Inner Speech: Philosophical and Psychological Investigations

by

Shivam Mahendra Patel

B.A. in Philosophy, University of California, Berkeley, 2011

Submitted to the Graduate Faculty of the

Dietrich School of Arts and Sciences in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2020

UNIVERSITY OF PITTSBURGH

DIETRICH SCHOOL OF ARTS AND SCIENCES

This dissertation was presented

by

Shivam Mahendra Patel

It was defended on

June 26, 2020

and approved by

James Shaw, Associate Professor, Department of Philosophy

Mark Wilson, Distinguished Professor, Department of Philosophy

Wayne Wu, Associate Professor,
Center for the Neural Basis of Cognition (Carnegie Mellon University)

Dissertation Director: Edouard Machery, Distinguished Professor,
Department of History and Philosophy of Science

Inner Speech: Philosophical and Psychological Investigations

Shivam Patel, PhD

University of Pittsburgh, 2020

This project investigates philosophical and psychological aspects of "that little voice in our head", or inner speech. Research on inner speech has been guided by the assumption that it is essentially a speech phenomenon. This dissertation argues against this assumption from three independent angles. In Chapter 2, I consider the idea that the content of inner speech is speech-specific. After arguing against this position, I go on to claim that the content of inner speech is vocalic. In Chapter 3, I examine models that treat inner speech as a prediction. I show that these models are problematic on both empirical and theoretical grounds. In their place, I argue for a model on which inner speech is treated as a goal state. In Chapter 4, I consider the popular idea that breakdowns in inner speech processing explain auditory verbal hallucination. I argue that inner speech-based explanations of AVH belong to a problematic class of psychological explanations, which confuse phenomenological and scientific understanding. My discussion of inner speech has implications for a wide range of topics, including metacognition, imagery, and the explanation of pathological mental states.

Table of Contents

List of Tables

List of Figures

Preface

I am indebted to many for this dissertation. My journey through philosophy started with Michael Holden and Bruce Hanson at Fullerton Junior College – from the former I learned to strive for clarity while from the latter I learned that philosophy can be edificatory. At UC Berkeley, I was fortunate enough to work with Barry Stroud and M.G.F. Martin, who remain for me exemplars of honesty and insight in philosophy. At the University of Pittsburgh, Edouard Machery appeared at a crucial time during my graduate career with generosity and humanity. A font of both professional and philosophical support, his quick-thinking and foresight have made this a better dissertation. My committee members have each in their own way contributed to my graduate career: Wayne Wu's seminars were highlights in pedagogy, James Shaw's honest feedback helped me avert potential slip-ups, and Mark Wilson's pastoral care was much appreciated. I also thank Peter Langland-Hassan for comments and a wonderful visit at the University of Cincinnati.

My friends and family have been there across the ups and downs. Daniel Foudeh has been an ever-present companion – there from the worried late-night call to the random gas station popsicle. Getting bites to eat and passing time with Cathy Carroll made graduate school go down easier. Going to ballet with LD and LT did as well. I would not be where I am without my parents. From my father, I received my initial philosophical temperament, and from my mother, I received unwavering love and support throughout my studies. I can say with confidence that without my mother, none of this would be. Finally, my partner, Zina Ward, has been a source of optimism and understanding throughout the completion of my dissertation. Her continuing excitement for the ideas in this dissertation has been more than enough to overcome my doubts.

1.0 Introduction

We pass much of our life doing something that resembles talking to ourselves. Whether in states of idle reverie, or focused, goal-directed thinking, we often engage in the activity of *inner speech*. Inner speech goes by a number of titles, including "thinking in words", "that little voice in your head", and "silent monologue". Sinclair Lewis points to the phenomenon in the course of describing the goings-on of one of his titular characters, George F. Babbitt:

> At the Nobby Men's Wear Shop he took his left hand off the steering-wheel to touch his scarf, and thought well of himself…and at the United Cigar Store, with its crimson and gold alertness, he reflected, "Wonder if I need some cigars – idiot – plumb – forgot – going t' cut down my fool smoking." (2010, p. 44)

Babbitt's one-word corrections – 'idiot' – his shortening of the preposition – 't'' – and his self-evaluations – 'fool smoking' – all point to the familiar phenomenon of thinking in words in one's head. Despite being widespread, it is not universal. Some people report no inner speech, while many of us who do engage in inner speech do not do so every moment of the day. Estimates have it that inner speech is present between 25-80% percent of our waking hours (Klinger and Cox, 1987; Heavey and Hurlburt, 2008). Despite its variable frequency, inner speech possesses a number of features that make its occurrence philosophically interesting.

Inner speech lies at a crossroads between thought, sensation, and language. As Peter Langland-Hassan and Agustin Vicente (2018a) point out, "inner speech is a paradigmatically *conscious* mental phenomenon, being the one obvious place where thought, language, and consciousness overlap" (p. 2). On the one hand, inner speech seems to be connected to our thinking in some way. After all, Babbitt was *thinking* about cigars and smoking when he engaged in his silent monologue. And, if someone asks what you are thinking, it is no accident that your answer – if truthful – will give expression to whatever is being logged by your inner speech. On the other

hand, inner speech also seems to involve sensation. The little voice often involves the experience of sounds. Alex Byrne (2009) goes so far as to claim that 'the unenlightened' – the folk – in fact believe that there *are* sounds occurring in their heads. Furthermore still, inner speech involves an element of language: syntax, speech sounds, and speech articulation seem to be at play. At first glance, then, inner speech combines thought, sensation, and language within a single mental state.

This combination makes inner speech ripe for philosophical study. Although philosophers have been interested in the trio of thought, sensation, and language, they have typically been interested in them two at a time. Thus, we find philosophers and cognitive scientists interested in the phenomenology of thought (thought and sensation), the relationship between the structure of thought and the structure of language (language and thought), and whether language abilities influence our perceptual experience (sensation and language). Part of the philosophical interest in inner speech, then, is that it combines these three domains into a single mental state. Moreover, it is precisely its connection with all three topics that seems to allow inner speech to play a role also in conscious thought and self-knowledge. The idea that inner speech has a role in conscious thought is nicely expressed by Socrates:

> Socrates: I speak of what I scarcely understand; but the soul when thinking appears to me to be just talking – asking questions of herself and answering them, affirming and denying. And when she has arrived at a decision, either gradually or by a sudden impulse, and has at last agreed, and does not doubt, this is called her opinion. I say, then, that to form an opinion is to speak, and opinion is a word spoken, –I mean, to oneself and in silence, not aloud or to another: What think you?
> Theaetetus: I agree. (*Theaetetus*, 190a)

Inner speech may not only make conscious thought possible, it also appears to ground self-knowledge. A version of this view has been attributed to Ryle (2009):

> We eavesdrop on our own voiced utterances and our own silent monologues. In noticing these we are preparing…to describe the frames of mind which these utterances disclose. (p. 165)

On Ryle's picture, inner speech seems to be revelatory of our thoughts in a way that makes them conscious and knowable. Another reason that inner speech is of philosophical interest, then, is that it seems to have direct bearing on topics that have been of perennial philosophical interest.

This dissertation will suggest that previous discussions of conscious thought and self-knowledge have faltered because philosophers have often operated with a naïve conception of inner speech. If we are to have any chance of understanding whether and how inner speech plays a role in conscious thought and self-knowledge, we first need to get a clearer understanding of the nature of inner speech and its role in functional architecture. These foundational topics will concern the Chapters of this dissertation.

So what is inner speech? If you concur with the suggestion I made above – that inner speech is at the crossroads between thought, sensation, and language – you might endorse the definition that Perrone-Bertolotti et al. offer in their 2014 review of the literature: inner speech is the mental simulation of speech (p. 221). Since actual speech production involves movement of the vocal tract and the generation of speech sounds, on Perrone-Bertolotti et al.'s definition, inner speech would be the simulation or representation of vocal tract movements and speech sounds. However, if this is intended as a definition that is supposed to capture all and only cases of inner speech, Perrone-Bertolotti et al.'s definition falls far from its intended target. The problem with the definition is that there seem to be forms of inner speech that implicate neither a representation of vocal tract movements nor speech sounds. For example, there seem to be forms of inner speech that represent words alone: one of Christopher Gauker's students reports "never hearing a voice" but that she only "experiences words – just words" (2018, p. 59). In a similar vein, Langland-Hassan (forthcoming) has recently discussed problems with the following definition of inner speech provided by Alderson-Day and Fernyhough (2015): inner speech is the subjective

experience of language in the absence of overt and audible articulation. Although the definition embraces abstract forms of inner speech that involve the representation of words only, it fails, as Langland-Hassan points out, to allow for the possibility of unconscious inner speech that may occur while we speak (Langland-Hassan, forthcoming). As Langland-Hassan notes, Alderson-Day and Fernyhough's definition is reduced to "inner speech is inner language". The definition fares no better than the colloquial characterizations with which the introduction opened, "thinking in words", "that little voice in your head", and "silent monologue".

Langland-Hassan's solution to the problem is to treat these not as theoretical definitions, capturing all and only cases of inner speech, but at most as "rough-and-ready" pointers that use certain salient or common features to pick out inner speech. I agree with Langland-Hassan that definitions of inner speech presented in the literature are at most "rough-and-ready" pointers. I think that the failure to develop a theoretical definition of inner speech also reveals a deeper anxiety about circumscribing the topic of inner speech. I think that part of the reason that theorical definitions of inner speech are problematic concerns the *plurality* of inner speech. There are at least two respects in which inner speech is pluralistic: in terms of its phenomenology and in terms of its empirical investigation.

The phenomenology of inner speech seems to be multifarious. Many forms of inner speech seem to be sensorimotor in character. Thus, when engaging in inner speech, I typically feel it to be auditory, and I even find that as I write this I am breathing and silently articulating in cadence with the words I am producing in inner speech. But, as I indicated above, at times our inner speech seems to be linguistic while being neither motoric nor auditory. In these cases, I have words popping into my head without a sense of pronouncing them. Finally, at the limit, there seem to be cases of inner speech that are 'unworded' – at most implicating a propositional content. Although

I noted that inner speech combines sensation, language, and thought in a single mental state, there are forms of inner speech that are reflective of one or another of these categories to the exclusion of the others.

The phenomenological plurality of inner speech is compounded by the fact that researchers do not agree on which tasks should be used to study inner speech. Many researchers use experimental protocols that track auditory and motor forms of inner speech. These protocols explicitly require subjects to imagine moving their mouth while imagining a speech sound (e.g., Tian and Poeppel, 2010). Even more common are tasks that measure only the auditory component of inner speech. Such tasks include homophone and rhyme tasks (Geva, 2011). In the former, subjects are presented with two written words that sound the same but are spelled differently and are asked to silently judge whether the two words sound the same, while in the latter, subjects are presented with two or more words and are asked whether they rhyme. These tasks presumably require representing the auditory properties of the words. Further still, a number of researchers use paradigms that remain neutral on the form of inner speech being studied. Central among these are silent reading tasks used in both behavioral and imaging paradigms (Kell, 2017).

