such exaggeration of differences occurs throughout perception (vision, audition, touch) and has the benefit of sharpening boundaries (Summerfield et al., 1984).

The spectral contrast account hinges on the difference between the third formant frequencies (F3) of /a/ and /e/. The F3 of /a/ is low, below the starting F3 of the /d/∼/g/ sound. The gap between the lower F3 of the /a/ and the higher F3

2 For example, Mann and Repp (1981) demonstrated another context effect in which the boundary between /t/ and /k/ is altered by a preceding fricative: An ambiguous sound is more likely to be heard as /k/ when it follows /s/.

3 A formant frequency is a resonance of some part of the vocal tract. These resonances are altered by movements of the articulators and form the primary cues to distinguishing many speech sounds. The first formant is the lowest (in frequency) of all the resonance frequencies, the second is the second lowest . . . . Note that I am using the IPA in this dissertation — /a/ is the phonetic symbol for a typical North American English “r”.

107
of the /d/~/g/ sound is exaggerated by spectral contrast and thus the F3 of the ambiguous consonant is perceived as higher than it really is. Since an important cue for the consonant /d/ is a relatively high F3, this increased perceived height of F3 in the ambiguous sound pushes it towards the percept /d/. The converse of this is that /l/ induces more /g/ percepts — the F3 of /l/ is higher than the F3 of the /d/~/g/ sound, and so, in mirror fashion to the case above, the F3 of /l/ pushes down the perceived height of the F3 of the ambiguous sound, moving it towards the percept /g/. These effects are roughly schematized in Figure 4.2 and Figure 4.3.

Figure 4.2: Schematic of Contrast Explanation of Mann effect (/al/ Influencing /da/~/ga/ to Sound More Like /ga/)

Figure 4.3: Schematic of Contrast Explanation of Mann effect (/ar/ Influencing /da/~/ga/ to Sound More Like /da/)
4.1.3 The Articulatory Explanation of the Mann effect

Mann (1980) does not argue for a spectral contrast explanation. Instead, she argues that this and other context effects are due to compensation for coarticulation. This idea is based on the fact that the pronunciation of /d/ and /g/ are altered by the preceding sound. Since /ʌ/ is pronounced by drawing the tongue back into the mouth, a /d/ pronounced after an /ʌ/ tends to be further back along the roof of the mouth than usual. This results in /d/ being slightly more similar to a canonical /g/ in this position. Conversely, since /l/ is pronounced at the front of the mouth, it has a tendency to pull a following sound forward in articulation and a /g/ will be slightly more like a canonical /d/ in this position. Mann argues that our speech-processing system is aware of this coarticulatory influence and compensates for it in perception, so that following an /ʌ/, listeners will automatically ‘front’ their perception of /d/ to counteract the ‘backing’ they subconsciously know will occur after /ʌ/. This compensation means that a sound ambiguous between /d/ and /g/ will tend to be pushed towards a /d/-percept in the context of /ʌ/, and towards a /g/-percept in the context of /l/.

Carol Fowler (Fowler, 2006; Fowler et al., 2000) has a similar, yet subtly different, proposal. Under her theory there is not an active process of compensation (the perceptual system does not alter anything about the ambiguous /d/∼/g/ sound), rather the perceptual system is simply (and correctly) categorizing the same acoustic signal differently in different contexts because one and the same acoustic signal is an appropriate indicator of different events in different contexts.

Under Fowler’s direct realist theory, the objects of perception are the distal events in the external world. The distal events structure the acoustic signal that we receive and our perceptual system pulls the necessary information from this signal. The acoustic consequences of an event change in different contexts and our auditory system is sensitive to that context-dependence. For example, the acoustic signal of a plate-dropping on a carpeted floor are very different from the acoustic signal of a plate dropping on a tile floor, and yet our auditory system is able to pull out the essential similarity of these two events — both involve a plate dropping. In the case of variations in production, the distal event has remained the same: a

---

4By “canonical”, I mean a word-initial /g/, where a preceding context cannot have an impact (at least when the /g/ is pronounced in isolation).
/g/ after /l/ is the same event as a /g/ in word-initial position (both are instances of the tongue-dorsum striking the roof of the mouth). It’s just that the tongue-dorsum strikes the roof of the mouth a little further forward when /g/ is pronounced after /l/. The acoustics have changed, but the event has remained the same; in this case the acoustics of a “post-/l/” /g/ happen to be similar to the acoustics of a word-initial /d/.

When an ambiguous sound that would have been perceived as /d/ in word-initial position occurs instead after /l/, then the perceptual system recognizes the acoustics as appropriate to a /g/ for that context. No compensatory modification of the /d/∼/g/ sound needs to be done, it’s just that /g/ is a bit fronted after /l/ so our auditory system recognizes the ambiguous sound as a typical and appropriate instance of /g/ for the context. The mirror situation holds for /s/ and /d/. For a visual analogy, look at the image in Figure 4.4.

![Figure 4.4: An Example of Orthographic Context Dependence](image)

The central symbol is “B” when viewed as part of the row (letter) context, but “13” when viewed as part of the column (number) context. It is not necessary to postulate a perceptual mechanism that makes the symbol appear more ‘closed’ (and so more like “B”) in the letter context but makes the symbol more ‘spaced’ (and so more like “13”) in the number context. It is simply that “B” and “13” have a range of realizations and the realization of “B” in some contexts happens to look like the realization of “13” in other contexts.

---

5This visual effect was developed by Edwin Boring.
Analogously in the Mann effect, /d/ and /g/ have a range of realizations. The appropriate realization varies according to context. The realization of /d/ after an /ʌ/ happens to be similar to the realization of /g/ in word-initial position, and so a sound that would be categorized as /g/ in word-initial position is correctly categorized as /d/ when it occurs after /ʌ/. A mirror situation holds for the realization of /g/ after /l/. Our auditory system does not need to actively compensate for coarticulation (according to Fowler).

Though Fowler’s theory is not really about compensation it is usually discussed under the same rubric as Mann’s theory, both being called compensation for coarticulation.

4.1.4 Comparison of Competing Explanations of the Mann Effect

Both of these explanations of the Mann effect have a large supporting literature. Lotto and Kluender (1998) did a series of experiments demonstrating that the Mann effect occurs even when there is a switch in speaker (including a switch from female to male) between the context-setting /ʌ/A or /ʌ/l and the ambiguous /da/∼/ga/ sound. That the effect still occurred in this context is potentially troublesome for an articulatory account since it would be odd (though not impossible) for listeners to compensate for articulation that clearly comes from different speakers. Lotto and Kluender (1998) also showed that pure tones can disambiguate the /da/∼/ga/ just like natural speech (the pure tones were set at the F3 frequencies of /ʌ/A and /ʌ/l). This suggests that the context does not have to be speech at all and that a speech-processing explanation is less tenable. The mirror side of this finding was demonstrated by Stephens and Holt (2003) who showed that the /ʌ/A and /ʌ/l context sounds could induce perceptual shifts in non-speech targets.

In a vivid demonstration of the general nature of the Mann effect, Lotto et al. (1997) demonstrated the effect in birds. They conditioned Japanese Quail to peck one key in response to the sound /da/ and another key in response to the sound /ga/. When presented with a /da/∼/ga/ ambiguous sound, these birds pecked more vigorously at the /da/ key if the ambiguous sound was preceded by /ʌ/A and more vigorously on the /ga/ key if the sound was preceded by /ʌ/l. This is very strong, but not incontrovertible, evidence that the Mann effect is a matter of funda-
mental properties of auditory perception and not due to knowledge of articulatory patterns.

Fowler may counter that, since the Mann effect is due to domain-general properties of the auditory system (that allow articulatory information to be extracted), it is quite possible for the quail’s auditory system to be susceptible to the same effect, since all animals have need for many of the same essential auditory-processing abilities.

Mann (1986) showed that native Japanese speakers, even though they are typically unable to distinguish English /æ/ and /l/, still show the same context-dependent shift in their perception of /dA/ and /gA/. In support of an articulatory origin of the effect, Fowler et al. (2000) superimposed video of a speaker saying /AôdA/, /AôgA/, /Alda/ or /Arla/ over a sound that was somewhere between /Alda/ and /Arla/; that is, the sound had two ambiguities: the first consonant was /l/ ambiguous and the second was /d/ or /g/ ambiguous. They found that when participants saw movies which had visual /Aa/ as the initial syllable, they tended to perceive the second (target) syllable as /dA/ more often than if the movie had visual /Aa/ as the initial syllable. This would appear to be strong evidence that the effect is not reliant on formant information of the context sound, since the formants remained the same throughout (only the video changed). However, there is a potential issue with the controls in this experiment. The video not only showed a face producing the context sounds, but also showed a face producing the target sounds. This means that it is possible that participants were being influenced to hear /dA/ or /gA/ because they saw the speaker’s articulators producing that target rather than because of visual information about the preceding context (thus making this a case of the McGurk effect, rather than a context effect). Indeed, Holt et al. (2005) showed that when this visual information about the target sound is removed (so only visual information about the context is available), the context effect disappears.

Viswanathan et al. (2009) showed that when presented in isolation (filtered out from the rest of the sound) the F3 region of /Aa/ and /Aa/ is not sufficient to induce the Mann effect, suggesting that a spectral contrast account cannot account for the data. Furthermore Fowler et al. (2000) have argued that the Mann effect induced by tones (Lotto and Kluender, 1998) was in fact due to the artificially high inten-

112
sity of these tones which masked the acoustic information in the following targets. Support for this idea was found by Viswanathan (2009); though on the other hand, in the original experiment of Lotto and Kluender (1998) the tones were matched for energy to critical bands in the original contexts, weakening the possibility that artificially high intensity is responsible.

Both the compensation for coarticulation and contrast effect explanations have strong evidence in their favour and these explanations are not necessarily exclusive — it may be that both purely auditory processes and processes relying on the extraction of articulatory information contribute to the overall effect. This dissertation does not take a position on the cause of the Mann effect, but merely uses it as a behavioural measure of the degree to which one sound influences another. This is used as a dependent measure in all experiments reported in this chapter. The idea is that if corollary discharge occurs during the context sound (/ə/ or /ɒ/), then the impact of this context sound on the following target will be attenuated (in line with the sensory attenuation function of corollary discharge). Thus, the Mann effect is predicted to be weaker in the presence of corollary discharge on the context sound. In the absence of corollary discharge, we might predict that mouthing in unison with the context sound would actually reinforce the Mann effect by way of increased attention to and/or priming of the context sound. Thus, this experiment provides crucial support for the corollary-discharge interpretation of the results of chapter 2 over an attention/priming interpretation.

4.1.5 The Neural Origin of the Mann Effect

Regardless of which explanation of the Mann effect turns out to be correct, the level of nervous-system processing responsible for the effect seems to be at least at the level of the brainstem and probably higher. Convincing evidence for this was given by Holt and Lotto (2002) who demonstrated that the Mann effect can be induced dichotically, that is with the context played exclusively to one ear and the target exclusively to the other. Since there is essentially no binaural connection below the level of the brainstem, this is strong evidence that the effect must be at least at that level, or even higher. This result was elaborated by Lotto et al. (2003)

6 The brainstem is, of course, not particularly high in the processing chain.
to non-speech context sounds, where again, the effect was induced dichotically. Holt (2005) also reported experiments that showed that the Mann effect can be induced by a series of tones whose average frequency, extending over more than 2 seconds, is appropriate to induce the effect, even when the immediately adjacent tones are kept constant between conditions. This type of long-term average computation is typical of processing at higher (cortical) levels rather than peripheral auditory processing. This point of the level of processing is important, because it helps determine the level at which the type of corollary discharge associated with enacted speech imagery has an influence on perception. The experiments reported below demonstrate that corollary discharge interacts with the Mann effect and so suggests that this form of corollary discharge influences higher centres of auditory processing.

4.1.6 Overview of Experiments

In the following three experiments, I test the claim that corollary discharge is present when people engage in enacted speech imagery (mouthing) and that this corollary discharge can induce an attenuation of the Mann effect. Experiments 3-1 and 3-2 demonstrate this predicted attenuation. Experiment 3-3 shows that this attenuation also occurs when, instead of mouthing, people speak aloud.

4.2 Experiment 3-1

The first experiment tests for the presence of corollary discharge in enacted speech imagery (mouthing) by measuring the strength of the Mann effect when participants mouth in synchrony with the context sound (but not the target). If corollary discharge is present when people mouth, any external sound that matches the mouthed sound should have its perceptual impact attenuated by corollary discharge.

The predictions of this experiment are straightforward. The context sound has a perceptual impact on the following target. Corollary discharge attenuates the impact of sensations, and so if corollary discharge is present in enacted speech imagery, then if participants mouth the context sound while hearing the actual context sound, corollary discharge should attenuate the impact of the context and thus
weaken the Mann effect.

It might be argued that mouthing is distracting or cognitively difficult and this distraction may weaken the Mann effect. So a control **Contrasting** condition was included in which participants mouthed /a1/ and /a1/, but this time they mouthed the sound that contrasted with the recorded context sound, mouthing /a1/ when the recorded context sound was /a1/, and mouthing /a1/ when the recorded context sound was /a1/.

As a baseline measure, a **Hearing** condition was included in which participants simply listened to the context and target sounds without mouthing anything.

The predictions of this experiment are that in the **Matching** condition participants will show a significantly attenuated Mann effect in comparison to both the **Contrasting** and **Baseline** conditions.

The predicted outcome of the **Contrasting** condition is ambiguous. If mouthing any sound does distract from the main task, and thus interfere with the Mann effect (the potentiality that motivated including this condition in the first place), then we may expect that there will be an attenuation of the Mann effect in this condition as well, though not as severe as in the **Matching** condition. An argument could also be made that, even though the corollary discharge signal will not match the context sound very well in the **Matching** condition, they are still somewhat similar: In all instances, both the context sound and the mouthed sounds are, phonetically speaking, liquids preceded by the same vowel. This partial overlap between the context sound and the corollary discharge (induced by the mouthed sound) thus might be expected to induce attenuation of the Mann effect in the **Contrasting** condition; though again, it would not be as strong an attenuation as in the **Matching** condition, in which the match between corollary discharge and context would be much better. This variable degree of attenuation depending on degree of fit between corollary discharge and sound has been demonstrated experimentally. See subsubsection 1.5.5 for a review of the relevant literature.

Alternatively, an argument could be made that the presence of corollary discharge in the **Contrasting** condition would actually amplify the Mann effect, since it might be thought that corollary discharge would selectively attenuate any slight “/a1/-ness” in the /a1/ context sound or any “/a1/-ness” in the /a1/ context sound, making the context sounds even more extreme versions of their respective cate-
As will be demonstrated below, in both this experiment and in Experiment 3-2 the effect of the **Contrasting** condition was to slightly attenuate the Mann effect (though in neither experiment did this tendency reach significance).

### 4.2.1 Methods

The three conditions (**Matching**, **Contrasting**, **Hearing**) and two context sounds (/ɑ/) and /æ/) means that there were 6 types of block in this experiment as shown in Table 4.1.

<table>
<thead>
<tr>
<th>Action</th>
<th>Mouthed / Heard Sound</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hearing</td>
<td>/ɑ/</td>
</tr>
<tr>
<td></td>
<td>/æ/</td>
</tr>
<tr>
<td>Matching</td>
<td>/ɑ/</td>
</tr>
<tr>
<td></td>
<td>/æ/</td>
</tr>
<tr>
<td>Contrasting</td>
<td>/ɑ/</td>
</tr>
<tr>
<td></td>
<td>/æ/</td>
</tr>
</tbody>
</table>

**Table 4.1:** Conditions and Context Sounds of Experiment 3-1

The 6 types of block were presented in random order. To prevent order or position effects, each block was presented 7 times (7 cycles of the 6 types of block), with a new randomization of the order of blocks on each cycle. Within each block, there were 15 trials, with each of the target sounds presented 5 times in random order (three target sounds were used). This means that participants performed a total of 105 categorizations in each type of block (or 210 in each condition). Left and right arrow keys were used to register participants categorizations of the target. The mapping of left/right arrow to categorization was counterbalanced across participants.

As with the experiments in chapter 2, it was assumed that the perceptual influences would be strongest at the perceptual boundary of the target sounds. Thus a 1001-step continuum was created ranging from /da/ to /ga/ and a pre-test was
used to determine where along the continuum the perceptual boundary occurred for each participant.

Eleven equally spaced points along the continuum were presented to participants. Since most of the variation in perception should occur near the middle of the continuum, there is more information about participants’ perceptual boundary there, so the central steps along the continuum were presented more often in the pre-test. The distribution of number of repetitions of each of the eleven steps along the continuum is presented in Table 4.2.

<table>
<thead>
<tr>
<th>Step on Continuum</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td># of Repetitions</td>
<td>8</td>
<td>8</td>
<td>16</td>
<td>16</td>
<td>16</td>
</tr>
</tbody>
</table>

Continued . . .

<table>
<thead>
<tr>
<th>Step on Continuum</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
</tr>
</thead>
<tbody>
<tr>
<td># of Repetitions</td>
<td>16</td>
<td>16</td>
<td>16</td>
<td>16</td>
<td>8</td>
<td>8</td>
</tr>
</tbody>
</table>

Table 4.2: Number of repetitions of Each Step along the /da/~/ga/ Continuum for Experiment 3-1 Pre-Test

As with all other experiments in this dissertation, a probit analysis was used to determine the 50% point along the continuum for each participant. The 35% and 65% points were also determined, giving three target stimuli (whose data were pooled in the analysis).

The structure of a trial in this experiment was similar to that used in the experiments in chapter 2. As with those experiments, a movie consisting of a flashing red ball was presented in time with the experimental sounds in order to facilitate rhythm.

Participants began each trial by pressing the spacebar. After 150ms the context sound (either /a1/ or /a1/) was played twice in sequence, with 150ms between the tokens. A red ball (whose diameter matched the amplitude profile of the context sound) flashed on the screen in time to the context sound. The context sound varied from block to block but was always the same within the 15 trials of a block.

In the Matching and Contrasting conditions, participants were asked to mouth

---

7In Hz
8This is similar to the method used in Bertelson et al. (2003)

117
a syllable in time with the context sound. In the **Matching** condition, they mouthed the same syllable as the context sound and in the **Contrasting** condition they mouthed the opposite sound to what they were hearing (so mouthing /æ/ when they heard /ɑ/ and vice versa). Participants were asked to mouth in time to both presentations of the context sound, but were told that the primary purpose of the first occurrence was to establish timing and that it was the second occurrence of the context sound that was of most importance. They were also told that it was crucial that their mouthing end in synchrony with the context sound.

120ms after the second token of the context sound, one of the target /dA/∼/gA/ sounds was presented. Participants were asked to categorize these sounds in a forced choice between “da” and “ga”. A schematic of the events in a single trial is presented in [Figure 4.5](#).

![Figure 4.5](#): This is a timeline of how stimuli were presented in each trial in Experiment 3-1. Each ‘frame’ of this timeline represents roughly 150 ms. In this example, the context sound is /æ/.

In the **Hearing** condition everything was exactly the same except that participants did not mouth anything.

The same cover story was used in this as in all other experiments in the dissertation — that the experiment was testing the impact on response time of triggering the *tensor tympani* middle-ear reflex.

A rough schematic of the structure of the experiment is presented in [Figure 4.6](#).

As with all experiments in this dissertation, participants were given an extended practice session and only moved on to the main experiment when both they and I felt they were ready. Before beginning each block, participants had to type in a two-letter code that represented the condition (**Hearing**, **Matching** or **Contrasting**)
and context sound (/α/ or /ɔ/) that they were about to perform. This was to ensure that participants were paying attention to the task. The experiment itself took about 27 minutes to complete. Participants were monitored via closed-circuit camera and microphone to ensure they were performing accurately.

4.2.2 Stimuli

For the target sounds, a continuum from a clear /də/ to a clear /ɡə/ was synthesized using Praat (Boersma and Weenink, 2001). The continuum was modelled on the stimuli used by Holt (1999). The sounds were 300ms in duration. The F1 and F2 formant trajectories were kept constant throughout the continuum, with only F3 varying.

For all steps on the continuum, F1 started at 200 Hz and rose linearly to a steady state of 750 Hz over an 80 ms transition; F2 started at 1650 Hz and fell to a steady state 1200 Hz. F3 varied across the continuum from a starting value of 2650
Hz for the clear /da/ end of the continuum to 1650 Hz for the clear /gα/ end. For all points on the continuum, the F3 moved linearly to a steady state 2450 Hz over the course of the 80 ms transition.

Between neighbouring steps of the continuum, there was a 1 Hz difference in the starting point of F3 (hence the continuum had 1001 steps). The pitch and intensity of the tokens was kept constant. The fundamental frequency started at 110 Hz and fell linearly to 100 Hz over the course of 250 ms with a faster decline to 95 Hz over the last 50 ms of the sound. The pitch and formant values of these target sounds were thus typical of a male voice.

Using such a finely divided continuum allowed for the target sounds to be closely matched to each participant’s perceptual boundary (as determined by the experimental pre-test).

As discussed above, only 3 tokens from this continuum were used for each participant. These were the tokens corresponding to that participant’s 50% point in perceptual space ± 15%; meaning the points along the continuum where the participant heard the token as /da/ 35% of the time, 50% of the time and 65% of the time.

For the context sounds, a male native speaker of English was recorded saying /Aô/ and /A/. Two tokens (one for each context sound) which were similar in pitch and intensity contours and free of artefacts were selected. These were trimmed to 365 ms and normalized to have the same average intensity.

Both context and target sounds were played over speakers. The context sounds were presented at c. 58 dB and the target sounds at c. 62 dB

### 4.2.3 Participants

There were 26 male participants (average age = 21.6 years; SD = 3.1 years), all were either paid or given course credit for their participation.

Male participants were selected for several reasons. Since corollary discharge is dependent on a degree of match between the predicted sound of one’s own voice and the external sound one hears, using stimuli of a male voice with female participants, or vice versa, might have introduced errors. Thus, it was best to select a single sex for the experiment (for both stimuli and participants). Males were
chosen because, to date, all replications of the Mann effect that I am aware of have used male target sounds and (with the exception of Lotto and Kluender [1998]) male context sounds. Thus, it was safest to use stimuli that were known to work when replicating the Mann effect. In a similar vein, prior to this experiment I conducted several pilot studies using male context and target sounds and had succeeded in replicating the Mann effect, thus I was confident that these stimuli were appropriate.

4.2.4 Results

The strength of the Mann effect was measured as %-difference in /dA/-categorizations between the /AÔ/ and /AL/ contexts. It was predicted that this strength would vary across conditions. The categorizations from all three target sounds were pooled for analysis.

Since there is a minority of the population that consistently shows an inverse Mann effect (Choe et al., 2009), participants who did not display a typical Mann effect in the Hearing condition were excluded (N = 4). A Grubbs test detected 2 outliers who were also excluded.

A repeated-measures ANOVA was performed on the %-difference scores (in /dA/-categorizations) between the /AÔ/ and /AL/ contexts. There were three conditions: Hearing, Matching and Contrasting and the results were significant [F(2, 38) 3.597, p = 0.03709]. Post-hoc tests (FLSD) found that the Matching condition showed a significantly weaker Mann effect than the Hearing condition, while the Contrasting condition did not. These results are shown in Figure 4.7.

The experiments reported in chapter 2 used a Holm-Bonferroni correction for pairwise comparisons. This correction is appropriate for experiments with multiple conditions (Experiments 1-1 and 1-2 had five conditions each), however with only three conditions to compare, FLSD is appropriate and so that is what was used here and in Experiment 3-2.
4.2.5 Discussion

These results demonstrate that the Mann effect is indeed weaker when participants are engaged in speech imagery if the content of that imagery matches the context sound they are hearing. This weakening was not found when the imagery did not match the heard context sound (though there was a similar trend in this condition). This strongly suggests that speech imagery induces corollary discharge, whose hallmark is an attenuated response to sensations which match its content.