The phenomenological and experimental pluralism associated with inner speech threatens not only definitions of inner speech, as noted above, but more fundamentally, whether inner speech is a *kind* at all. In the first place, it is difficult to see what is in common between representations of propositions, representations of words, and representations of motor and sensory information. If the phenomenology is to be trusted, inner speech just seems to be a disjunction of language-related mental states (cf. Cho and Wu, 2014). In addition, if researchers use disparate tasks in measuring inner speech, this raises the concern that the object of investigation is shifting across different empirical studies of inner speech. The plurality of forms of inner speech threatens the

idea that inner speech is a kind of mental state in its own right as opposed to a mere disjunction of mental states.

The threat that the plurality of inner speech poses to its kindhood has been underexplored in the literature. The central guiding question of the dissertation concerns how the plurality of inner speech is to be reconciled with the unity of inner speech. On this guide to reading the dissertation, Chapter 2 provides an account of what unifies inner speech despite its plurality, while Chapters 3 and 4 provide a defense of this account by blocking traditional explanations of inner speech and auditory verbal hallucination that would undermine it.

In Chapter 2, I provide an account of the content of inner speech. Existing views about the content of inner speech are unable to account for what unifies the plurality of inner speech. These accounts share the common assumption that the contents of inner speech are derived from the run-up to speech production: propositions, syntax, phonemes, phones, and vocal gestures. However, once this assumption is made, we face the problem that there is nothing about representing a proposition or representing syntactic structure that would count these episodes as being instances of inner speech. In addressing this problem, I argue that we have to broaden our view of the processes that underlie inner speech. In particular, I argue that inner speech is a product of a complex set of processes that target *voice*. On this basis, I claim that the content of inner speech is vocal: what unifies inner speech is that we represent voices communicating information.

However, there remain obstacles to providing an answer of this shape. The most serious obstacle stems from predictive models of inner speech. In Chapter 3, I take on predictive models directly, while in Chapter 4, I challenge accounts of auditory verbal hallucination that provide indirect support for them. Predictive models of inner speech hold that inner speech is identical to an auditory prediction that is otherwise used to guide online speech production. One fundamental

feature of this prediction is that it concerns events in the world: were such and such movement made by the speech articulators, a prediction is generated that such and such auditory event will occur in the world. In the first instance, predictive models of inner speech fail to accommodate the plurality of inner speech, since the predictions they identify with inner speech are auditory predictions. Thus, predictive models of inner speech are at best models of auditory forms of inner speech. Moreover, it is unclear how these models could be extended to cover more abstract, higher-level forms of inner speech, such as syntactic inner speech, whose contents have no realization in the world to which it can possibly correspond. Thinking of inner speech in terms of a prediction is therefore incompatible with the commitment to the plurality of inner speech I adopt in Chapter 2.

Although predictive models fail to accommodate the plurality of inner speech, it has been thought there is a wealth of empirical support in their favor. Chapter 3 argues that predictive models of inner speech are empirically unsupported and theoretically problematic. I argue that empirical evidence in support of the position is thus far non-existent, and that empirical considerations in fact bear against such models. I then move to evaluate a subclass of predictive models of inner speech, predictive processing models. I show that if inner speech, and imagery more generally, are characterized in terms of predictive processing models, these models will have to give up their ability to account for perception and action.

Chapter 4 takes on another obstacle to adopting the pluralist position set out in Chapter 2. Most theorists account for auditory verbal hallucination (AVH) in terms of predictive models of inner speech. According to these accounts of AVH, a breakdown in predictive processes leads to the feeling that one's inner speech is authored by some external agent. The widely accepted

success of these inner speech models of AVH provides indirect support for predictive models of inner speech, which, in turn, puts pressure on the results of Chapter 2.

Chapter 4 argues that explanations of AVH in terms of inner speech are part of a broader class of problematic explanations of AVH. This class, which I label 'substrate-based explanations', seeks to explain AVH in terms of normal mental state kinds whose phenomenology matches that of AVH. Although this style of explanation allows us to come to grips with the phenomenology of AVH – it allows us to grasp what it is like to undergo an AVH – it does not provide for an explanation of AVH. I claim that substrate-based explanations rest ultimately on a confusion between scientific and phenomenological understanding of mental illness. By disarming inner speech models of AVH, we are able to fend off remaining indirect support for predictive models of inner speech, and thereby rescue the shape of account provided in Chapter 2.

The dissertation can thus be read as an attempt to accommodate the plurality of inner speech without giving up on its kindhood. It does this by broadening our view of what unifies inner speech, while also defending the plurality of inner speech from modes of explanation that would reduce it to a phenomenon with only a single manifestation.

2.0 From Speech to Voice: On the Content of Inner Speech

2.1 Introduction

What goes on when we think? James Joyce comes pretty close to capturing it in his *Ulysses*. Take a peek into the head of Molly Bloom:

> …let me see if I can doze off 1 2 3 4 5 what kind of flowers are those they invented like the stars the wallpaper in Lombard street was much nicer the apron he gave me was like that something only I only wore it twice better lower this lamp and try again so as I can get up early… (p. 930)

Joyce provides us with an unvarnished glimpse into Bloom's train of thought by allowing us to eavesdrop directly on it. The feeling that we are eavesdropping is partly underwritten by the fact that much of our own thinking is as fragmentary and disjointed as Bloom's. Thought whose aim is sleep gives way to a meandering fantasy about wallpaper and an apron only to return, after a brief hiatus, back to the subject of sleep. We also have the feeling of eavesdropping because we ourselves – although not all of us[1] – often think in words. Thinking often involves a little voice in our head, or what philosophers and psychologists have come to call "inner speech".

Given his medium, Joyce is forced to present inner speech as being clear and consistent: Bloom is presented as thinking in whole words, one after another, each pronounced to its completion. However, as I think Joyce would attest, the fragmentary and disjointed character of our thought is matched by the diverse forms of our inner speech. Sometimes when I write, I type out a sentence while moving my mouth in unison; but a few words into the sentence, my lips now

[1] Depending on the protocols being used, authors estimate that we engage in inner speech up to 80% of the time (Klinger and Cox, 1987). Heavey and Hurlburt (2008), however, claim that inner speech is present only 25% of the time.

pressed together, I find whole words popping into my head without any corresponding motor sensation; soon enough I have a full-blown auditory experience, as I return to a word to utter it in my head just as I would have heard it aloud; and finally, stepping back to consider the whole of what I have just written, in my head I affirm the bare thought to some authoritative, generalized other, but without an auditory or linguistic garb. In sum, inner speech is more a shape-shifting menagerie, taking on different forms as it unfolds, than a consistent march of fully pronounced words.

In this Chapter I will show that this plurality of inner speech tells against a standard picture of its content. According to a widely accepted view of speech production, in the run-up to generating an utterance we first select a proposition, then select words to express the proposition, then speech sounds to express those words, then motor commands to create those speech sounds, and, finally, we execute those commands, thereby generating an utterance of the original proposition. Speech production thus involves a hierarchical process from the selection of a proposition through to the execution of motor commands (see, e.g., Levelt (1993)). This picture of speech production shapes a corresponding standard picture of inner speech. According to this picture, inner speech is a truncated version of speech production: we start with the selection of a proposition and move down the hierarchy, but unlike speech production, processing is cut short. The result is that I have the conscious experience of speech, but do not produce speech. As Oppenheim and Dell (2010) and Perrone-Bertolotti et al. (2014) note:

> Inner speech is generally thought of as the product of a truncated overt speech production process. Theories differ, however, about where this truncation lies… (Oppenheim and Dell, 2010, p. 1147)
>
> Inner speech can be seen as truncated overt speech, but the level at which the speech production process is interrupted (abstract linguistic representation vs. articulatory representation) is still debated. (Perrone-Bertolotti, 2014, p. 235)

As a truncated version of outer speech, according to these authors, inner speech engages a subset of the processes implicated in outer speech.

A background assumption of this standard picture of inner speech is that the content of inner speech is speech-based. The assumption seems to be a natural one: if inner speech is a truncated version of outer speech, then inner speech represents just those contents represented in the run up to outer speech. My aim in this Chapter is to argue that this standard picture of the content of inner speech fails to capture its highly variegated character. In short, if the contents of inner speech are speech-based, then we are unable to account for how there could be abstract forms of inner speech, including propositional and syntactic inner speech. By broadening our view of the processes that are implicated in inner speech, I will contend that inner speech is marked not by the possession of speech-related content, but by structurally complex, voice-based content.

In Section 2.2, I explain in more detail the notion of a speech processing hierarchy, and characterize three views concerning the current debate about the content of inner speech – *concretism*, *abstractionism*, and *standard pluralism*. In Section 2.3, I show that empirical evidence is split between the two monist positions, concretism and abstractionism. In Sections 2.4 and 2.5, I provide theoretical challenges to Peter Langland-Hassan's argument for concretism and Christopher Gauker's brand of abstractionism. This leaves standard pluralism as the most plausible view about the content of inner speech, which I explain in Section 2.6. However, in Section 2.7, I argue that even standard pluralism fails to account for the sheer diversity of inner speech. In Section 2.8, I show that the existing menu of views about the content of inner speech go wrong in their too-narrow focus on the resources of the speech processing hierarchy. In an attempt to expand our view about the contents of inner speech, I show that speech processing is only one component of a number of processes centered on voice. In light of this broader view, in Section 2.9, I present an alternative position, *vocalism*, according to which the content of inner speech is vocalic: during an inner speech episode one represents a voice communicating

information. With the new position in hand, in Section 2.10, I close by showing how vocalism captures both the plurality and the unity of inner speech.

2.2 A Menu of Views

Existing views concerning the content of inner speech are best understood against the background of the speech processing hierarchy. The speech processing hierarchy is a bidirectional hierarchy of processing levels implicated in the production and perception of speech (see Figure 1) (e.g., Levelt, 1993).[2] Top-down processing subserves speech production, while bottom-up processing subserves speech perception.

[2] To be clear, the notion of a speech processing hierarchy I am employing is a bit of an idealization. Psycholinguists tend to hold different views about how information flows through the hierarchy. For example, some psycholinguists hold that the hierarchy is serial while others that it is parallel, some that the information flow is feedforward others that it also feedback. In fact, there is also disagreement about the exact order of operations and sub-operations within the hierarchy (e.g., Fromkin, 1971; Dell, 1986, Bock and Levelt, 1994). Despite these differences, psycholinguists tend to agree on the organization presented in Figure 1.