It seems that by mouthing speech sounds, participants were generating corollary discharge, which is essentially a prediction of hearing the sound that is being mouthed. When an external sound is played in time to this corollary discharge, the prediction is fulfilled and corollary discharge can execute its function, which is to attenuate the impact of the sound. I have argued that this attenuation may
be achieved by channelling the external sound into a ‘self-produced’ processing stream. If correct, this would mean that, to a degree, participants were treating the externally presented context sounds as feedback of their own (unspoken) voice.

This experiment was not without problems, though. While there was a large difference between the Matching and Contrasting conditions, this difference did not reach significance. This is an important issue and is taken up in Experiment 3-2, which is essentially a replication of Experiment 3-1 but with tighter controls aimed at achieving greater experimental power.

It should be noted that while the Contrasting condition was not significantly different from the Hearing condition, the means were substantially different. This issue is taken up in the discussion section of Experiment 3-2 (subsection 4.3.5).

4.3 Experiment 3-2

As discussed above, Experiment 3-1 established that enacted speech imagery attenuates the Mann effect when the imagery matches the inducing context sound. This is in line with the prediction of corollary discharge. However, Experiment 3-1 did not find a clearly significant difference between the Matching and Contrasting conditions. This is a crucial distinction, since the Contrasting condition is an important control, establishing that the attenuation found in the Matching condition is not due to distraction, or some other factor associated with mouthing itself. So Experiment 3-2 replicates Experiment 3-1, but with more experimental power.

4.3.1 Methods

Experiment 3-2 is identical to Experiment 3-1, but with the following changes:

- The pre-test used to determine each participant’s perceptual boundary between the target /da/ and /ga/ sounds was extended from 11 steps to 17 steps (the distance between steps thus being reduced). More steps along the /da/~/ga/ continuum were presented to participants and there were more tokens to categorize in total. Thus, the accuracy of the pre-test (and by extension the sensitivity of the experiment) was significantly improved.

- The /da/~/ga/ continuum was resynthesized, this time with 2001 steps be-
tween the clear /da/ and /ga/ ends. This allowed for a more precise choice of participants’ target sounds and thus would reduce noise in the data.

- Rather than randomize both the condition and the context sounds across multiple cycles, the order of conditions and context sounds was strictly counterbalanced across participants. Thus, with three conditions and two context sounds, this means there were 12 versions of the experiment needed to achieve complete counterbalancing. In addition, the side of the response-buttons was again counter-balanced across participants and so there a total of 24 versions of the experiment needed for complete counterbalancing.

- Each block was shortened from 15 trials to 12 trials.

- There were more repetitions of each type of block: each type of block was repeated 9 times (as compared to 7 times in Experiment 3-1). It was hoped that more repetitions of shorter blocks would result in more spread out fatigue and learning effects.

- The trials were made slightly shorter (the gap between context sounds was reduced from 150ms to 130ms and the gap between the final context sound and the target was reduced from 120ms to 115ms).

- The total number of categorizations was increased (very slightly) from 105 to 108 categorizations in each type of block (meaning an increase from 210 to 216 categorizations in each condition).

- The perceptual spacing between the three target tokens (as determined by the pre-test) was reduced. The three target tokens used in this experiment corresponded to the participant’s 50% point along the continuum ± 10% (so, 40%, 50% and 60%). This means that the targets were more ambiguous in this experiment, thus potentially increasing their susceptibility to being influenced by context.

All of these differences were intended to make the Mann effect slightly stronger (and thus give more scope for detecting its attenuation by corollary discharge) and
to reduce the variance in participant responses. Other than these differences, the set-up was identical to Experiment 3-1.

The order of conditions was counterbalanced across participants, as was whether they started with the context sound /æ/ or /æ/. The context sounds alternated back and forth in strict succession throughout the experiment.

Seventeen points on the /da/∼/ga/ continuum were presented in the pre-test. As with Experiment 3-1, the central points on the continuum were over-represented. The steps along the continuum used and the number of repetitions of each step is given in Table 4.3.

<table>
<thead>
<tr>
<th>Step on Continuum</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td># of Repetitions</td>
<td>5</td>
<td>5</td>
<td>10</td>
<td>10</td>
<td>15</td>
</tr>
<tr>
<td>Continued . . .</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>Continued . . .</td>
<td>12</td>
<td>13</td>
<td>14</td>
<td>15</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>15</td>
<td>10</td>
<td>10</td>
<td>5</td>
</tr>
</tbody>
</table>

Table 4.3: Number of repetitions of Each Step along the /da/∼/ga/ Continuum for Experiment 3-2 Pre-Test

4.3.2 Stimuli

The same Praat script was used to generate the /da/∼/ga/ continuum as in Experiment 3-1. So the stimuli were identical, except that in this experiment 2001 steps were generated between the endpoints, rather than 1001. Thus there was a $\frac{1}{2}$ Hz difference in F3 starting point between neighbouring points on the continuum. Such a fine-grained continuum is perhaps excessive, but generating the continuum is quick and simple and having a finer-grained continuum is potentially beneficial (the chosen target sounds can more accurately match the participant’s predicted perceptual boundaries).

Of course, participants were never aware of the size of the continuum; as with Experiment 3-1, only three target sounds were used for each participant. These
corresponded to the participant’s perceptual 40%, 50% and 60% points along the continuum (as determined by the pre-test).

The context stimuli were exactly the same as in Experiment 3-1, see subsection 4.2.2).

4.3.3 Participants

As discussed above, a complete counterbalancing of conditions required 24 participants, thus there were 24 male participants (average age = 20.7 years; SD = 2.2 years). All were either paid or given course credit for their participation.

4.3.4 Results

As with Experiment 3-1, the strength of the Mann effect was measured as %-difference in /dʌ/-categorizations between the /ʌl/ and /ɒl/ contexts. It was predicted that this strength would vary across conditions. The categorizations from all three target sounds were pooled for analysis.

A repeated-measures ANOVA was performed on the %-difference scores (in /dʌ/ categorizations) between the /ʌl/ and /ɒl/ contexts. There were three conditions: Hearing, Matching and Contrasting and the results were significant [F(2, 46) 7.866, p = 0.00115]. Post-hoc tests (FLSD) found that the Matching condition was significantly different from both the Hearing and Contrasting conditions. These results are shown in Figure 4.8.
4.3.5 Discussion

These results replicate the difference between Matching and Hearing found in Experiment 3-1 and also demonstrate that the difference between Matching and Contrasting is significant. This is crucial since it demonstrates that the weakening of the Mann effect found in the Matching condition is not due to mouthing per se, but due to the match between what is mouthed and what is heard. This is exactly what a corollary discharge account of speech imagery would predict (see subsection 1.5.5).

As with Experiment 3-1, there was a large, but not quite significant, difference between the means of the Contrasting and Hearing conditions. This points to the possibility that corollary discharge is attenuating the context sound even in the Contrasting condition, but that the match between corollary discharge and heard
sound is not close enough for the attenuation to be particularly strong, and that is why the difference in means does not reach significance. This is exactly in line with the predictions of corollary discharge, in which it is typically found that the degree of sensory attenuation is correlated with the degree of match between the sensation and the content of the corollary-discharge signal (the imagery in this case). This aspect of corollary discharge is discussed in subsection 1.5.5.

Another interpretation of the trend towards attenuation in the **Contrasting** condition is due to the distraction caused by mouthing. Perhaps doing any concurrent task will attenuate the Mann effect, and this is why we see attenuation in this condition. Crucially, this distraction explanation *cannot* explain the results for the **Matching** condition, in which the distraction of the concurrent task would presumably be weaker, and yet the attenuation is significantly larger.

The evidence from these experiments is compatible with either the *spectral contrast* or *compensation for coarticulation* explanations of the Mann effect.

There are at least two interpretations compatible with the *spectral contrast* explanation. First, if we view corollary discharge as simple attenuation, we could argue that the frequency region appropriate to the context sound’s F3 is attenuated by corollary discharge, which thus weakens the impact of that F3 on the following target sound.

Alternatively, if we view corollary discharge as performing a ‘channelling’ function, we could argue that there is partial stream-segregation being induced by corollary discharge. This would (partially) segregate the context sound from the target and so attenuate the impact of the context sound’s F3 on the following target sound.

Under a *compensation for articulation* account, the attenuation of the Mann effect can be explained using the ‘channelling’ interpretation of corollary discharge. Participants may fail to compensate for coarticulation when the context and target come from different speakers (but see Lotto and Kluender, 1998) and so, in this theory, corollary discharge would cause participants to attribute the context sound (partially) to a ‘self-caused’ perceptual stream. In this scenario, the context sound would be treated as originating from a different speaker and so not an appropriate source of coarticulatory influence on the target */dA/~/ga/ sound. Thus, there would be less of a tendency to compensate for coarticulation.
Under both explanations, auditory streaming could play a role. The idea is that the corollary discharge mechanism is essentially channelling the context sound (/au/ or /al/) into a ‘self’ auditory stream that insulates it from the target (/da/∼/ga/). This isolation could mean either that the F3 of the context is isolated from that of the target or that the context sound is isolated from the target in terms of being a likely source of coarticulation; either way the effect could be due to auditory streaming. This is the same process that I proposed in subsection 1.5.5 to explain the attenuation of tickle when the stimulation is self-caused. Essentially, I am claiming that corollary discharge is preventing the context from tickling the ear in these experiments.

The possibility that auditory stream segregation attenuates the Mann effect was suggested by the findings of Lotto and Kluender (1998), who showed that the Mann effect still occurred when the context sound was female and the target sound male, but that the effect was weaker than when both context and target were male. This is perhaps because some degree of stream segregation isolated context from target. It is possible that corollary discharge is having the same effect in these experiments, causing participants to channel the context sound into a ‘self’ stream. Of course such a channelling of the context sound would not be complete — it is not as if participants were fooled into believing they were hearing their own voice. However, despite their conscious awareness that it was not their own voice, it is possible that subconscious auditory processing was fooled — to a degree — by the corollary discharge signal.

4.3.6 Mouthing vs. Pure Imagery

It may be asked why I concentrated on enacted speech imagery (mouthing) in this experiment and did not include a non-enacted (pure) imagery condition. This was done for several reasons. First, the experiment was already quite long and complicated and adding another condition would simply have been pragmatically impossible. Second, the impact of non-enacted speech imagery is anticipated to be significantly weaker than that of enacted speech imagery. This was shown in chapter 2, in which enacted imagery had a significantly stronger impact on perception than non-enacted speech imagery; thus, a much larger experiment would need to
be run to detect the effects of non-enacted speech imagery. Third, even if a much larger experiment were run, and succeeded in demonstrating a qualitatively similar attenuation of the Mann effect, it could always be argued that the attenuation was due to micro-movements of the articulators and not due to ‘pure’ (completely non-enacted) speech imagery.

This is not a problem for my theory of speech imagery, since micro-movements of the articulators is a quite common component of imagery (see section 1.4 and section 1.3), but their presence means that it is difficult to test whether ‘pure’ non-enacted speech imagery also involves corollary discharge. I would not be surprised if the presence of corollary discharge is dependent on the degree of articulator engagement, in line with the ‘flexible abstractness’ of inner speech discussed in section 1.3. If this is the case then I suspect that the degree of sensory content in inner speech tracks the degree of articulator engagement as well — future research will address this question. This experiment is intended to establish the possible presence of corollary discharge in imagery and so examined the circumstances where corollary discharge would be most easily detected.

### 4.4 Experiment 3-3

Experiment 3-2 provided strong evidence for the presence of corollary discharge in inner speech, however this conclusion is based on the assumption that the corollary discharge would indeed attenuate the Mann effect. This is a reasonable assumption to make and Experiment 3-3 tests its validity by comparing the strength of the Mann effect across two conditions: one in which participants speak the context sound (/ʌ/ and /əl/) aloud and another in which the same sounds (recorded when participants spoke them aloud) are played back to participants to act as the context sounds. Since corollary discharge is undoubtedly present in real speech (see subsection 1.5.5) the Mann effect should be weaker when participants speak the context sounds aloud than when participants hear recordings of their own voice. This is what was found.