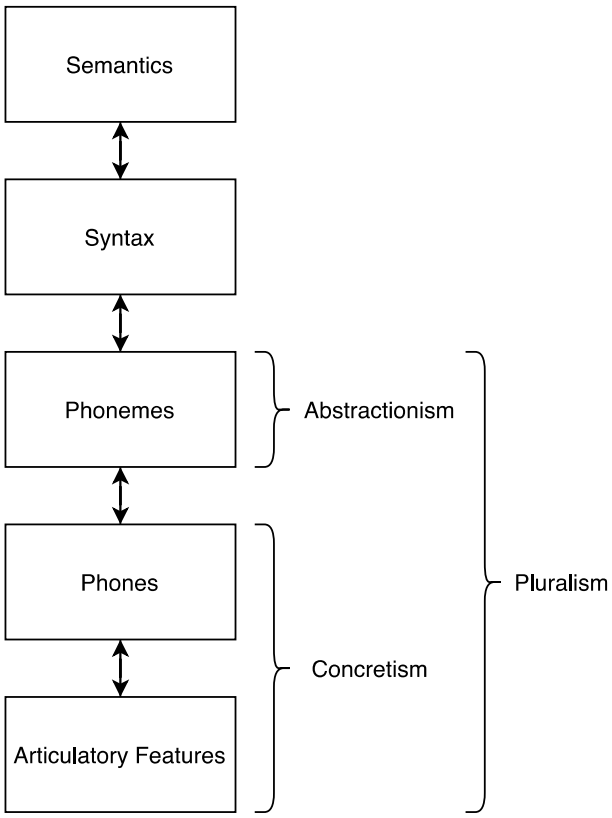

Figure 1: The speech processing hierarchy with views about the contents of inner speech

I will understand the topmost level of the hierarchy – 'Semantics' – as generating propositional contents. These can either be complete contents – JOHN IS AT THE MEETING – or partial contents – JOHN IS AT…. The next level – 'Syntax' – generates an abstract frame populated with words along with a specification of their syntactic roles. We thus have a content of the following (rough) form: *John* (Subject) *is* (Verb) *in* (Preposition) *the meeting* (Object). In the next level – 'Phonemes' – words are populated by phonemes. I will understand a phoneme as the smallest unit of speech that distinguishes one word from other words within a language. For example, the word *pat* is distinguished from the word *cat* by the phoneme /p/. Although there is much controversy about the nature of phonemes, I will understand a phoneme as a set of similar speech sounds. Given that a phoneme can be accessed independently of accessing any of its

member speech sounds (see Figure 1), I will understand a phoneme as non-sensory.[3] In the next level – 'Phones' – phonemes are further specified in terms of phones. A phone is a speech sound that is a member of a phoneme. Where *leaf* and *pool* both contain the phoneme /l/, each uses a different phone – *leaf* uses [l] (clear l) and *pool* uses [ɫ] (dark l). Finally, articulatory features are sets of motor commands for producing a phone. For example, the instruction for producing the auditory characteristics of [p] is {[labial, -round], [-voice], [+stop]}. That is, [p] is produced when both lips are pressed together and there is stoppage, build up, and abrupt release of airflow without vibration of the vocal folds.[4]

Each level of the speech processing hierarchy is thus associated with a particular type of content. Perhaps the most intuitive view about the content of inner speech is that it is phonetic and/or articulatory. Langland-Hassan (2018) notes that this view has seemed to many authors to be a "truism, a platitude of common sense" (p. 79). According to this view, which I will label *concretism*, inner speech episodes represent speech sound/articulatory content (i.e., sensorimotor content) (Langland-Hassan (2014, 2018)). Though naïve introspection seems to reveal that inner speech is phonetic and/or articulatory, some have thought that introspection is not a reliable basis on which to determine the contents of inner speech. According to a number of views, which I will group under the label *abstractionism*, inner speech episodes represent phonemic contents (i.e., non-sensorimotor content) but nothing lower down the speech processing hierarchy (e.g., MacKay (1992) and Gauker (2018)). In contrast to both of these positions, some authors have been open to the possibility that inner speech takes on different forms depending on context: during default contexts, inner speech is phonemic, but during "stress and cognitive challenge," inner speech

[3] See Langland-Hassan (2018) for a contrasting position on phonemes.
[4] Although articulatory features are often characterized as 'phonetic features', this has to do with the fact that one strand of phonetics concerns articulatory phonetics while another concern acoustic phonetics.

becomes phonetic and/or articulatory (Alderson-Day and Fernyhough, 2015, p. 933). According to *standard pluralism*, inner speech episodes have speech sound/articulatory content in some contexts, while in others they have phonemic content (e.g., Oppenheim and Dell (2010) and Alderson-Day and Fernyhough (2015)).

The speech processing hierarchy, and the levels of content it makes available, thus forms the background of the debate over the content of inner speech (see above, Perrone-Bertolotti, 2014, p. 235). The debate is over whether inner speech engages sensorimotor levels (concretism and abstractionism) and whether the views are exclusive (standard pluralism).[5] In this Chapter, I will not adjudicate the debate between concretism, abstractionism, and standard pluralism. Rather, I will reject the assumption that serves as common ground for the debate: that the contents of inner speech are derivable from the speech processing hierarchy.

2.3 Empirical Evidence is a Draw: Assessing Concretism and Abstractionism

I first consider the two monist positions – concretism and abstractionism. These views have standardly been adjudicated using empirical evidence. In this Section, I suggest that the conflicting empirical evidence regarding concretism and abstractionism seems to provide some initial motivation for standard pluralism.

[5] I will remain neutral on the question of whether distinct types of content are matched by distinct types of vehicles. At one extreme, one may posit a distinct vehicle type for each distinct content type. On this view, propositional contents are subserved by one vehicle type, while phonemic contents would be subserved by another vehicle type, etc. At the other extreme, one may posit a single vehicle type for each distinct content type. I will use the generic term 'representation' (as opposed to 'x-type representation', e.g., auditory representation) so as to remain neutral on this issue.

On the one hand, there is much empirical evidence of the presence of sensorimotor contents in inner speech. Using electromyography (EMG), McGuigan and Dollins (1989) found differential activity when children produced "P" and "T" in inner speech: reading "P" elicited lip movements, whereas "T" elicited tongue movements. Given that these articulatory gestures mirror those used to generate the sounds aloud, this result suggests that inner speech represents articulation. In other behavioral evidence, Scott et al. (2013) found that subjects who produce a segment in inner speech (e.g., [p]) while hearing an ambiguous speech sound tend to report hearing the identity of the ambiguous speech sound as that of the segment produced. This result suggests that inner speech represents speech sounds. Such a conclusion is also supported by a convergence of fMRI data showing that inner speech activates the superior temporal sulcus (STS) and other areas related to auditory processing (e.g., see Kell (2017)).

Despite such evidence in favor of concretism, there is also evidence that inner speech does not possess sensorimotor contents. Bishop and Robson (1989) have shown that children who are unable to control their vocal tract from birth, and so unable to refine speech motor commands through use, are nevertheless able to engage in intact inner speech. One plausible interpretation of this evidence is that although motoric contents may often accompany inner speech, such contents are not part of inner speech proper. In another influential behavioral study, Oppenheim and Dell (2008, 2010) reasoned that if inner speech represented speech sounds, then given that speech sounds are distributed in a similarity space, we should expect errors on tongue twisters recited in inner speech to reflect that similarity space. The researchers did not find this pattern, but did find substitutions at the level of phonemes. Their results have suggested to some that inner speech does not possess sensorimotor contents, but nevertheless does possess abstract, phonemic contents.

On the neuroimaging front, there has been much research comparing inner speech with 'inner signing', which is present in deaf people. In one such study, Rudner et al. (2012) uncovered broad overlaps in representations of phonemes across both speakers and signers. The authors found that any differences were "located anterior to sensory and motor areas", and claimed that "it seems likely that [phonological] differences relate to abstract rather than surface differences" (p. 664) (see also McGuire et al. 1997). These lines of evidence suggest, contrary to concretism, that inner speech possesses non-sensorimotor contents.

On its face, this conflicting empirical evidence appears to make both forms of monism – concretism and abstractionism – untenable. As such, it provides some motivation to embrace a standard pluralist position about the content of inner speech. However, proponents of the individual monist positions can always try to reinterpret and explain away inconvenient empirical results. In Sections 2.4 and 2.5, I look at specific versions of concretism and abstractionism and show why the positions are theoretically problematic. The arguments to follow will show that recent defenses of monism flout existing theoretical work in psycholinguistics.

2.4 Against Langland-Hassan's Concretism

Peter Langland-Hassan (2018) has argued that inner speech always has an "auditory-phonological" or speech sound component.[6] Although this view has been taken to be a "truism, a platitude of common sense", Langland-Hassan seeks to provide an argument in its favor (p. 79).

[6] Recall that Langland-Hassan (2018) believes that phonemes are auditory. Although I have denied this (see Section 2.2), for the sake of the present argument, I will use 'phonological' in the sense that Langland-Hassan intends in this Section.

Langland-Hassan starts with the fact that we know which language our inner speech is in, e.g., whether it is in English, French, Spanish, etc. Langland-Hassan then engages in an inference to the best explanation, seeking to explain how it is that we know the language of our inner speech. He considers "the most salient features of words and sentences and [asks] whether those features might reveal to us the language in which they occur" (p. 83). Langland-Hassan runs through four features: semantics, syntax, phonology, and graphology. The semantics of a sentence cannot ground knowledge of the language of inner speech, since, according to Langland-Hassan, semantics is held constant across languages. Moreover, the syntax of a sentence is unable to ground such knowledge, since syntactic frames cannot distinguish between different sentences across certain languages. Langland-Hassan asks us to "imagine that we were able to "see directly" the [syntactic] structure of a sentence, abstracting away from its specific words" (p. 83). In this context, although we would know that a given sentence had a subject-verb-object (SVO) structure, we would not know the words that fill in that structure (e.g., we would not know that the frame was filled in by *John*, *likes*, and *ice cream*). Given that the syntactic frame is shared across 'SVO' languages, one would not know whether the sentence in question is one of English, Spanish, French, or a number of other 'SVO' languages. Thus, according to Langland-Hassan, syntax is unable to ground knowledge of the language of inner speech. Having excluded semantics and syntax, Langland-Hassan moves to consider graphemes. Graphemes, according to Langland-Hassan, cannot explain how we know the language of our inner speech since grapheme identification is visually-based, whereas inner speech is not visually-based. This leaves only one possibility: there must be an auditory component of inner speech that accounts for our knowledge of which language our inner speech is in.

This argument is not compelling, however, because Langland-Hassan is mistaken in thinking that syntax cannot ground our knowledge of the language of inner speech. Langland-Hassan's discussion seems to stem from W.J.M. Levelt's classic model of speech production, which has been echoed in a number of more recent models (Levelt (1993); see also (Roelofs et al. (1998)). Levelt distinguishes between two types of representation of a word: a lemma and a lexeme. The lemma is a representation of a word's semantic and syntactic structure, while a lexeme is a representation of a word's morphophonological form. According to Levelt, lemmas are selected prior to the selection of lexemes, and so are pre-phonological/auditory. Although Langland-Hassan fails to mention the lemma/lexeme distinction, he would argue that lemmas represent only the semantic and syntactic structure of a word, but not the *word* whose semantic and syntactic structure it is (Langland-Hassan, personal communication). For example, the lemma for *cake*, according to Langland-Hassan, would represent its referent (semantics) and that it is a noun (syntax), but would not represent the identity of the word whose semantic and syntactic properties are in question – that the word in question is *cake*. This is important for Langland-Hassan because if lemmas did represent the identity of the word, then it would follow that knowing which lemmas occur in one's inner speech would be sufficient for knowing which language one's inner speech is in.

The problem is that this view of lemmas is contradicted by existing psycholinguistic work. Theorists like Levelt believe that lemmas do represent the identity of words alongside their semantic and syntactic properties. Consider a concrete example quoted verbatim from Levelt (1993):

> give: conceptual specification:
> CAUSE (X, (GOposs (Y, (FROM/TO (X, Y)))))
> conceptual arguments: (X, Y, Z)
> syntactic category: V
> grammatical functions: (SUBJ, DO, IO)

relations to COMP: none
lexical pointer: 713
diacritic parameters: tense
aspect
mood
person
number
pitch accent
Figure 6.3
Lemma for *give* (p.191)

Notice that the lemma for *give* represents not only its semantic and syntactic properties, but also the word itself. Indeed, if the identity of the word were not represented, then it is difficult to understand how there could be such a thing as selecting the correct set of phonemes for a given lemma. That is, if all one knows is that something refers to a particular set of items and that it is a noun, it is difficult to see how one would be able to even get a start on figuring out which word to pronounce. Therefore, against Langland-Hassan, it seems that pre-phonological/auditory words provide a possible ground for knowing the language of one's inner speech.

But suppose we concede to Langland-Hassan that lemmas do not represent word identity. After all, psycholinguists tend not to be entirely clear in their informal discussion about whether lemmas represent word identity. Still, there are a number of speech-related phenomena that put pressure on the idea that the auditory contents of inner speech ground our knowledge of its language. In the tip-of-the-tongue (TOT) phenomenon, speakers have the compelling feeling of either knowing a word or almost retrieving a word but being unable to "get it". Although speakers sometimes subpersonally represent auditory information during TOT, including stress, meter, and the initial consonant of the word, there are cases of TOT in which there is no representation of auditory information about the word (Schwartz, 2001). In these latter cases of TOT, speakers are nevertheless able to identify semantic and syntactic properties of the word, and, most importantly for our purposes, *that the word is in a particular language*. Now, whatever explanatory posit is responsible for the speaker's knowledge that the word is in English when undergoing a TOT, that

same posit is also sufficient to explain that knowledge when the word is actually being pronounced. Because the explanatory posit is consistent with a TOT state it must be non-auditory. If we assume that this non-auditory posit is present during inner speech as well, then, just as our knowledge of the language of outer speech can be explained in terms of it, so too can our knowledge of the language of inner speech. A representation with non-auditory content therefore explains our knowledge of the language of inner speech.

Finally, Langland-Hassan's inference to the best explanation is itself suspect. There is nothing in Langland-Hassan's argument that bars its application to (outer) speech production. But if we apply the argument to speech production, we are led to a bizarre conclusion: that I know that I am currently speaking English because I make the discovery that the auditory stream I produce is in English. After all, it seems that I can know that I am speaking an English sentence even if my ears are completely plugged and my facial bones are unable to conduct energy. (The sentence may end up being garbled due to the lack of feedback, but I presume it would still count as a sentence of English and I would know it to be a sentence of English.) Now, if Langland-Hassan's argument seems suspect when applied to outer speech, I see no reason it should be compelling for inner speech.[7] I therefore conclude, on both psycholinguistic and philosophical grounds, that Langland-Hassan's argument fails to show that auditory contents are always present in inner speech.

[7] This line of argument puts into relief a plausible alternative for knowing the language of my inner speech: my knowledge that I am speaking English during inner or outer speech is *non-observational* in just the way that my knowledge that I am grabbing a glass may be non-observational (see Anscombe (1957) and a healthy lineage that has followed). On this alternative explanation, I know that my inner speech is in English because I *use* English words in my inner speech – and this knowledge need not be grounded in knowledge of auditory sensation. Although Langland-Hassan seems to restrict the knowledge under discussion to knowledge gained by introspection (p. 92), the alterative I have mentioned here rejects that restriction. Though inner speech may indeed be a phenomenon open to introspection, it would not follow that all of its properties – e.g., that it is in English – need to be *known* via introspection.

2.5 Against Gauker's Abstractionism

In contrast to Langland-Hassan, Christopher Gauker has argued for a form of abstractionism, according to which inner speech never possesses auditory content. Whereas Langland-Hassan is moved by the introspective character of inner speech, Gauker claims that introspection fails to distinguish between inner speech, on the one hand, and the auditory imagery *of* inner speech, on the other hand. Gauker motivates his position by emphasizing the analogy between inner speech and (outer) speech production. Just as the *production* of speech involves the production of sound waves and the *perception* of speech involves processing those sound waves and thereby hearing speech, so too we should distinguish between *inner speech*, on the one hand, and the *perception* or *auditory imagery* of inner speech, on the other. On Gauker's so-called "perception theory of the auditory imagery of inner speech", the auditory imagery of inner speech misrepresents inner speech as involving speech sounds. The inner speech itself, however, never represents speech sounds.[8]

I raise both empirical and theoretical problems with Gauker's position.

For Gauker auditory contents are associated with the *auditory perception* of inner speech, and are not present in inner speech as such. Hence, we can challenge Gauker's brand of abstractionism if we find that deficits in auditory perception do not lead to deficits in auditory inner speech. Indeed, there are a number of case studies of "pure word deafness" that exhibit this

[8] There is some difficulty in interpreting Gauker's position on precisely what constitutes inner speech. On the one hand, Gauker claims that "auditory imagery of inner speech systematically represents inner speech as *sounds*" (italics mine, p. 60). This suggests that, according to Gauker, inner speech is not constituted by sounds. The problem is that all theorists are in agreement that inner speech does not involve actual sounds, and so, if this interpretation were correct, then all theorists would be error theorists like Gauker. A more plausible, alternative interpretation is that Gauker claims that inner speech does not *represent* speech sounds. This claim is controversial, and so sets Gauker off from other theorists. I will interpret Gauker as claiming that inner speech does not represent speech sounds.

structure. People with pure word deafness are able to hear environmental sounds, but are unable to hear speech – to them, speech sounds like "mumbling", "noise", or "a foreign language" (e.g., see Klein and Harper, 1956). Despite this deficit in auditory-verbal perception, people with pure word deafness nevertheless seem to have intact auditory inner speech. As Hemphill and Stengel (1940) note, "in the picture of word-deafness there were no paraphasias and the inner speech remained undisturbed" (p. 251). And the authors go on to discuss a patient with word-deafness where "the inner language is totally unaffected" (p. 258). Furthermore, Rapin (1985) claims that word-deaf patients often report being unable to understand themselves speak, but nevertheless have an "internal language", including the ability to read silently with comprehension, an indication of the presence of inner speech. Although these authors did not perform tests for auditory inner speech, Marshall et al. (1985) tested for auditory inner speech in a patient with auditory agnosia. The authors found that their subject was able to silently judge which of a set of written words rhymed with a target written word. Rhyming tasks plausibly require the representation of speech sounds, and so seem to be an adequate test for the presence of auditory inner speech (Langland-Hassan, 2014). In conflict with this data, Gauker's theory predicts that people with pure word deafness and auditory agnosia should not be able to engage in inner speech with associated auditory contents, given that the auditory centers of speech perception are corrupted.

Gauker's theory is also contradicted by research in psycholinguistics. Much theoretical work in psycholinguistics suggests that inner speech does represent auditory contents independently of the auditory perception of inner speech. According to Levelt's influential theory of speech control, a phonetic plan is generated by the *speech production module*, and serves as input to the speech perception module, which assesses the plan for accuracy (p. 470). Levelt's

picture suggests that representations of speech sounds are present in the speech production module independent of the speech perception module. In fact, Levelt explicitly identifies inner speech with the phonetic plan as it arises in speech production. Therefore, although Gauker may be correct to draw a distinction between inner speech and the auditory imagery of inner speech, that distinction does nothing to support the view that inner speech does not represent speech sounds on its own.

2.6 Standard Pluralism

I have so far shown that the empirical data does not support monism (Section 2.3), and, in addition, that the most prominent monist positions concerning inner speech seem to flout existing psycholinguistic theory (Sections 2.4 and 2.5).

The failures of monism help explain why standard pluralism is currently the most popular view of the contents of inner speech. According to standard pluralism, in some contexts, inner speech possesses contents characterized in terms of speech sounds and/or vocal tract articulation (i.e., sensorimotor contents), and, in other contexts, inner speech represents phonemes (i.e., non-sensorimotor contents). For example, in their recent review of the literature, Alderson-Day and Fernyhough (2015) conclude that "the core of inner speech is [an] abstract code containing a combination of semantic, syntactic, and phonological information" that can nevertheless be "unpacked" in terms of representations of speech sounds and articulation (p. 950). Oppenheim and Dell (2010) hold a similar position. According to them, a "shortcoming" of concretism and abstractionism is that both views "[conceive of] inner speech as a stable, consistent phenomenon" (p. 1150). In line with their "flexible abstraction hypothesis", the authors "hypothesize that

activation could be restricted to the level of phonemes in one situation…but strongly activate articulatory features in another" (p. 1150). In contrast to various forms of monism, standard pluralism is popular because it allows for a diversity of inner speech contents that seems to reflect the empirical evidence.

2.7 Moving Beyond the Speech Processing Hierarchy

In the current Section, I will argue that phenomenological considerations regarding inner speech not only tell against standard pluralism, but also against the guiding assumption that the content of inner speech is to be derived from the speech processing hierarchy. As we will see, phenomenological reports of inner speech suggest that there are forms of inner speech that are propositional and syntactic without engaging levels lower down the hierarchy. However, as we will find, the resources of the speech processing hierarchy are unable to account for such forms of inner speech. Thus, although the argument that follows nominally targets standard pluralism, it more fundamentally targets a common assumption shared by concretism, abstractionism, and standard pluralism, namely, that the contents of inner speech are to be derived from the speech processing hierarchy.