This experiment had just two conditions: **Speaking** and **Listening**. Participants performed a block of speaking each of the context sounds (/ʌ/ and /əl/) and then a block of hearing each of the context sounds in turn. Thus, there were
four types of experimental block as shown in Table 4.4.

<table>
<thead>
<tr>
<th>Action</th>
<th>Spoken / Heard Sound</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speaking</td>
<td>/a1/ / AI/</td>
</tr>
<tr>
<td>Listening</td>
<td>/a1/ / AI/</td>
</tr>
</tbody>
</table>

**Table 4.4:** Conditions and Context-Sounds of Experiment 3-3

The basic design of this experiment is very simple: participants spoke /a1/ or /AI/ aloud in a rhythm after which they were presented a target /dA/~//gA/ sound which they categorized in a forced-choice between “da” and “ga”. There were 12 repetitions of this within a block. After doing two blocks (one for each context sound) participants switched conditions and instead of speaking, they listened to a recording of themselves from the previous condition saying the context sound (/a1/ or /AI/ depending on the block), after which they again categorized a /dA/~//gA/ target. All of the 12 trials in each Speaking block were recorded and these recordings were played back in random order for the 12 trials of each Listening block. This alternated back and forth for a total of 16 cycles. The experiment predicted that the Mann effect would be weaker when participants actually produced the context sound (and so engaged corollary discharge) than when they simply heard a recording of their own voice (in which the acoustics were identical, but no corollary discharge was present).

Despite the simplicity of the basic design, the details required a fairly complex implementation. These are discussed below.

### 4.4.1 Methods

The structure of this experiment was very similar to that of Experiments 3-1 and 3-2. As with those experiments, participants were presented with a flashing red ball which determined a rhythm in which they were to perform.

Each trial began by pressing the spacebar. After 250 ms, a sequence of two movies of the bouncing red ball were played, with a 250 ms gap between the
movies.

In the Speaking condition, participants were instructed to say the context sound (/ɑ/ or /ɑ/, depending on the particular block) in time to the red ball. 150 ms after the second red-ball movie was presented, participants heard a target /dɑ/~/gɑ/ sound and categorized it in a forced choice between “da” and “ga”.

In the Listening condition, participants merely listened to recordings of the context sound (recordings of their own voice taken during the Speaking condition) being played in time with the red ball followed by a target sound to be categorized\[10\]. Participants completed 12 trials in each block. After a block for each context sound had been completed in one condition, the participant then completed a block for each context sound in the next condition. Participants alternated back and forth between these 4 types of block, completing a total of 16 repetitions of each type of block, for a total of 192 categorizations within each type of block (or 384 categorizations in each condition).

Because the Listening condition consisted of playing back sounds recorded in the Speaking condition, Listening had to follow Speaking. Thus counterbalancing the order of conditions was impossible. To deal with any order or position effects, the blocks were made very short and participants alternated frequently between the two conditions, switching back and forth between the conditions 16 times. The order of context sounds was counterbalanced across participants (half of participants got the order /ɑ/, /ɑ/, /ɑ/, /ɑ/ . . . half got the reverse order). The side of the response buttons (left and right arrow keys) was also counterbalanced across participants.

A rough schematic of Experiment 3-3 is provided in Figure 4.9.

A picture of the headphone/microphone set-up is shown in Figure 4.10.

The context and target sounds were played over headphones\[11\]. This is different from Experiments 3-1 and 3-2, and was done to limit the difference in localization of the context sounds in the Speaking vs. Listening conditions. Obviously the sound of one’s own voice is unavoidably localized to the middle of one’s own face (though with some acoustic diffusion). Thus headphones were used to deliver the

---

10 The recording was done with a Shure WH30 head-mounted microphone with the microphone approximately 3 cm from the participants’ lips.

11 This is the only experiment in this dissertation in which sounds were not presented in free field.
recorded context sounds in the Listening condition in roughly the same localization — because of binaural presentation, headphones make sounds appear to be localized to the mid-line of the head (Warren 2008).

However, this presented a problem in that headphones would isolate the ears somewhat to the sound of the participant’s own voice in the Speaking condition. This was overcome by using ‘open’ headphones (Grado Headphones, model SR60). Open headphones have an open casing (no covering on the back of the speaker), and so cause minimal isolation. However, for strict parity between the two conditions, the recorded context sounds were filtered (via Praat script) to ensure that the sounds in the Listening condition matched the small amount of isolation that would occur from the headphones in the Speaking condition. The isolation profile for the headphones was obtained from:

http://www.headphone.com/headphones/grado-sr-60i.php

Another discrepancy between the sounds in the Speaking and Listening conditions is created by the response characteristics of the headphones. No speaker
diaphragm can respond to all frequencies equally, and so the response of the headphones will slightly alter the spectrum of a recorded sound. The recorded sounds were filtered to correct for the response characteristics of the headphones. The response profile for the headphones was obtained from:


Microphones also have response characteristics that will alter the difference between recording and original sound. The microphone used for this experiment was a head-mounted Shure WH30 whose frequency response (provided by the company) was counteracted via filtering of the recording.

This leads to the issue of bone conductance. Approximately half of the sound we experience from our own voice is in fact due to the sound of our voice being conducted through the bone of our skull to our ears, rather than being conducted through air — which is how we hear other people’s voices (Shuster and Durrant, 2003). This leads to a significant quality change between how we hear our own voice and how others hear it. Shuster and Durrant (2003) conducted a series of experiments to determine the filtering characteristics of this bone conductance and
came to the conclusion that the best perceptual match between a recording of our voice and how we hear our own voice is to filter the recording at 1000 Hz with a 3 dB boost to frequencies below this frequency, and a 3 dB reduction of frequencies above it. This standard has been used repeatedly since (Ford and Mathalon, 2005; Heinks-Maldonado et al., 2006; Kaplan et al., 2008; Mathalon et al., 2005). Thus, in addition to the filtering done to compensate for the isolation characteristics of the headphones, the recordings were also filtered (using this ± 3 dB filter) to compensate for bone conductance.

This means there were four different stages of filtering:

1. Filtering to compensate for the isolation of the headphones (Grado sr60).
2. Filtering to compensate for the response characteristics of the headphones.
3. Filtering to compensate for the response characteristics of the microphone (Shure WH30).
4. Filtering to compensate for bone conductance (± 3 dB centred around 1000 Hz).

At the end of each Speaking block, Psyscope automatically called a Praat script which performed all of these filterings as well as trimming the start and end of each recording so that there would be no ‘popping’ sound when they were played back (this trimming was simply to the nearest zero-crossing in the waveform and so only altered the duration of the sound file by a fraction of a millisecond). In each trial, the recording started 250 ms before the first context sound was to be spoken and continued 150 ms after the last one was to end, so that if participants made minor mistakes in the timing of their speaking, this would not create problems. While this may seem like a lot of filtering, each of the filters was a very small adjustment and together their perceptual impact was slight.

\[12\] Since only about half of the auditory input when hearing one’s own voice comes from air-conduction (Shuster and Durrant, 2003) (the other half comes from bone conductance), the compensation-for-isolation was divided in half (after proper conversion to deal with the fact that the decibel scale is logarithmic) so that there was no over compensation — only the air-borne portion of the sound of one’s own voice would need to be filtered to compensate for the isolation characteristics of the headphones; the bone-conductance portion would be unaffected.
In addition to this filtering another Praat script was called which measured the duration of each of the recordings to determine if a computer malfunction or excessive computational load (this was a memory intensive experiment) might have caused recording to start late or end early (and thus produce an inappropriate recording for use in the \textbf{Listening} condition). An elaborate system was coded that would check for recording duration and substitute an unaffected recording from a different trial in such a condition, while eliminating the affected trial from the data for both the \textbf{Speaking} and \textbf{Listening} conditions. In the event, this elaborate precaution was not needed. There were no computer problems and all trials were properly recorded and played.

The next issue was the timing of the playback with respect to the timing of the original speaking. In the \textbf{Speaking} condition, participants said the context sounds in time with a flashing red ball (as in all other experiments in this dissertation). In the \textbf{Listening} condition the playback of the recordings also had to be timed to occur in synch with the red ball. This was achieved by calibrating the computer through repeated testing.\footnote{The testing was done by embedding sine waves in the experiment script at the times when the participant would be speaking and at the point when the playback of the recording would occur. The experiment was then run and the sounds produced were recorded on a separate computer. The sine waves were of different frequency and so the difference in their timing was easily measured on a spectrogram and the position of sounds could be repeatedly adjusted until the timing between them was essentially zero.} With this iterative testing, the playback of recordings was reliably in the same temporal position (vis-a-vis the red ball) as the original speaking. The standard deviation in the timing of the playback with respect to the time of speaking was just 5 ms (with 0 ms average difference). Thus the context sounds were in an equivalent temporal relationship with the target sounds in the \textbf{Speaking} and \textbf{Listening} conditions. This yoking of speaking with playback means that if a participant was slightly off rhythm in the \textbf{Speaking} condition, exactly the same off-rhythm relationship would hold in the \textbf{Listening} condition, and thus the two conditions were perfectly controlled with respect to timing (Participants were monitored for severe timing problems and none occurred).

The last equalization issue between recording and playback is the volume of playback. Simple equalization of the volume is inappropriate since it is essentially impossible to determine the intensity of self-produced speech at the inner
ear (especially with the confound of bone conductance). A perceptual equivalence is more appropriate, so before the experiment started, each participant was asked to speak into the head-mounted microphone and were immediately played back a recording of their speech. Participants adjusted the gain on the microphone pre-amplifier until they felt that the playback volume was equal to the volume of their own voice when they spoke. It has been demonstrated that given such a set-up people usually set the playback volume at a value that is *lower* in absolute terms than the intensity of their voice as measured at their lips (Toyomura et al., 2009). This is because people typically hear their own voice as less intense than it really is (perhaps, again, due to the effects of corollary discharge).

This is not an issue for this experiment, because perceptual equivalence is more important (especially since the Mann effect seems to be a higher-level process — see subsection 4.1.5). Furthermore, a lower playback intensity of the context would only lead to a less intense Mann effect (Viswanathan et al., 2009) and thus would work against my prediction. Therefore, it is not a confound (remember, my prediction is that participants will show a weaker Mann effect in the **Speaking** condition, not in the (playback) **Listening** condition). In the event, all participants set the gain at exactly the same value. The relative intensity of speech versus playback for this level of gain was measured inside the ear canal (of the author’s ear) using an **Audio Scan Verifit** Analyzer. The intensity of speech in this set-up was c. 8.5 dB louder than the intensity of playback — in line with predictions and confirming the results of Toyomura et al. (2009) that people perceive their own voice as quieter than it is. The intensity of the target sounds was consistent at c. 78 dB. The intensity of the context sounds, since they were produced in real time by the participants, varied from trial to trial and from participant to participant; but the difference between speaking and playback was consistent (at c. 8.5 dB) and the intensity of the targets was consistent. As a point of reference, my own voice (recorded at my ear canal) was c. 79 dB, the playback of my voice was c. 71.5 dB, and the target sound was c. 78 dB.

The same pre-test as in Experiment 3-2 was used to determine each partic-

---

137

14 This experiment showed that lowered intensity of F3-like tones led to a weaker context effect.
15 Thanks to Susan Small for lending her lab equipment and Kelly-Ann Casey for doing this measurement for me.
participant’s perceptual boundary along the /da/~/ga/ continuum. Exactly the same procedure was used as in Experiment 3-2, including the choice of targets (corresponding to participants’ 40%, 50% and 60% points on the continuum).

As with Experiments 3-1 and 3-2, participants were given an extended practice session to familiarize them with the experimental set-up and only moved on to the main experiment when both they and I felt they were ready. Before beginning each block, participants had to type in a 2-letter code that represented the Condition (Speaking or Listening) and Context sound (/aʊ/ or /al/) that they were about to perform. This was to ensure that participants were staying on task. The experiment took about 30 minutes to complete. Participants were monitored via closed-circuit camera and microphone to ensure they were performing accurately.

4.4.2 Stimuli

Participants provided their own context sounds in this experiment, as discussed above. The target /da/~/ga/ sounds were taken from the same 2001-step continuum used in Experiment 3-2. As with Experiment 3-2, the 40%, 50% and 60% perceptual points on the continuum were used — different for each participant and determined by pre-test.

4.4.3 Participants

There were 12 male participants (average age = 22 years; SD = 3 years). All were either paid or given course credit for their participation.

4.4.4 Results

As with Experiments 3-1 and 3-2, the strength of the Mann effect was measured as %-difference in /da/-categorizations between the /aʊ/ and /al/ contexts. It was predicted that the Mann effect would be weaker in the Speaking condition than in the Listening condition. The categorizations from all three target sounds were pooled for analysis.

A repeated-measures ANOVA was performed on the %-difference scores (in /da/ categorizations) between the /aʊ/ and /al/ contexts. There were two conditions: Speaking and Listening and the results were significant [F(1, 11) = 14.891,
p = 0.00265]. These results are shown in Figure 4.11\textsuperscript{16}

As can be seen in Figure 4.11 the mean of the \textbf{Speaking} condition is not just lower than the \textbf{Listening} condition, it is below the zero-line, suggesting that the Mann effect is not just weaker when participants speak the context themselves, but is in fact reversed. A one-sample t-test was performed comparing the mean of the \textbf{Speaking} condition with a predicted mean of zero (\(\mu = 0\), meaning no Mann effect at all). With a Bonferroni correction to compensate for this being the second comparison of the data, the results were significant (\(t = -2.224\), df = 11, \(p = 0.02403\)); indicating that the Mann effect actually reverses in this condition.

\begin{figure}[h]
  \centering
  \includegraphics[width=\textwidth]{figure411.png}
  \caption*{\textbf{Figure 4.11:} Results of Experiment 3-3. Standard-error bars are shown.}
\end{figure}

\textsuperscript{16}Of course, an ANOVA with only two conditions is mathematically equivalent to a t-test.
4.4.5 Discussion

In this experiment, with the context sounds kept as acoustically identical as possible, the strength of the Mann effect was still significantly weakened when participants spoke the sounds themselves as compared to when they heard a recording of their speaking. In fact, the weakening of the Mann effect was so severe that in the Speaking condition, the effect was actually negative, that is, the normal pattern of results for the Mann effect was reversed. These results are strongly suggestive of corollary discharge, the hallmark of which is an attenuated impact of self-produced sounds.

The results of this experiment suggest that the corollary discharge explanation of the weakened Mann effect in Experiments 3-1 and 3-2 was correct.

As with Experiments 3-1 and 3-2, these results are compatible with a spectral contrast account of the Mann effect, but less clearly with a compensation for coarticulation account.

The reversal of the Mann effect in this experiment fits well with a view of corollary discharge as a matter of sensory attenuation. In the case of speaking aloud, the attenuation seems to be sufficiently strong as to reverse the usual impact of the context on following sounds. Such after-effects are relatively common. An example often used in classroom demonstrations is to expose students to a long-duration vowel sound with a constant spectrum and then switch to white noise — most people will hear the converse of the spectrum (its ‘photographic negative’) in the white noise as an after-effect of the exposure (Summerfield et al., 1984).

The relationship to a channelling view of corollary discharge is less clear. Since channelling sensations into different perceptual streams should isolate those streams from interference, a channelling explanation is compatible with a weaker Mann effect, but in this experiment, an inverse Mann effect was found. It is not clear how channelling the context into a different perceptual stream could lead to a reversal of the usual pattern of interference.

It should be borne in mind, though, that despite auditory streaming, perceptual streams can still interact to a degree. For example in dichotic-perception experiments in which a tone is played in one ear and an ambiguous syllable in the other — even though the tone is streamed as belonging to a different source from the
syllable, it can have an impact on the identification of the syllable, a phenomenon known as *duplex perception* (Rand, 1974).

A further possibility to consider in this regard is that perceptual streaming involves a *corrective* mechanism that corrects for inappropriate interactions between perceptual streams. This would mean that lower-level auditory interactions between sounds that do not come from a common source would be *undone* by higher-level corrective mechanisms. Under this proposal the corrective mechanism would simply be over-correcting in Experiment 3-3 and this is why we find an inverted Mann effect.

While the finding of a reversal of the usual Mann effect in this experiment is suggestive, I would not want to draw too much from it. The result should be replicated using other context effects. The difference between the **Listening** and **Speaking** conditions was much more robust than the reversal effect and is the more important result of this experiment, establishing that the Mann effect is weaker when the context is self-generated and thus that this weakening is likely to be due to corollary discharge.

### 4.5 Conclusion

The experiments reported in this chapter provide compelling evidence that the impact of sounds is attenuated when those sounds are accompanied by enacted speech imagery (mouthing). Since sensory attenuation is the hallmark of corollary discharge, these experiments support one of the central claims of this dissertation: that the sensory content of speech imagery is constituted by corollary discharge.

In the first two experiments, participants mouthed a syllable in time to a recording of a syllable. When the content of what was mouthed matched the recording, then the perceptual impact of the recording (as measured by the strength of the Mann effect) was attenuated. This attenuation was not significant when the mouthed syllable was different from the recording.

In the third experiment this finding was extended to natural speech, where it was found that the usual pattern of the Mann effect was not just attenuated, but actually reversed when the context sound was spoken aloud in comparison to when it was simply a recording.
The presence of corollary discharge in speech imagery also supports the claim that speech imagery has *sensory*, and not just *phonemic*, or category-level information. Corollary discharge is inherently a *sensory* signal. Its purpose, throughout the animal kingdom, is to map incoming sensations onto sensory predictions — the predictions being about which of the sensations are self-caused. While it is not clear at what level of abstraction this sensory prediction is coded, it *is* clear that it is a *sensory* signal and not a symbolic (in the sense of abstract categories) signal. This fits with the claim that inner speech has sensory content.

As stated before, there is an inherent limitation to this line of argument. While the experiments above provide strong evidence for the presence of corollary discharge in inner speech, there is always room for the counterargument that while corollary discharge is *present* in inner speech it is not *constitutive*. For example, we know that corollary discharge accompanies external speech (as shown by Experiment 3-3), but it is not the case that corollary discharge *constitutes* external speech; and someone may argue that the same relationship holds for inner speech.

While such a counterargument is perfectly valid, I believe that inference to the best explanation points to corollary discharge as the bearer of sensory content in inner speech.
Chapter 5

Further Discussion

The theory put forward in this dissertation is at the crossroads of several disciplines and several conceptual issues. There is not enough space in a single dissertation to address even a small subset of these issues. This is, I think, a very lucky circumstance, since it allows me to use the work I have done for this dissertation as a staging area for research into many domains of interest.

In this chapter I would like to address a few of the issues that are related to the topic of this dissertation but are not on the ‘main line’ of the research and so were not dealt with earlier.

I will first deal with the relationship between the theory I am proposing and the Motor Theory of Speech perception. This is an important point since the experiments reported in chapter 2 could be interpreted as consistent with the Motor Theory. Furthermore, I believe that the theory I am proposing in this dissertation informs the debate over the validity of the Motor Theory of Speech Perception.

Following the discussion of the Motor Theory, I present a series of speculative ‘afterthoughts’ that relate the topic of this dissertation to other areas. These sections contain ideas and material too tangential and hypothetical to include in earlier discussions and so have been collected together to form a speculative addendum. I do not claim to have experimental support for the ideas suggested in these sections — they are merely speculations that may serve as the basis for future research.
5.1 Relationship to Motor Theories of Speech Perception

*every representation of a movement awakens in some degree the actual movement which is its object* — William James (1890)

This dissertation is not arguing for or against a motor theoretic view of speech perception. However, the topics and experiments presented here are obviously closely related to those discussed in motor-based theories, so I should specify the relationship between motor-based theories and what I am proposing.

There are a range of views about what gets perceived in speech perception. The classical view is that the objects of speech perception are auditory events. However, there are alternative views that argue for a role of production in speech perception. The strongest claim is that of the Motor Theory of Speech Perception (Liberman and Mattingly, 1985) which claims that speech perception is achieved via a specialized module which extracts the intended speech gestures from the acoustic signal. A similar view is proposed by Fowler (1986) who offers a Gibsonian approach to speech perception in which it is speech gestures that are recovered in perception but, in contrast to the Motor Theory, this recovery is through general auditory mechanisms rather than a biologically specialized speech-perception mechanism. The view that the fundamental units of speech perception (and production) are gestures is also shared by the Gestural Phonology approach (Goldstein and Fowler, 2003).

There is a lot of evidence to support such views. First, there is the evidence presented by Liberman that when acoustics and gestures diverge, our perception tracks the gesture and not the acoustics. For example, the same stop burst of 1400 Hz will be perceived as a /p/ before high vowels, but as a /k/ before a low vowel (Liberman, 1952). This means that our auditory system correctly attributes the same acoustics to two different gestures in the different environments, suggesting that our speech perception is more interested in what the gesture was than what the acoustics were. This is part of a general issue known as the ‘lack of invariance’, which refers to the fact that there is no invariant acoustic cue to the identity of a speech sound, rather all cues are heavily context-dependent. This has led many to argue that what is invariant is the gesture that produced the variable acoustics. For example, I am a relatively large adult male, my production of the syllable /ma/
is going to be acoustically very different from the ‘same’ syllable spoken by a small child. Yet despite the acoustic differences, our productions are perceived as being, in some sense, the same. The motor-based theories of speech say that what is shared is the gesture we produced (both involved a closure at the lips and an opening of the nasal passage etc.).

The theory that the motor system is involved in speech perception also finds significant support from brain-imaging studies. Using fMRI, Wilson et al. (2004) found that similar areas of motor cortex become activated during speech perception and speech production. Pulvermüller et al. (2006) extended this result, finding more specificity in this production-perception overlap. They found that labial motor areas were more activated than tongue areas when perceiving labial sounds and that hearing ‘tongue’ sounds produced the reverse pattern of activation.

However, despite the evidence that I have just surveyed, I am not arguing for (or against) a motor-based theory of perception.

The experiments reported in this dissertation all involve speech perception occurring simultaneously with silent speech production. I argue that this silent speech production engages the forward models of motor control, discussed in subsection 1.5.1, which influence perception through corollary discharge. Thus, these experiments are only informative about the unique situation in which forward models are engaged during speech perception, and so cannot be extended to ‘normal’ perception situations. They do not, therefore, provide support for motor approaches to speech perception.

In fact, I would argue that much of the evidence usually adduced to support motor-theoretic claims can be more parsimoniously explained with the forward model/corollary discharge mechanisms discussed in section 1.5. For example, D’Ausilio et al. (2009) found that TMS applied to articulator-specific areas of motor cortex (e.g., the area responsible for controlling the lips) had a facilitatory effect on the perception of sounds involving that specific articulator, but not on the perception of unrelated sounds. For example, labial sounds such as /p/ are more easily heard when the cortical area that controls lip movement is stimulated. While this could be argued to support a motor theory approach to speech perception, I think it is better seen as the result of forward models — by stimulating the motor cortex for a particular articulator, the forward-model system is engaged, as it is when any
motor command is triggered, and corollary discharge appropriate to that articulator is produced. This corollary discharge then influences the perception of external sounds.

This alternative explanation is available for many other experiments and points to a common problem with such experimental designs. Since the forward-model system is bound to influence perception via corollary discharge, and since forward models are engaged whenever speech production occurs, any perception experiment that involves triggering speech production always has a confound. It cannot determine whether the resulting effects on perception are representative of what happens during normal perception, or whether the effects are simply an artefact of the presence of corollary discharge. These reported effects could simply be ‘fridge-light phenomena’ — that is, phenomena that are only there when you go looking for them because the experimental set-up induces them, but are not there outside of the experimental context. I believe that this is not a confound for the experiments reported in this dissertation, since these experiments are not about normal perception, but about the effects of forward models. Thus the fact that such forward models are absent in normal perception is moot for my purposes.