Theorists are beginning to take seriously the thought that inner speech is not a phenomenological monolith, but possesses a phenomenology that is highly variegated. Some of this work is based on Descriptive Experience Sampling (DES), a phenomenological protocol developed to avoid problems with naïve introspection by appealing to random sampling of introspective data, diary keeping, and weekly interviews (Hurlburt, 2011). Using DES, Hurlburt

et al. (2013) report that although inner speech often seems to be auditory-phonological, at times inner speech seems to be "missing all of its words" (p. 1483). Hurlburt and Heavey (2018) thus

> call into question [an] aspect of inner speech that Vigliocco and Hartsuiker take for granted: *Of course* "it is phonetic in nature." DES shows that sometimes (when it is unworded) inner speaking is not at all phonetic. (p. 183)

The phenomenology of this so-called 'unworded' inner speech seems to be propositional, not word-like. A similar form of inner speech seems to be echoed in Vygotsky's classic work *Thought and Speech* when he says that "inner speech is to a large extent thinking in pure meanings" (1986, p. 249).

In addition to 'unworded' inner speech, at other times inner speech seems to involve words, but does not seem to be phonological, phonetic, or articulatory. Take Gauker (2018):

> One of my students reports that she never hears an inner voice. But when I ask her what she experiences when she plans out what she is going to say, she says she experiences words – just words (p. 59).

The phenomenology of the inner speech of Gauker's student seems to be confirmed by Wayne Wu's review of the phenomenological literature on inner speech (Wu, 2012). Wu goes so far as to say that inner speech is "often abstracted from an auditory format, namely without representation of audible properties" (p. 96). This form of inner speech seems to also be acknowledged by MacKay (1992), who writes that "many aspects of the acoustics of overt speech are normally absent from our awareness of self-produced internal speech" (p.128). For MacKay, this is phenomenological evidence that inner speech is neither phonetic nor articulatory. Thus, at times inner speech can seem to involve propositions without involving words, and at other times it can seem to involve words without possessing speech-based characteristics. Therefore, alongside the often-reported auditory and articulatory forms of inner speech, there are phenomenological reports of inner speech that are merely propositional or syntactic.

I will assume that these phenomenological differences are to be explained by differences in representational content. If we can use phenomenology as a guide to content, it follows that

inner speech can take up a variety of contents corresponding to the levels of the speech processing hierarchy. Based on the above reports, there are forms of inner speech whose contents are propositional, syntactic, phonological, auditory, or motor. Though this phenomenological diversity is pluralistic in spirit, it stands in tension with standard pluralism. According to standard pluralism, the contents of inner speech come from the levels of the speech processing hierarchy that are distinctively speech-based: phonological, phonetic, and motor. The problem is that standard pluralism does not aver the existence of inner speech that might be merely propositional or merely syntactic. Standard pluralism therefore fails to capture the variety of forms of inner speech.

This phenomenological diversity also poses a more fundamental threat. Even if we develop an expanded version of standard pluralism that agrees that there are propositional and syntactic forms of inner speech, such a view would not be able to distinguish these forms of inner speech from other mental states. Propositional contents are truth-evaluable contents, and so representations of them are akin to believing, entertaining, or just plain thinking a propositional content. But these would not amount to cases of inner speech. Syntactic contents are contents that specify the grammatical relations that hold among words in a sentence. But, to represent the grammatical relations of words in a sentence is not to engage in inner speech. After all, a representation of the grammatical relations of words in a sentence is neutral as to whether those words are graphically or auditorily specified. Thus, there is nothing inherent in representations of either propositional or syntactic contents that mark them off as cases of inner speech. Therefore, appealing to the speech processing hierarchy on its own fails to explain why propositional and syntactic forms of inner speech are forms of inner speech rather than some other mental state. This suggests that if we are to allow for propositional and syntactic forms of inner speech, we must not

only reject standard pluralism, but, more fundamentally, we must look for the contents of inner speech outside of the speech processing hierarchy.[9]

2.8 Toward Vocal Content

The debate about the content of inner speech has reached a dead-end due to its narrow focus on the resources made available by the speech processing hierarchy. To resolve this situation, I suggest that we broaden our view of the processes that are implicated in inner speech. The key is to realize that the speech processing hierarchy is just one component of a broader range of processing: *voice processing*. Pascal Belin and colleagues have been at the forefront of delineating the structure of voice processing in the brain. As Belin et al. (2004) note, "the voice not only contains speech information, it can also be viewed as an 'auditory face', that allows us to recognize individuals and emotional states" (p. 129). These three different facets of vocal signals – speech, affect, and identity – are captured by Belin and colleagues' model of voice processing. According to this model of voice processing, a vocal signal is independently processed for speech, affect, and identity (see Figure 2).

[9] One might attempt to resist this conclusion by claiming that phenomenological reports about inner speech are unreliable. Such an objector might go on to claim that we have empirical evidence for motoric, auditory, and phonemic forms of inner speech, but not syntactic and propositional forms. I agree that we are lacking empirical evidence for the latter two forms of inner speech. However, the lack of such evidence is not meaningful since researchers have failed to develop the paradigms needed to test for syntactic and propositional inner speech in the first place. Moreover, the reason that we have developed paradigms to test for motor, auditory, and phonemic inner speech, but not syntactic and propositional inner speech, seems to be that researchers have assumed that inner speech is essentially a speech phenomenon. However, this background presupposition is, contrary to the objection, grounded in the phenomenology of inner speech – it *feels* like speech, so that's how it's been studied.

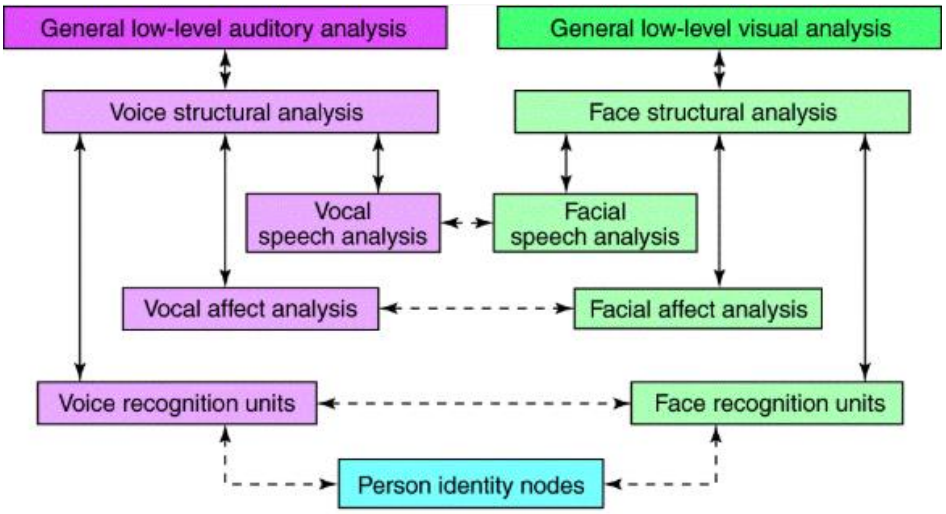

Figure 2: Models of face and voice recognition. Reproduced from Belin et al. (2004).

In the first stage, the vocal signal is processed for auditory properties that are on a par with non-vocal, environmental sounds. During the second stage of processing – structural encoding – a representation is generated of a human voice. It is hypothesized that at this stage a representation is produced of the distance in auditory space between the detected human voice and a normal or average human voice (Belin, 2019). Structural encoding then leads to three independent processes: speech processing, affect processing, and identity processing. Speech processing encompasses the speech processing hierarchy already presented in Figure 1. Affect processing involves processing the vocal signal for information relating to affect and other socially relevant properties, including weight, gender, competence, personality, and the like. Identity processing extracts from the vocal signal the identity of the speaker, e.g., that it is *Sam's voice*. Finally, on the basis of the representation of vocal identity, a representation of the amodal identity of the speaker is generated. The main takeaway from the model is that speech processing shows up as only a single component within a complex range of processes centered on voice.

I want to suggest that vocal *processing* is mirrored in vocal *content*. The assumption here is that psychological processes fix the content of the states that result from them. An implication

is that, in a standard case of speech perception, we not only represent the motoric, phonetic, phonological, syntactic, and semantic properties of someone's utterance – those that make up the speech processing hierarchy – but we also represent *how* the utterance was made (angrily, sadly, etc.) and *whose voice* made it (Sam's voice). Although it remains true that in *speech* perception we strictly speaking represent only those contents contained in the speech processing hierarchy, once speech perception is viewed as only one aspect of an encompassing perceptual activity centered on voice, the structure of the content that results from that activity becomes more complex.

I will characterize 'vocal content' as follows:

(1) Communicate (Voice, Information)

The idea behind (1) is that when we hear a voice, we do not simply represent speech information – propositions, words, etc. – but also a voice communicating such information. The view of inner speech I will be presenting in Section 2.9 claims that the content of inner speech is vocal content. A crucial aspect of the view is that what unites the plurality of cases of inner speech – from propositional inner speech to lower level, motoric inner speech – is that their contents include *voices*. In the rest of this Section, I explain each component of (1), starting with 'Communicate'.

There is some difficulty in determining the exact relation that voice bears to information. One option is to say that voice *carries* or *expresses* information. However, as it is typically used, *express* is a relation between states of information, not between voices and states of information. Thus, speech sounds express a proposition, but voices do not. Appealing to the *carriage* of information fares no better: although tree rings *carry* information about age, it is not clear that voices similarly carry conventional information, e.g., information about words. These problems

might motivate one to reject the idea that there is a relation between voice and information. On this alternative, vocal processing at best produces conjunctive contents: a voice and some information. However, there seems to be some sense in which voice *subserves* information in those cases in which we represent both.

I favor a strategy on which voice is thought of as a *channel* of information.[10] A channel communicates a signal from a source to a receiver; it is the medium through which one communicates an informational signal. When Bob says the words, *the meeting is cancelled*, Bob is the source, the words are the signal, and the channel over which these words are communicated is Bob's voice. As a channel, voice communicates a number of different kinds of information at once: auditory information, phonemic information, syntactic information, and propositional information. For those of us who speak, voice is very often the preferred channel over which complex informational signals are to be communicated. Once we think of voice as a channel – and this is the important point – then the most natural relation that voice bears to information is that of *communication*. Voice communicates information. In whatever way the relation of communication is to be understood, as we shall see in Section 2.9, the unity of inner speech stems from representing the channel through which communication happens – *voice*.