The brain-imaging evidence showing that motor areas of cortex are activated by speech perception could be explained with internal models as well. In this case by inverse models. The function of inverse models is to provide the motor commands needed to achieve a particular goal. It is possible that speech perception automatically recruits inverse models so that a person can quickly repeat a sound they have just heard. It is known that we are able to repeat speech sounds at a remarkably short latency (Porter and Castellanos, 1980); and imitation is certainly a basic aspect of human cognition, as shown by Meltzoff and Moore (1977) who found that newborns (between 12 and 21 days old) were able to imitate facial and manual gestures. The fundamental tendency of humans to imitate is also seen in evidence from patients with frontal lobe damage who seem unable to suppress a tendency to imitate (De Renzi et al., 1996). There is also evidence from the phonetics literature showing that people have a tendency to take on the phonetic characteristics of others automatically (Goldinger, 1998). In order for this to occur, the phonetic targets of others must be translated into motor coordinates, which is the job of inverse models.
Under this hypothesis, the motor system (the inverse model component) is triggered after speech perception to allow for immediate repetition/imitation of what is perceived, and is not part of the perceptual process. This would explain why perception activates the motor system. Again, I am not arguing for this view, merely presenting it as an alternative to the theory that the motor system is a component of speech perception.

While I have presented alternative interpretations for much of the evidence used to support motor theories, I am not taking a stance against motor theories. It is possible, as argued by Pickering and Garrod (2007), that when perception is faced with a difficult task, top-down information in the form of motor-predictions is used to ‘fill in the gaps’. Under such a proposal, speech perception would be primarily an auditory process, as argued in the classical view, but when the perception is particularly difficult the motor system makes predictions about what is to come and in so doing constrains the possibilities that the auditory system has to entertain and thus eases the computational load. This would be a hybrid motor/auditory view of speech perception. Skipper et al. (2006, 2007) have proposed a similar view, again arguing that speech perception is not necessarily a matter of coding the incoming sound into a gesture code, but that engagement of forward models can be used to aid the speech-perception task. Skipper et al. argue that such forward model engagement is what underlies the influence of vision on speech perception, such as in the McGurk effect (McGurk and MacDonald, 1976). This idea finds support from an MEG study showing suppression of the N100m auditory-cortex response to tones when participants were lip-reading (Kauramäki et al., 2010), suppression being a hallmark of the corollary discharge output of forward models (as discussed in subsection 1.5.5).

These motor-helping-hearing theories are also quite similar to that proposed in the Perception-for-Action-Control Theory (Schwartz et al., 2010). This theory argues that speech perception is not motor-based, but that speech gestures do define equivalence classes for speech sounds. The idea is that the motor system helps establish which sounds count as members of the same category, membership being determined by sharing a common method of production. However, once the sound classes are set, the motor system is not used online in the act of perceiving the members of these classes, such online perception being achieved by the auditory
system. Schwartz et al. (2010) would argue that the lack of online motor-system involvement has one exception — when auditory perception is made difficult because of missing information, the motor system can be used to ‘fill in’ the missing information (this is similar to the proposal of Pickering and Garrod (2007)).

5.1.1 Auditory vs. Gestural Coding

The discussion of motor theories above is closely tied to the issue of the domain of speech coding. Is the neural code for speech a motor code? or an auditory one? or some hybrid?

There is necessarily a tight communication between auditory and motor representations of speech since we use auditory feedback to guide our speech production at a very short latency. This suggests the need for a ‘common currency’ between auditory and motor representations of speech — a hypothesized coding that can be used by both systems (Goldstein and Fowler 2003). Representing speech in a common, amodal, code would also more easily allow other modalities to contribute to perception, which is important since there is ample evidence that visual (e.g., Jääskeläinen et al., 2008; McGurk and MacDonald, 1976) and tactile (e.g., Gick and Derrick, 2009; Gick et al., 2008) information are used in speech perception.

In a sense, the distinction between auditory and motor codings is a matter of deciding on labels — deciding where, along a continuum of processing, perceptual processing ends and action processing begins.

There are animal models that highlight the problem of drawing a sharp distinction between auditory and motor coding. For example, many species of moth have evolved an auditory system that is exclusively used for bat avoidance. By comparing the intensity of the bat’s echolocation call between its left and right ears, the moth determines the location of the bat and alters its course accordingly. In these moths, the auditory system consists of only two neurons in each ear and the output of these neurons goes directly to the motor system to alter flight behaviour (Madsen and Miller, 1987). Thus, the output of the ear could be considered a motor-command signal, making the distinction between motor and auditory repre-

\footnote{This is known as \textit{phonotaxis} and is not exclusive to moths.}

\footnote{There are complications to the story: When the bat is very close (and so very loud), the moth’s behaviour changes from a course-correction to sudden, seemingly random, evasive manoeuvres.}
sentations largely an arbitrary decision — is the output of the moth’s four auditory neurons an auditory or motor representation? Either could be an appropriate categorization and nothing in nature hangs on our decision.

The situation is not exactly parallel to that of speech perception/production. In speech, the motor and auditory codings are more clearly of the same thing (speech), rather than of two different things as in the case of the moth in which what is perceived (the bat) is very different from the subsequent motor act. However both situations highlight the fact that there is one end of the processing chain that is clearly auditory and another that is clearly motor and between there is an abstract neural coding that bridges between those ends. Is this abstract coding motor or auditory? I would argue that the decision is largely arbitrary and picking one or the other label says more about our theoretical perspective than about the nature of the system under examination. This is essentially the issue discussed in the Theory of Event Coding (TEC) (Hommel et al., 2001).

According to the TEC, as the processing of a sound (or other sensation) proceeds up the processing chain from sound-transduction in the cochlea, through the subcortical nuclei, all the way to the auditory cortex, the representation of sound becomes increasingly abstract and thus increasingly concerned with ‘events’ as opposed to ‘features’. For example, the cochlea essentially performs a Fourier analysis of a sound; however, at the level of conscious awareness, this information is not accessible, and instead we are simply conscious of the event that caused the sound. If we hear a coin drop on a hard floor, we are primarily aware of the event of the impact and not of the acoustic features that informed us about the event.

This basic-to-abstract chain of processing is reversed in the preparation of actions by the motor system, where higher levels are more abstract — concerned more with goals than with the details of how these goals are to be achieved. The details of motor implementation are dealt with by a nested hierarchy of lower and lower level control centres, eventually cashing out in commands to muscles to contract. The TEC suggests that the abstract coding of the late stages of perception is the same as the coding for the early stages of motor processing. Under this theory,

3 Though, of course, our awareness may not be particularly detailed — we may be aware that there was a collision but not aware that it was a collision of a coin with the floor. Also, though we are primarily aware of the event, we can focus our attention on the low-level features of a sound.
events are what are of the greatest ecological concern to an organism and so it is events that we perceive and events which our motor system plans.

5.1.2 The Motor/Sensory Content of Corollary Discharge

An obvious question when discussing the sensory estimate contained in corollary discharge is: what is being estimated? Does the corollary-discharge signal estimate such low-level aspects of the signal as formant values? Or does it specify a more abstract level of auditory processing?

Since corollary discharge is compared against a sensory signal (that is its function), it must contain a comparable type of information as that sensory signal. However, there is an increasing level of abstraction as a sensory signal moves up the processing chain, and it is unclear at what point in the processing chain the comparison with corollary discharge occurs. Therefore it is also unclear how abstract the sensory representation in corollary discharge is. Of course, if the Motor Theory approach to speech perception is correct, then as the auditory processing becomes more abstract it becomes more motor-based and so the corollary-discharge signal would be a motor-based code.

This dissertation does not take any position on what level of sensory content is carried by the corollary discharge signal and to what degree this content should be characterized as motoric.

5.1.3 The Analogue vs. Propositional Debate

The question of the level of sensory representation in corollary discharge is related to the question of analogue vs. propositional representation of imagery. There is an ongoing debate in the imagery literature about the fundamental nature of images: Are they symbolic (propositional) or are they depictive (analogue)? A symbolic representation would be one in which the image is not tied to the original sensory experience, not containing any of the original sensory detail. So, for example, a word like “rose” can represent an object without having any of its characteristics (the word itself is not thorny or red). Whereas a depictive image would include information similar to an actual sensory experience.

This debate has been conducted primarily in the area of visual imagery where
the issue has largely been settled in favour of depictive accounts, though not without dissenters (see Dennett, 1992). The question has been less hotly contested in the inner speech literature, possibly because there is simply less literature in this field. A symbolic account of inner speech would mean that our experience of inner speech is largely a matter of manipulating abstract symbols (possibly phonemes, syllables or lexemes) without any accompanying phonetic representation. For example, MacKay (1992) argues that inner speech does not contain phonetic detail. As evidence for this view he cites the fact that speech errors that occur in inner speech typically involve phonemic substitutions and lexical errors, but never involve phonetic mistakes. Furthermore, he remarks that many phonetic components of external-speech are typically absent in inner speech:

Many aspects of the acoustics of overt speech are normally absent from our awareness of self-produced internal speech. To illustrate, consider loudness and fundamental frequency, integral characteristics of the acoustics of overt speech. Unlike overt speech perception, awareness of the loudness and fundamental pitch of words produced internally is normally absent. Moreover, speakers normally fail to note the absence of these omnipresent characteristics of overt speech (MacKay, 1992, p.198).

Introspectively it may seem that there must be sensory content to inner speech, after all, we can ‘hear’ the characteristic quality of our own voice and what could this mean other than that the sensory details of our voice are present? However, introspection is a dangerous guide: most people are unaware that their peripheral vision is virtually colour-blind or that they have a blind-spot near the centre of their field of vision. Introspectively, we are unaware of such absence of content, but the content is absent nonetheless. The same may hold true of inner speech: ‘hearing’ our own inner voice does not necessarily mean that our inner voice carries the sensory detail that an external sound does.

To be very clear, my dissertation argues that inner speech can contain sensory detail, and what is more, that this sensory detail can influence the perception of external sounds.
5.1.4 Relationship to Schizophrenia

Schizophrenia is one of the most devastating mental illnesses. It affects about 1% of the population globally and is characterized by a suite of symptoms. The positive symptoms of schizophrenia include the hearing of voices and feelings of external control, as if one’s body were being controlled by others (Frith, 1992).

The feelings of control and the voice hallucinations have been theorized to stem from a common problem, corollary discharge malfunction.

The review of corollary discharge in section 1.5, did not discuss one putative function of corollary discharge: generating a sense of agency. It is argued that when the comparison of corollary discharge with reafference is successful (i.e., the two signals match quite closely), the reafference is perceived as self-caused; thus corollary discharge underlies the feeling of agency. The role of corollary discharge in agency is implicit in the discussion of streaming in subsection 1.5.5.

In relation to schizophrenia, Frith (1992) has proposed that an error in the corollary discharge mechanism leads to problems with feelings of agency in patients, which is why they do not self-attribute actions the way non-schizophrenics do, and hence feel that their actions are being controlled by others. Under Frith’s hypothesis, when schizophrenics hear voices, they are simply talking to themselves in their head (as we all do), but they incorrectly attribute the voice to someone else.

The link between inner speech and hallucination was first proposed by Kandinskii in 1890 (cited in Frith, 1992), and has since been strongly supported by both behavioural and brain imaging studies.

Some studies have found that if a sensitive microphone is used, some hallucinating schizophrenics can actually be heard speaking to themselves in a very low voice, with the content of the recording corresponding to the reported content of the hallucination (Gould 1949, Green & Preston 1981; both cited in Frith, 1992).

\[\text{As opposed to the negative symptoms, such as anhedonia, that are an absence or reduction of normal experience}\]

\[\text{Though corollary discharge is presumably not the only source of the feeling of agency, for a review of various sources of the sense of agency, see Sato} (2009).\]

\[\text{It is interesting to note that the earliest discussion of inner voice in the Western intellectual tradition comes from Socrates (as portrayed by Plato), who is known to have heard voices in his head (which he attributed to his} \text{daemon and because of this is sometimes thought to have been schizophrenic} \text{(Leudar and Thomas, 2005)}.\]
Bick and Kinsbourne (1987) found that hallucinations were reduced if patients kept their mouths wide open (preventing them from performing subvocal articulation).