Turning to the 'Information' component of vocal content, voice can communicate any of the contents of the speech processing hierarchy – those contents relevant to language – as well as contents relating to affect and identity. My hushed voice might communicate secrecy or fear as it also communicates the words *don't tell anyone*, all the while communicating my identity to the receiver. In normal cases, there is a dependency between the types of information communicated

[10] Although the 'channel' terminology is most closely associated with Shannon (1948) and Dretske (1981), I do not intend to use the term in the formalized sense intended by these authors.

by voice. Thus, voices communicate a proposition typically by communicating words, communicate words typically by communicating speech sounds, and so on. Thus, in normal contexts, there is no such thing as a voice *directly communicating* a proposition. However, as we shall see, inner speech violates this downward dependency: in inner speech we can represent voices directly communicating propositions and words.

Central to vocal content is the 'Voice' slot. I noted that voices are channels of communication. But there are multiple ways a channel can be characterized. Voices may be characterized in terms of either *auditory* or *amodal* properties.

Voice processing involves both telling different voices apart and 'telling together' variations on the same voice. On the one hand, representing a voice as belonging to Bob requires that I distinguish it from voices belonging to other subjects, while on the other hand, representing a voice as belonging to Bob requires also representing variations of his voice as belonging to Bob's voice. Both of these representations of voice have been shown to rely on representations of a multidimensional auditory space. Latinus et al. (2013) argue that in distinguishing a particular person's voice from others, we represent the distance in auditory space of that voice from a prototype voice whose auditory value is reached by averaging across experienced voices. The greater the distance from the prototype voice, the more distinctive the voice is perceived to be. In lumping together variations of a single voice, Lavan (2019) argues that we represent a prototype version of that voice whose auditory value is reached by averaging across experienced samples of the voice. Subsequent samples whose auditory properties are nearby those of the prototype are taken to be produced by the voice whose prototype it is. Therefore, in coming to represent voices, we represent points and regions within a multidimensional auditory space. I will call this characterization of voice an 'auditory characterization of voice'.

But there is also evidence of representations of voice that are not specified in terms of auditory properties, but are rather modality-independent. Awwad Shiekh Hasan et al. (2016) found that fMRI activity generated in response to hearing voices could also be used to correctly classify the faces of corresponding individuals. This suggests that higher stages of voice processing "become increasingly abstracted from input modality" (p. 5). In replicating this finding, Tsantani et al. (2019) conclude that right posterior superior temporal sulcus (rpSTS) was "able to discriminate familiar identities based on modality-general information in faces and voices" (p. 201). In addition, these authors also found that rpSTS represented the identity of a single person across both videos and recordings of the person. This sort of evidence suggests that, at some stage, representations of voice are not representations of points within an auditory space. Rather, we are representing an integration of face and voice information or else information that is "modality-general". (Of course, it remains an empirical question exactly what amodal properties are being tracked in representing voice under such a specification.) I will call this characterization of voice 'an amodal characterization of voice'.

In sum, vocal content is more complex than the contents of the speech processing hierarchy. Where the latter include propositions, syntactic structure, speech sounds, etc., the former involve a relation – that of communication – holding between voice and information. Moreover, this structural complexity of vocal content is compounded, as I have shown, by a plurality of substitution instances of 'Voice' and 'Information'. Voices can be characterized either auditorily or amodally, while the information communicated by voices includes that of speech, affect, and identity.

2.9 Vocal Content in Inner Speech

We are now in a position to reassess the content of inner speech in light of my proposed account of vocal content. To this end, I first present evidence that inner speech not only implicates speech processing, but also vocal processing more generally.

Belin and colleagues have shown evidence for the existence of so-called temporal voice areas (TVA) in the temporal lobe: areas that are sensitive to voice and vocal identity in a way that is assumed to be analogous to the selectivity of the fusiform face area (FFA) for faces and facial identity (e.g., Belin et al. 2000). Building on this research, Yao et al. (2011) used fMRI in a silent reading task involving either indirect or direct reports of speech (e.g., "Sam said that the car is red" versus "Sam said, 'The car is red'"). The authors found that TVA activation was present in both conditions, but that activation increased in the direct speech condition. Moreover, Perrone-Bertolotti et al. (2012) found similar activations in TVA during silent reading of a story in patients being prepped for brain surgery. In addition to this imaging data, Kurby et al. (2009) have presented behavioral evidence of vocal processing during inner speech. The authors had subjects first listen to an enactment of a script and then read the script silently to themselves. Auditory probes were presented during parts of the silent reading task, and it was found that subjects reacted faster to a probe if it matched the voice of the character being read than if it did not match. This priming effect suggests that subjects' inner speech is in the voice of particular characters as they silently read texts (see also Alexander and Nygaard, 2008). The takeaway from this evidence is that inner speech implicates not just speech processing, but voice processing, more broadly.

We are now in a position to use vocal contents to understand the structure of inner speech episodes. According to my proposal, *vocalism*, the content of inner speech is vocal content. The basic strategy in what follows is to derive the forms of inner speech observed in Section 2.7 by

replacing (1) with a specification of voice and a specification of the content of the speech processing hierarchy (see Table 1).

Propositional inner speech is the 'unworded' inner speech we noted in Hurlburt et al. (2013). According to vocalism, during propositional inner speech, one represents an amodal specification of a voice communicating a proposition. For example, I may represent my voice communicating THE MEETING IS CANCELLED.

Syntactic inner speech is the sort of inner speech reported by Gauker's student who "experiences words – just words". During syntactic inner speech one represents an amodal specification of a voice communicating words in an order indicative of their syntactic roles. For example, I may represent my voice communicate *the meeting is cancelled*.

Phonemic inner speech is the sort of inner speech found in Oppenheim and Dell (2013): "there is just one level for inner speech – a phonological level" (p. 1157). During phonemic inner speech one represents an amodal specification of a voice communicating phonemes. For example, I may represent my voice communicate /the meeting is cancelled/.

Table 1: The structure of different forms of inner speech

	Voice	**Information**
Propositional inner speech	Amodal specification of voice	Propositional content
Syntactic inner speech	Amodal specification of voice	Lemmas and syntactic properties
Phonemic inner speech	Amodal specification of voice	Phonemes
Phonetic inner speech	Auditory specification of voice	Phones
Articulatory inner speech	Motor specification of voice	Vocal tract movements

Phonetic inner speech is the sort of "auditory-verbal" inner speech that Langland-Hassan claimed was definitive of inner speech. According to vocalism, during phonetic inner speech one represents an auditory specification of a voice communicating speech sounds. For example, I may represent my own voice in an auditory register communicate [the meeting is cancelled].

Articulatory inner speech is the sort of inner speech exhibited in the minute movements of the vocal tract observed by McGuigan and Dollins (1989). Articulatory inner speech is marked by a representation of a motoric specification of voice communicating vocal tract movements.[11] For example, I may represent my own voice in a motor register communicate {…}, where '…'

[11] It might be objected that it is incoherent to say that vocal tract movements are *communicated*. However, there is evidence that vocal tract movements are communicated for the sake of speech perception. Although it may not be the case that vocal tract movements are the primary unit of speech, as posited by the motor theory of speech perception (see Liberman and Mattingly (1985)), there is ample evidence that vocal tract movements assist in speech perception. The McGurk effect shows that speech sounds are interpreted differently based on visual perception of motor gestures (McGurk and McDonald, 1976). Moreover, there is also evidence that hearing speech sounds activates correlative motor regions in the absence of visual perception of the motor gestures (see Pulvermuller et al., 2006).

indicates the complex array of motor gestures need for the voice to communicate [the meeting is cancelled].

There are two important points to make about this lineup of forms of inner speech. First, notice that propositional, syntactic, and phonemic inner speech each implicate the same substitution instance for 'Voice'. This is because none of these three forms of inner speech is auditory in character. The best option in accounting for these forms of inner speech is in terms of an amodal specification of voice. Second, notice that I have introduced the notion of a *motor specification* of voice in order to account for articulatory inner speech. Although this specification of voice was not discussed in Section 2.8, it can be thought of as a specification of voice in terms of those motor features that would, if executed, give rise to an auditory characterization of voice.

What makes each member of the lineup (including propositional and syntactic inner speech) a case of inner speech is that it possesses vocal content – Communicate (Voice, Information). Thus, the same feature that makes phonetic inner speech a case of inner speech also makes propositional inner speech a case of inner speech. It is not that phonetic inner speech has more of a claim on being inner speech because it also possesses phonetic contents. That would be to assume just the picture that I am resisting in this Chapter: that the contents of the speech processing hierarchy are somehow definitive of inner speech. Instead, once we adopt vocalism, a variety of forms of inner speech are on a par in virtue of exemplifying vocal content. Of the positions we have considered, vocalism is best able to account for the plurality of forms of inner speech.

2.10 Escaping a Dilemma: Unity and Plurality in Inner Speech

An overarching theme of this Chapter concerns reconciling the plurality of forms of inner speech with its underlying unity.

On the one hand, inner speech is highly variegated. Although most discussion of inner speech has concerned the auditory variety, there are forms that extend across the speech processing hierarchy, including propositional and syntactic inner speech. The variety of forms of inner speech poses a problem, however, since it can seem that they do not share anything in common. As a result, accounts of inner speech are forced to adopt a disjunctive view: on the one hand, there are standard forms of inner speech with speech-specific contents, and, on the other hand, there are forms of inner speech with non-speech-specific contents, e.g., propositional and syntactic inner speech. But there seems to be nothing that unites both forms as instances of inner speech. By emphasizing the plurality of inner speech, we fail to provide a general account of the mental state.

On the other hand, in order to secure a general, unified account of inner speech, it seems we need to exclude those forms of inner speech – propositional and syntactic – that seem to be the cause of trouble. This is the route taken by participants in in the current debate over the content of inner speech. They adopt a picture of the content of inner speech as speech-specific: concretism characterizes the content as phonetic and/or articulatory, abstractionism characterizes the content as phonemic, while standard pluralism marks the disjunction of these views. But, once this picture is adopted, the existence of abstract forms of inner speech is no longer recognized. By emphasizing the unity of inner speech, we fail to account for its extension.

This dilemma results from adopting the speech processing hierarchy as a framework for thinking about the content of inner speech (Section 2.2). Endorsing that framework forces us to choose between the plurality and unity of inner speech: emphasizing the former puts a general

account of inner speech out of grasp, while emphasizing the latter excludes non-canonical forms of inner speech. I have suggested that escaping this dilemma requires us to widen our view to take into account voice processing, of which the speech processing hierarchy is but one component (Section 2.8). Taking this wide-angle view on inner speech shows that we can embrace both the radical plurality and the overarching structural unity of inner speech (Section 2.9).

3.0 Imagination and Prediction: The Test Case of Inner Speech

3.1 Introduction

Predictive processing is an ambitious theoretical framework that attempts to analyze the mind in terms of predictive processes alone. According to the framework, different psychological capacities – perceiving, acting, believing, desiring, among others – are analyzed as variations on a single kind of predictive process: the minimization of prediction error. These capacities emerge as different ways of reconciling predictions about the world and data coming from the world so as to minimize the difference between them. This framework has been used to model a wide range of mental states, including binocular rivalry (Hohwy, 2008), action (Hohwy, 2013; Clark, 2015), attention (Feldman and Friston, 2010), and a number of mental disorders, including hallucination and delusion (Sterzer, 2018). As the scope of the framework has grown, supporters have touted predictive processing as a "grand theory" (Dehaene, 2008, p. 30): a framework "able to unify…diverse aspects of our mental lives under one principle" (Hohwy, 2013, p. 5), "offer[ing] the best clue yet to shape a unified science of mind and action" (Clark, 2013, p. 181).

Though the predictive processing framework has universal ambitions, there has been little attention paid to predictive models of imagination and imagery. Discussion of imagery is scant because supporters assume that an account of imagery is "straightforward" on a predictive processing framework, part of a "package deal" with the framework itself (Clark, 2015). According to predictive processing theorists, states of imagery are identical to predictions. What makes a prediction fit to serve as a state of imagery, according to these theorists, is that predictions

just are images. The idea is present in the work of Andy Clark, a vocal proponent of predictive processing:

> Animals capable of forming rich, world-revealing percepts are, if the predictive processing story is on track, animals that understand their worlds and that are poised to imagine them too. The argument for this is straightforward. An important feature of the internal models that power such approaches is that they are generative in nature. That is to say, the knowledge (model) encoded at an upper layer must be such as to render activity in that layer capable of predicting the response profile at the layer below. That means that the model at layer N + 1 becomes capable, when operating within the context of the larger system, *of generating the sensory data (i.e., the input as it would there be represented) at layer N (the layer below) for itself*. Since this story applies all the way down to the layers that are attempting to predict activity in early processing areas, that means that such systems are fully capable of generating 'virtual' versions of the sensory data for themselves. (italics added) (2015, p. 93)

The predictive processing framework posits a hierarchy of predictions. Conflict between a prediction and incoming data results in prediction error, which is passed up the hierarchy leading to revision of the prediction the next level up. This requires that each level predicts 'the response profile' of the level just below it. In this context, Clark's basic idea is that a prediction predicts a lower level of the hierarchy to the extent that it generates "for itself" the sensory data that would otherwise be present in that lower level.[12] Predictions thus generate "'virtual'" sensory data. According to Clark, predictions are simulations of data, and as a result, imagery is "'part and parcel' of learning to perceive" (Clark, 2013, p. 198). A similar line to Clark's is advocated by many other proponents of predictive processing:

> An important implication of [predictive processing] is the existence of a strong continuity between perception and imagination or imagery…to be able to perceive an object requires a generative model capable of autonomously creating, in a top-down fashion, fictive (i.e., surrogate) sensory signals that could originate from that object… (Seth, 2014, p. 101)

> perception and imagination are the psychological results of the brain endogenously generating its own sensory inputs top-down (Kirchhoff, 2018, p. 752)

> What we have so far is an internal model that generates a hypothesis – we might call it a fantasy – about the expected sensory input….A particular fantasy might do a fine job at matching the incoming sensory input, and thus should determine perception. (Hohwy, 2013, p. 54)

[12] 'Sensory data' does not have its typical connotations within the predictive processing framework. Within predictive processing, 'sensory data' is simply the data in light of which predictions are updated (cf. Clark's unpacking above). As such, there is no commitment to the data having a phenomenological character. Predictive processing thereby purports to encompass both 'sensory' and 'non-sensory' forms of imagination and simulation.

The basic idea is that predictions are themselves images: 'virtual', 'surrogate', 'endogenous', or 'fantasy' versions of sensory data. Theorists are thus united in the idea that a model of imagery is supposed to be "straightforward" on the predictive processing framework.

This Chapter will challenge the idea that imagery should be identified with a prediction. Though my challenge will have bearing on imagery in general, my focus will be on inner speech in particular. While inner speech involves a number of linguistic features that are not associated with imagery, including propositional and syntactic contents, my focus will be on the *auditory verbal imagery* that is most closely associated with inner speech. I will therefore use 'inner speech' to refer only to auditory verbal imagery.[13] Inner speech is an apt test case to assess predictive models of imagery because predictive models of inner speech have been more thoroughly developed than predictive models of other kinds of imagery (see Swiney, 2018). Predictive processing models of inner speech were inspired by earlier predictive models that stem from standard models of motor control. For this reason, in addition to predictive processing models of inner speech, I will be focusing also on these earlier, more standard predictive models of inner speech.

In Section 3.2, I present two models of inner speech – the 'Standard Model' and 'Radical Model' (Section 3.2). The Standard Model identifies inner speech with a prediction of the sensory consequences of executing speech motor commands. The Radical Model is similar but uses only the minimal resources of the predictive processing framework. I will challenge the Standard and Radical Models of inner speech on both empirical and theoretical grounds. In Section 3.3, I explain

[13] From the point of view of Chapter 2, 'inner speech', as it is used in the current Chapter, denotes only a *component* of inner speech. That is, from the perspective of Chapter 2, I am focusing only on the 'Information' slot within the form Communicate (Voice, Information), and even then, only on the auditory instantiation of 'Information'. This terminological shift reflects deference to the way theorists use the term 'inner speech' within the literature on predictive models.

a swath of studies that are taken to support both models. I argue that these studies not only fall short of supporting these models, but also that existing empirical evidence converges to tell against predictive models of imagery more generally. Section 3.4 takes up predictive processing models of imagery by focusing on the Radical Model of inner speech. I show that the predictive processing framework is unable to account for imagery without undermining its account of how subjects go about perceiving and acting in the world. Having criticized predictive models of imagery, in Section 3.5 I sketch an alternative, non-predictive model of inner speech, thereby illustrating the viability of non-predictive models of imagery more generally. I close in Section 3.6 by marking out a significant implication of the argument: if imagery is not to be identified with a prediction, then we need to rethink the nature of the predictions on which predictive models are based. Contrary to the claims of predictive modelers, predictions are not images – 'virtual' versions of sensory data – but are at most psychological kinds that represent statistical properties of events.

3.2 Two Models of Inner Speech

In this Section, I first present two theoretical frameworks, the standard framework of motor control and the predictive processing framework. I then show how these frameworks give rise to the Standard and Radical Models of inner speech, respectively.

The standard framework of motor control implicates a series of comparators that compare intentions, predictions, and sensory feedback in an effort to control and guide action. According to this framework, limb movement begins with an intention to move a limb, which is transformed by an 'inverse model' into a set of motor commands, which are then sent to the limbs for execution (see left-hand side of Figure 3). Prior to the execution of the movement, however, a copy of the

original motor command is generated – an 'efference copy' – which is transformed by a forward model into a prediction of the sensory consequences of executing the limb movement (right-hand side of Figure 3).[14] At Comparator 2 (bottom of Figure 3), this prediction is compared to the actual sensory feedback generated from executing the limb movement. If the prediction and the actual sensory feedback match, then the limb movement is considered successful. However, if there is a mismatch at Comparator 2, the difference between the predicted and actual sensory feedback is fed back into the inverse model, where motor commands are fine-tuned and the limb movement is re-executed.[15]

[14] Sometimes authors use the term 'forward model' to refer to what I will call a 'prediction of sensory consequence'. However, this usage is not standard. As I will use the term here, a forward model is a *process* that transforms motor representations into sensory representations and the *output* of this process is a prediction of sensory consequence.

[15] For the purposes of this Chapter, Comparators 1 and 3 will not be of primary significance. Comparator 1 computes the difference between the predicted sensory state and an intended sensory state. This comparator allows for the detection of errors prior to the execution of an action. Comparator 3 computes the difference between actual sensory feedback and the intended sensory state. This comparator allows for the construction of novel motor commands on the basis of sensory feedback.

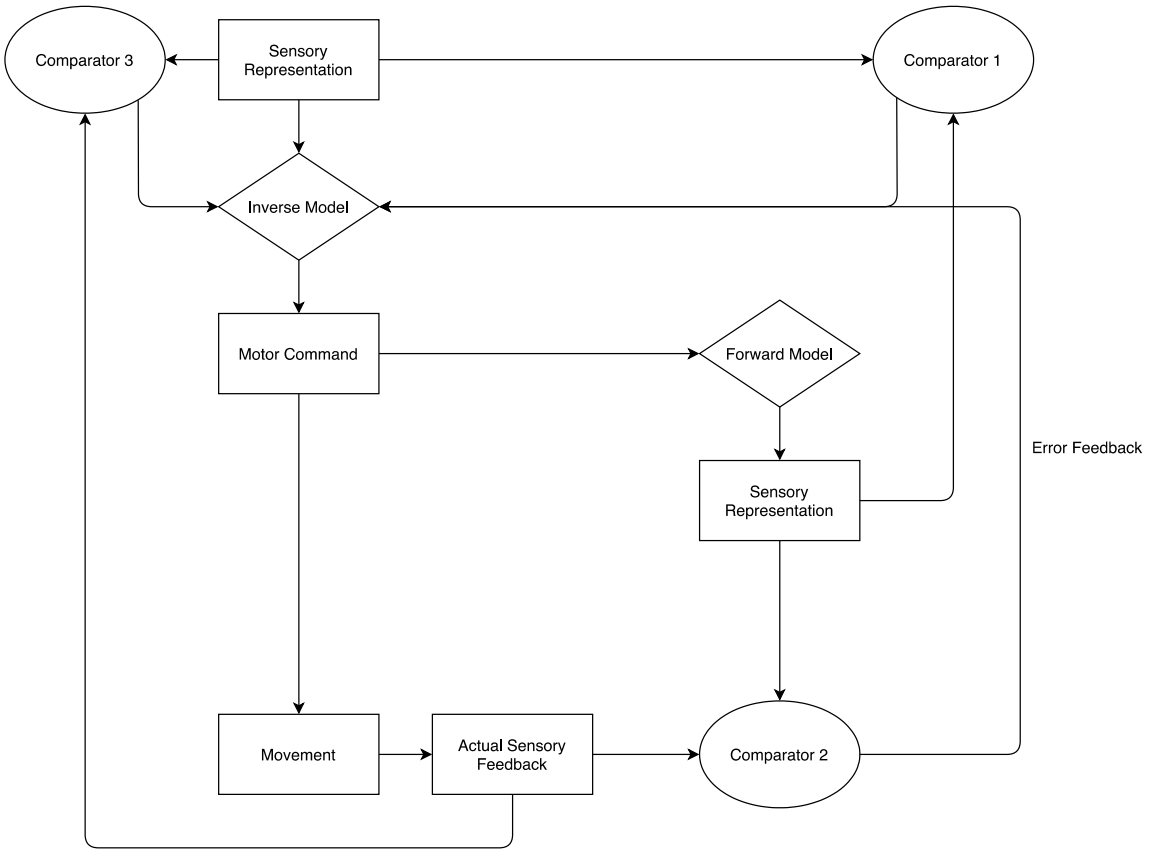

Figure 3: Standard comparator model of motor control (see, e.g., Frith, et al., 2000; Wolpert and Flanagan, 2001; Swiney and Sousa, 2014; Loevenbruck, et al.., 2018))