Recalling from subsection 1.6.2 that the cerebellum appears to be implicated in corollary discharge, it is interesting to note that Shergill et al. (2000) found (using fMRI) that patients with a history of hallucination showed a reduced cerebellar activity when asked to produce auditory imagery. Shergill et al. (2003b) extended this result, examining the influence of speaking rate and again found that schizophrenics with a history of auditory hallucinations showed a reduced activation in the cerebellum in comparison to healthy controls when asked to repeat a word in their inner voice at a high repetition rate.

Consistent with this corollary discharge explanation of hallucinations, Ford and Mathalon (2004) found that schizophrenics showed less sensory attenuation to sounds when talking. This lowered level of attenuation was found by measuring the N1 ERP to auditory stimulation in talking versus silent conditions. This result was replicated in Ford and Mathalon (2005).

Johnson (2005) provides a thorough overview of the brain imaging literature on the relationship between corollary discharge and schizophrenia.

My Proposal for the Failure of Corollary Discharge in Schizophrenia

A simple view of what causes the loss of the sense of agency in schizophrenia is that corollary discharge is simply absent during periods of hallucination. If this were the case, it would present a problem for my theory — corollary discharge cannot provide the sensory content of a hallucination if is absent during hallucinations. This is only superficially a problem. It is unclear what exactly the forward model/corollary discharge problem is in schizophrenia and it is not necessarily the case that hallucinations are caused by the absence of corollary discharge. The problem could be elsewhere in the forward model system.

I will offer one purely speculative suggestion, merely as a demonstration that a malfunction of corollary discharge is not incompatible with corollary discharge providing the sensory content of both speech imagery and auditory hallucinations.

Since forward models need to provide corollary discharge in multiple modalities (auditory, visual, kinaesthetic, vestibular), and since the time-course for reaf-
ference is different in these different modalities (see subsection 1.5.1), the corollary-discharge signals would not all be timed simultaneously, but would instead have to be in a precisely staggered relationship with each other. It is possible that it is the correct timing relationship between these different corollary-discharge signals that causes the feeling of ownership for self-caused actions. Under this theory, when corollary discharge in a particular modality is not timed correctly with respect to other modalities, the action that produced the corollary discharge is less likely to be considered self-caused. Thus, it would be a disruption of corollary-discharge timing relationships that causes the external control symptoms of schizophrenia. If this were true, then it would not be that auditory corollary discharge is absent in schizophrenics when they are hallucinating, but that the auditory corollary discharge is not properly timed with respect to other corollary discharge signals. This would allow corollary discharge to constitute the sensory content of speech imagery and auditory hallucinations while at the same time playing a role in the misattribution of the source of that content.

5.1.5 Relationship to Reading

Reading is often accompanied by the experience of our own voice in our head and also seems to involve subvocal articulation. This fits with the theory of inner speech presented in this dissertation.

Most of us will have had the experience of hearing our own voice in our head as we read. The link between inner speech and reading has been established for a long time. Young readers typically mouth the words they are reading and even adults, who no longer make such visible movements, can ‘sound out’ words in their head, as demonstrated by the fact that they can understand sentences like: “If yew kann sowned owt thiss sentunns, ewe wil komprihenned itt” (Baddeley et al., 1981, p.439). Baddeley et al. (1981) provide evidence that adults still rely on subvocal articulation for some components of reading. When engaging in articulatory suppression (counting silently) their ability to detect anomalous words is compromised, though their reading speed is not affected.

Abramson and Goldinger (1997) performed two lexical decision experiments on visually presented words and found that the phonetic length of the to-be-categorized
words influenced response time (even though orthographic length and number of phonemes was controlled), suggesting that the phonetic content of the words are accessed in reading. This result was replicated for phonetic differences in vowel duration (e.g., the durational difference between ‘plead’ and ‘pleat’) by Lukatela et al. (2004).

Eiter and Inhoff (2010) found that reading of words could be disturbed if a phonologically similar spoken word is presented simultaneously; however this disruption does not occur if the articulators are already engaged in a different task (articulatory suppression). This suggests that subvocal articulation is used in the normal course of reading and that when it is not available (under articulatory suppression), phonological distraction is not as potent.

In line with the corollary discharge theory of inner speech, Buchsbaum et al. (2005) found deactivation of some auditory areas in single-word reading, which they tentatively attribute to corollary discharge.

5.1.6 Relationship to the Phonological Loop

The part of working memory we use when maintaining a phone number in memory long enough to dial (silently repeating it in our head) is called the ‘phonological loop’ (Baddeley, 1983). This is the component of working memory that lets us maintain snippets of language in short-term memory by rehearsing the snippet using subvocal articulation. The rehearsal is heard internally as our inner voice (Baddeley, 1981). The phonological loop is just one use of the general phenomenon of inner speech.

The literature on the phonological loop is vast, but the most relevant finding for this dissertation is that the phonological loop is dependent on subvocal articulation. This is supported by two types of evidence.

The first line of evidence is that the amount of content that can be held in memory is strongly dependent on the articulatory duration of the content — for example, Welsh speakers can hold fewer numbers in their short-term memory than English speakers because Welsh numbers take longer to articulate (Ellis and Hennelly, 1980). Such a duration effect has been replicated for sign language where the number of signs that can be maintained in memory (without overtly producing the
signs) is dependent on how long it takes to actually produce the signs, independent of the structural complexity of the signs (Wilson and Emmorey, 1998).

The second line of evidence that our rehearsal is based on subvocal articulation is that some aspects of the phonological loop can be disrupted if the articulators are distracted (by articulatory suppression) (e.g., Jones et al. 2004).

5.2 The Speculative Addendum: Observations and Speculations

The following are observations and speculations on the experience of inner speech and the possible connections between corollary discharge and other aspects of cognition. These sections are meant to be read merely as speculation for further research.

5.2.1 Why Your Voice Sounds Weird in Recordings

It is a common observation that our voice sounds different to us when we hear a recording of it compared to when we are speaking. This difference is usually attributed to the effects of bone-conduction which acts as a low-pass filter for our voice and thus makes it seem ‘bassier’ to us than it does to others. Maurer and Landis (1990) (cited in Shuster and Durrant 2003) examined this low-pass filtering to determine its perceptual characteristics; however their results were inconclusive. Shuster and Durrant (2003) also used perceptual measures to determine the degree and nature of filtration. While they had some success with filtering recordings with +3 dB for frequencies below 1000 Hz and -3 dB for frequencies above 1000 Hz (the filtration used in Experiment 3), there was still significant mismatch between this filtration and participants’ perception of their own voice when they spoke. In their conclusion they write “Given our results and those of Maurer and Landis (1990), it is clear that the transfer function is not easy to characterize.”

I would suggest that the problem is not quantitative but qualitative. The presence of corollary discharge when you speak will make the perception of your own voice a qualitatively different experience from the perception of any recording of your voice, no matter how accurate the filtering. This is similar to the example given in subsubsection 1.5.5 of the difference between someone else tickling you
and you trying to tickle yourself. Even when you touch yourself in exactly the same way as you were touched by someone else, the experience never feels the same because of the presence of corollary discharge when you perform the action yourself. The same situation holds with speaking — corollary discharge channels the sound of your own voice into a different, ‘self-caused’, processing stream and thus it can never be perceived in the same way as an external presentation of your voice. Essentially, external sounds are able to ‘tickle’ your ear, but you will never be able to ‘tickle’ your own ear with the sound of your own voice. This means that all attempts to equate a recording with the sound of our voice as we hear it are probably doomed to failure. While we can make the experience more similar, we will never be able to cross that qualitative gap.

5.2.2 Some ‘Parlour Tricks’ with Inner Speech

The following is a collection of observations (‘parlour tricks’) about the phenomenology of inner speech. They support the claim that inner speech is generated by the motor system and is a prediction of the sound of one’s own voice. These are merely demonstrations and are not meant as conclusive evidence. They are also untested experimentally (except for asking lab-mates if they experience the same phenomenon) and perhaps a little odd, which is why they are included as part of this speculative addendum for those interested in seeing if they have the same experience.

I do not expect that everyone will experience these ‘parlour tricks’ in the same way that I do, and I do not want to make too much of them, but they are suggestive of the theory proposed in the dissertation and so I present them here. These demonstrations are very much in the vein of the introspective studies of the early psychologists, with the same potential for being misled by wishful thinking, and so should be taken with a grain of salt. The best that can be hoped for in these cases is ‘intersubjectivity’, namely finding that most people report the same introspective experience.

---

7Unless, perhaps, you mouth along with it and generate corollary discharge in that way.
• **A pencil in the mouth perturbs inner speech**

This demonstration shows that our inner speech does seem to be a prediction of the sound of our own voice, even in cases where our normal voice is perturbed.

**Instructions**

Hold a pencil in your teeth so that it obstructs your normal pronunciation patterns. If you attempt to ‘mouth’ a sentence — moving your mouth as if you were speaking normally, but not actually producing any sound — you will probably find that your inner auditory experience is perturbed; it will probably sound as if you have a pencil in your mouth. This does not seem to work if you do not actually move your mouth (imagining your voice without moving doesn’t seem to have the effect).

This experience supports the idea that inner speech has phonetic content (how else can it have disturbed phonetic content?) The phonological structure has not changed, just the details of how it is implemented at the phonetic level. Furthermore this fits with the corollary discharge account which says that the phonetic content of inner speech is a prediction of the sound of one’s own voice. In this example, the forward models are generating a prediction of what your voice would sound like if you spoke with a pencil in your mouth.

You can compare the inner auditory experience under perturbation with the sound of your voice when you actually speak with the same perturbation. I find (and I suspect you will too) that the auditory experience of the disturbed inner speech is remarkably similar to the sound of the disturbed external speech, even when the perturbation is novel.

• **‘McGurk’ inner-speech percepts**

Sartre remarked that one can never be surprised by one’s own imagery (Sartre, 2004). A similar idea is what underlies the contentious claim that imagery is
inherently about something (the philosophical notion of intentionality) and so can never be ambiguous.

However I believe it is possible to generate an unexpected imagery experience in the auditory domain and this is how:

**Instructions**

Repeat the syllable /ta/ several times so that you are comfortable with knowing where your tongue is positioned at the start of this syllable.

Place your tongue in the position to say /ta/ again, but do not say it. Simply hold your tongue in that starting position for /t/. With your tongue held against the roof of your mouth, repeat in your inner speech the word “mom” several times in a rhythm (you may find it difficult to manage this, but you should eventually be able to ‘hear’ the word “mom” while holding your tongue against the roof of your mouth). Now, as you are repeating the word “mom” in a regular rhythm, drop your tongue exactly in time with the start of one of the repetitions of “mom”. You will probably hear /nam/, which is a combination of the nasal and voicing components of /m/ with the place of articulation of /t/. This suggests that the auditory experience we have in inner speech is dependent on the action of our articulators.

This ‘mixed’ percept arising from conflicting sources of information is similar to the McGurk effect (McGurk and MacDonald 1976).

- **Whispered and nasal inner voice**

  This demonstration also shows a dependency of inner speech on motor components.

**Instructions**

Count to ten inside your head, trying to hear your inner voice as a whisper. Now hold your breath and try to do the same thing again. If your experience is like mine, you will be unable to achieve a whispered inner voice with your
breath held. Now release your breath and try again, this time paying attention to the state of your larynx. You may find (as I do) that you can feel the tension in your larynx that normally accompanies whispering — suggesting that this atypical form of phonation can only be experienced in inner voice if the motor system is in a configuration appropriate for generating it.

A comparable situation holds for imagining one’s voice as very nasal. For me, this can only be done with the velum dropped, which is the appropriate position for generating nasal sounds.

- **Cupping your mouth while mouthing**

  This demonstration shows that the perceived location of your inner voice can be altered.

  **Instructions**

  Mouth the numbers from one to ten. Now repeat the action, this time cupping your mouth with one hand as if you were whispering in someone’s ear (without actually whispering); then switch hands, cupping your mouth with your other hand. You may find that you hear your inner voice switch back and forth between your ears as if it were being directed to the non-occluded ear. This suggests that your inner voice is a prediction of the consequences of your actions — when you cup your hand in a way that would direct your real voice to one ear, your inner voice follows suit.

### 5.2.3 Speculation on Music and Dance

There is a debate about the evolutionary origin of music and dance and whether these universal aspects of human culture are in anyway useful or whether they are merely useless by-products of mental capacities that evolved for other purposes. I would like to offer one suggestion as to why we enjoy these activities and a possible social role for them. The account, of course, centres around corollary discharge.

A discussed in subsection 5.1.4, one of the functions of corollary discharge is to produce a feeling of agency over self-caused sensations. It is this function that
appears to be impaired in schizophrenics, which is why their sense of agency is disrupted leading to feelings of external control and the experience of hallucinations.

When performing an action in time to a predictable external sound (as with the beat in a piece of music), the sound itself\(^8\) is predictable and occurs in time with the kinaesthetic feedback of the performed action. This means that it is possible for the person dancing (or simply tapping their foot), to start to incorporate the predicted sound into their self-produced perceptual-feedback stream. This would mean that the person may start to feel a sense of agency over the sound, even if all they are doing is dancing/tapping along with it.

This hypothesized extended sense of agency is similar to the extended sense of ownership seen in illusions like the rubber hand illusion, in which correctly timed feedback is sufficient to cause a person to incorporate a foreign object (a rubber hand) into their body-image (Botvinick and Cohen [1998]). Furthermore, this proposed extended sense of agency for music finds support in the behaviour of teenagers, who regularly engage in ‘air-guitar’ performances in time to music; an activity that suggests that they derive enjoyment from pretending to be the agent of the music (even when all they are doing is moving in time to it).

A further aspect of this extended sense of agency that might play a social role is the possibility that it leads to a shared sense of agency between all those dancing. When everyone is performing an action in rhythm, then each person can generate a sense of agency over the whole event (thanks to corollary discharge), but since each person sees all others performing in time to the same event, each person can also get a sense that others are agents of the same event. This may lead to a sense of shared agency and so in this way create a social bond (‘we are all the same’) between all those dancing.

All of the above is of course pure speculation; but I think it is interesting enough to consider.

\(^8\)At least the occurrence of the beat, the actual notes may not be predictable in an unfamiliar piece of music.
5.2.4 Why Self-Caused Sensations are Perceived as Earlier in Time

Sensations that are self-caused are perceived as occurring earlier in time than externally-caused sensations, that is, the time between an action and a sensation is perceived as shorter when the action is believed to have caused the sensation (Eagleman and Holcombe, 2002; Haggard et al., 2002). I would like to propose (merely as speculation) that this is due to the monitoring function of corollary discharge.

As discussed in section 1.5 one function of corollary discharge is to provide faster feedback about the results of an action than are possible through the normal sensory channels. This means that, by definition, the sensory consequences of a self-produced action are available, via corollary discharge, sooner than the sensory consequences of a non-self produced action. So, for example, when seeing two rocks collide we have visual information about the collision only after the collision has occurred. However, when we strike one rock with another, corollary discharge provides a prediction of the collision before the collision occurs.

We do not seem to be aware of the corollary-discharge signal, that is when we perform an action we do not perceive two versions of the same event: first a predicted sensation of the event (corollary discharge), followed by the real sensation (reafference). Why not?

One possibility is that corollary discharge is inherently a subconscious process and never reaches the level of awareness. Another possibility is that corollary discharge does reach the level of awareness, but that the memory of corollary discharge is erased (perhaps by the arrival of reafference). A third possibility, and the one that I am proposing here, is that the awareness of reafference is fused with the awareness of corollary discharge and so the perceived timing of the reafference is blended with the perceived timing of the corollary discharge, producing a compromise perception of timing. Since corollary discharge is available before reafference, that would explain why self-caused sensations are perceived as occurring earlier in time.

---

9 Though this dissertation would argue that corollary discharge does reach conscious awareness in the case of imagery. However, in this dissertation it is argued that it is the sensory attenuation form of corollary discharge that constitutes imagery, not the monitoring form.

10 In its monitoring capacity, but not its sensory attenuation capacity.
This theory may seem to be assuming some sort of central ‘theatre of the mind’ where all sensory events get evaluated in strict linear order, and so potentially generating a homunculus fallacy. While this is a valid concern it is not a particularly strong counterargument. We may not know exactly how timing is represented and perceived in the mind, but it is not unreasonable to assume that, in general, the order of events in ‘real’ time will be reflected in the perceived order of events.

I should point out that while I came to this idea independently, I have since found a similar idea proposed in the literature. In one of the first demonstrations of a shorter temporal gap between action and self-caused sensation, Haggard et al. (2002) suggested that a forward model may be involved (though they did not provide any details on how a forward model would produce this effect).
Chapter 6

Conclusion

For many of us, moments of apparent silence are in fact not silent at all. Instead, these moments are filled with a nearly constant stream of internal sound: inner speech. This internal monologue is such a regular component of our daily lives that it seems to have been largely overlooked as an issue worthy of scientific investigation.\(^1\) This ‘hiding in plain sight’ feature of inner speech was remarked on by the 19th century philosopher/psychologist Victor Egger:

\[
\text{Et voilà pourquoi la parole intérieure a échappé à l’attention de la plupart des psychologues; faute d’être reconnue, elle passe inaperçue; elle est comme ces personnes actives et modestes qui, dans une famille ou dans une société, rendent mille services sans exiger de retour, dont chacun subit la bienfaisante influence et auxquelles personne ne fait attention.}
\]

This is why inner speech has escaped the attention of most psychologists; it is like those active and modest people who, in a family or society, render a thousand services without demanding anything in return, from whom everyone receives beneficial influences and to whom no one pays attention.

Yet inner speech, or speech imagery, is a central component of our mental lives and so a complete understanding of cognition is impossible without examining it.

\(^1\)Though it has been the subject of philosophical discussion.
This dissertation is a step toward filling this gap in our understanding of cognition. In this dissertation I have proposed that humans have co-opted an aspect of the motor-system (corollary discharge) to serve the function of providing the auditory content of inner speech.

Corollary discharge is a fundamental component of the motor system; not just of humans but of any animal that moves. Corollary discharge is a sensory prediction; it represents the motor system’s prediction of the sensory consequences of its actions. Corollary discharge is generated by the motor system to fulfil two functions:

1. Providing short-latency feedback about the consequences of actions.
2. Segregating (or ‘filtering’) self-caused sensations from externally-caused sensations.

This dissertation argues that it is the second of these two functions that has been ‘recycled’ in human cognition to provide the ‘sound’ of inner speech.

The existence of corollary discharge was initially postulated in the motor-control literature for purely theoretical reasons — there seemed to be no way to achieve the ends of motor control without assuming that there was a sensory signal fulfilling these two functions. Later research found strong evidence to support this theoretical claim. Now we have behavioural, brain imaging, and neuroanatomical evidence supporting the existence of corollary discharge in a range of animals (from tadpoles to humans) in a variety of sensory modalities (from electrosensation to hearing; hearing being the modality of interest for this dissertation). A subset of this literature is reviewed in section 1.5.

The theory proposed in this dissertation is very simple. In speech, corollary discharge is produced by the motor system and this corollary discharge is a prediction of the sound of one’s own voice. Thus, corollary discharge is a priori very much like speech imagery — it is an internal sensory signal constituting a prediction of the sound of our own voice. This dissertation points out this similarity and

---

2 I am anthropomorphizing the motor system for the sake of exposition.
3 The motor system needs to predict the sensory consequences of its actions not just for sound, but for all sensory modalities. Thus, there are several modalities of corollary discharge. This dissertation only deals with the use of auditory corollary-discharge in speech imagery. It has been proposed (Grush 1995) that kinaesthetic corollary-discharge could serve a similar function in motor imagery.
argues that it is due to the fact that corollary discharge is what provides the sound of our voice when we engage in inner speech.

There are several consequences of the claim that inner speech involves corollary discharge.

Corollary discharge is a prediction of external sounds, thus its presence should influence the perception of external sounds. This is what was shown in the experiments of Chapter 2. These experiments demonstrated that when imagining speech sounds, such as /a'ba/ or /a'va/, the perception of ambiguous sounds (ambiguous between /a'ba/ and /a'va/) was altered to be perceived in line with what was being imagined. Similarly, the perception of these ambiguous sounds was altered by imagery of /a'pa/ or /a'fa/: the ambiguous sounds were perceived as being more similar to the imagined sound (so more /a'ba/ were heard when people imagined /a'pa/ and more /a'va/ were heard when people imagined /a'fa/).

The experiments described in Chapter 3 examine the duration and extent of the influences of this imagery by examining how speech imagery interacts with the effects of recalibration and adaptation. It was found that the impact of imagery on perception can be seen to linger even after the imagery has ceased.

The strongest evidence for the presence of corollary discharge in inner speech is presented in Chapter 4. Here, the sensory attenuation of corollary discharge is demonstrated to occur when people engage in speech imagery. The ‘filtering’ function of corollary discharge, by which self-caused sounds are segregated from externally caused sounds, is usually demonstrated experimentally by showing that an animal’s response to self-caused sounds is attenuated in comparison to an equivalent external sound. This is known as sensory attenuation and is considered the hallmark of corollary discharge.

In Chapter 4, I show that when imagined speech coincides with ‘real’ speech, then the impact of that ‘real’ speech is attenuated. This attenuation was measured by using a context effect, in which one sound alters the perception of its neighbours. The strength of this context effect was attenuated when speech imagery matched the context sound. Proper controls rule out the possibility that it is speech imagery per se that is responsible for this attenuation, and that it must be the match between the imagined and real sound that is responsible for this effect.

This is strong evidence that corollary discharge is present in inner speech.
Bibliography


Baart, Martijn, and Jean Vroomen. 2010. Phonetic recalibration does not depend on working memory. *Experimental Brain Research* 203:575–82. \(\rightarrow\) pages 71, 78

Baciu, Monica V., Christophe Rubin, Michel A. Décorps, and Christoph M. Segebarth. 1999. fMRI assessment of hemispheric language dominance using a simple inner speech paradigm. *NMR in Biomedicine* 12:293–298. \(\rightarrow\) pages 13


Bell, Curtis C. 1981. An efference copy which is modified by reafferent input. *Science* 214:450–453. \(\rightarrow\) pages 28


Bristow, Davina Josephine. 2006. Monitoring and predicting actions and their consequences. Ph.d., University College London. → pages [27]


*Developmental Review* 28:225–262. → pages 2


175


Ito, Masao. 2008. Control of mental activities by internal models in the

Formant transition-specific adaptation by lipreading of left auditory cortex N1m. *Neuroreport* 19:93–7. → pages[148]


Jacobson, Edmund. 1931. Electrical measurements of neuromuscular states
during mental activities. VIII. Imagination, recollection, and abstract thinking

Holt and Co. → pages[8]

Jaynes, Julian. 1976. *The origin of consciousness in the breakdown of the

In *Electroreception*, ed. Theodore Holmes Bullock, Carl D. Hopkins, and

Johnson, Keith. 2005. Neural correlates of inner speech and auditory verbal
hallucinations: a critical review and theoretical integration. *The handbook of
speech perception* 27:363–389. → pages[153]

Jones, Dylan M., William J. Macken, and Alastair P. Nicholls. 2004. The
phonological store of working memory: is it phonological and is it a store?
*Journal of experimental psychology. Learning, memory, and cognition*
30:656–674. → pages[156]

during phonation: studies of Mandarin tone production. → pages[17, 20]

The self across the senses: an fMRI study of self-face and self-voice

MRI study on the neural substrates for writing. *Human brain mapping*
13:34–42. → pages[41]


181


Plato. ???? Sophists. → pages


Quintillian. 1856. *Institutes of oratory*. London: Haddon Brothers. → pages


186


Watson, John B. 1913. Psychology as the behaviorist views it. Psychological Review 20:158–177. → pages[8, 12]