The Standard Model of inner speech is inspired by the standard model of motor control. It identifies inner speech with the output of a forward model, namely, with a prediction of the sensory consequences of executing speech motor commands (see Figure 4). According to the Standard Model, an intention to produce speech sounds is transformed into a set of speech motor commands, but unlike normal speech production, the commands are suppressed at the vocal tract (gray). Despite this suppression, an efference copy of the motor commands is still generated, which is transformed by a forward model into a prediction of the speech sounds that would have been produced by those motor commands. This prediction of speech sounds in the absence of actual feedback is experienced as inner speech (green) (e.g., Swiney and Sousa (2014), Carruthers (2018), and Loevenbruck et al. (2018)).

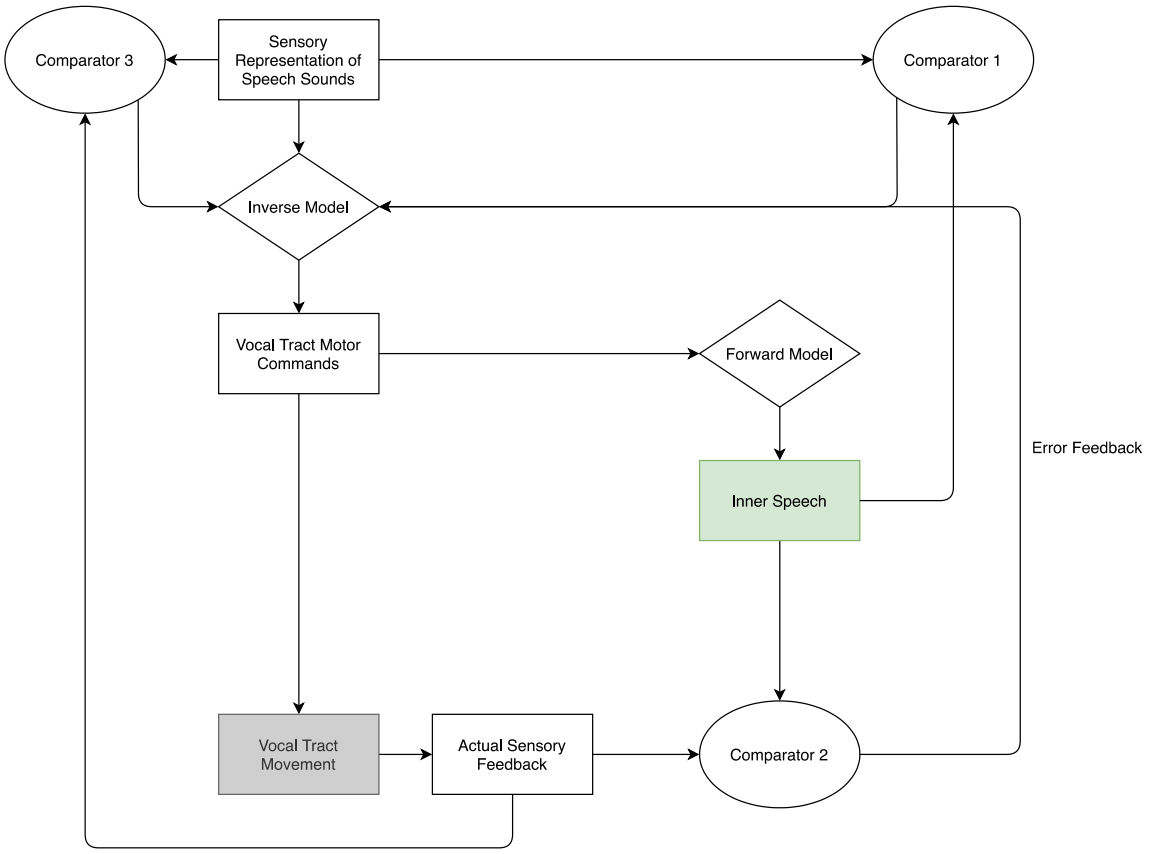

Figure 4: The Standard Model of inner speech (e.g., Pickering and Garrod (2013); Swiney and Sousa (2014); Carruthers (20018); and Loevenbruck et al. (2018))

Although the Standard Model remains the most popular model of inner speech, several influential theorists have been inspired to model inner speech in accord with the predictive processing framework (Wilkinson and Fernyhough, 2017). Where predictive processes feature as only one type of process alongside a number of others in the standard motor control framework, the predictive processing framework seeks to understand motor control, and the mind more generally, in terms of predictive processes alone.

There are three basic features that make up the predictive processing framework: *prediction error minimization*, *prediction error*, and *precision*. First, according to predictive processing, the mind has the sole aim of *minimizing prediction error*. Though the standard framework also

implicates the minimization of prediction error, it is only one process among a number of others. Second, *prediction error* is determined by the difference between prediction and incoming data. Prediction error is also implicated in the standard framework (see Comparator 2). However, unlike the standard motor control model, predictive processing emphasizes a statistical gloss on the notion. Prediction and data are represented by probability distributions, while prediction error is computed as the difference between the means of the distributions. The mind minimizes prediction error by updating either its predictions or data in a way that minimizes the difference between the means of the respective probability distributions.

The third component marks one of the novelties of predictive processing, and will become important in our assessment of its ability to account for imagery. The *precision* of a signal is represented by the probability distribution's degree of concentration over the mean. In effect, the precision of a signal determines the *weight* or *importance* of that signal in updating prediction or data. If the precision of a prediction is higher than the precision of the data, the update will be more influenced by the prediction than the data, and vice versa if the precision of the data is higher than the precision of the prediction. In effect, the precision of a signal represents how confident we are in that signal, and our updates are shaped by that confidence. Using these three components alone, the predictive processing framework attempts to model a diverse range of mental states and processes.

Perception and action, according to predictive processing, are simply ways in which prediction error is minimized. Given some prediction-data pair, one way of minimizing prediction error is by revising the prediction to fit the data. This form of prediction error minimization corresponds to a perceptual, or mind-to-world direction of fit. In contrast to perception, in action we minimize prediction error by changing the data to fit the prediction. This form of prediction

error minimization corresponds to a world-to-mind direction of fit. The world is acted upon so as to bring the world in line with one's prediction. For example, in the case of speech production, predictions of the presence of incoming speech sounds in one's own voice generate downstream predictions of proprioceptive sensations at one's vocal tract, both of which are then confirmed via movements of the vocal tract that bring about just those speech sounds and proprioceptive sensations. Action is not caused by independent motor commands, as on the standard motor control framework, but are rather elicited as a means of confirming prior predictions, thereby minimizing prediction error. In short, perception occurs when there is an update of prediction in light of data, whereas action occurs when there is an update of data in light of prediction.

But what determines whether a subject should perceive or act, on a predictive processing framework? According to most theorists, action is generated when the *precision* of the prediction is sufficiently high relative to that of the incoming sensory data, such that the prediction becomes nearly unrevisable (e.g., Hohwy, 2008). In effect, one is sufficiently confident in the prediction that one does not revise it, and one lacks sufficient confidence in the data that it becomes open to revision. The result is that one is forced to minimize prediction error by revising the data through action. Whether one engages in perception or action is therefore a function of one's relative confidence in the prediction and data. Whereas the standard framework of motor control views action as generated by motor commands and only guided by predictions, the predictive processing model views action as generated by the predictions themselves (see also Garrod and Clark (2014)).

Inspired by the predictive processing framework, Wilkinson (2014) and Wilkinson and Fernyhough (2017) adopt what I call the Radical Model of inner speech. Their account stems from Clark's predictive processing model of imagery:

> The proposal is that the brain, in order to simulate future unfoldings, must mute the weighting on select aspects of the proprioceptive prediction error signal. Suppose this is done while simultaneously entering a high level neural state whose rough-and-ready folk-psychological gloss might be something like "I reach for

> the cup." Motor action, on the PP [predictive processing] account, is entrained by proprioceptive expectations and cannot here ensue. But all the other intertwined elements in the generative model remain poised to act in the usual way. The result should be a "mental simulation" of the reach and hence an appreciation of its most likely consequences. (p. 2, 2013)

Imagery is achieved, according to Clark, by generating a high-precision prediction – *there are sensations of reaching for the cup*.[16] In the case of action, the high-precision prediction is compared to the data – *there are no sensations of reaching for the cup* – and the subsequent prediction error is then minimized by reaching for the cup. However, according to Clark, in the case of *simulating* reaching for the cup, I 'mute the weight' on the prediction error, such that the prediction error is not minimized by updating either the prediction or the data. In other words, by 'muting the weight' on the prediction error, the prediction error is deemed unimportant and thereby does not need to be minimized by updating either the prediction or the data. Nevertheless, given that the precision is high for the prediction *that there are sensations of reaching for a cup* and low for the incoming sensory signal *that there are no sensations of reaching for a cup*, I enjoy motor imagery of reaching for a cup. The crucial feature of Clark's account is that the precision on prediction error can be modulated, such that I can minimize prediction error without updating either the prediction or the data.

Wilkinson and Fernyhough (2017) adopt this style of account for inner speech:

> [imagery in inner speech]…is the prediction itself, or, more specifically, a decoupled hypothesis that entails a bunch of deliberately unfulfilled (but prediction-error minimized, through down-modulation [of the weighting/precision of prediction error]) predictions. (p. 296)

Analogous to Clark's account of mental simulation, inner speech, according to Wilkinson and Fernyhough, is identical to a high-precision prediction – *there are incoming speech sounds in my own voice*. In the case of speech production, this prediction is compared with the incoming sensory

[16] My presentation might suggest that the prediction is propositional. However, the prediction might also be a sensory state of imagery. Nothing will turn on that difference in what follows.

data – *there are no incoming speech sounds in my own voice* – and the subsequent prediction error is minimized via production of those speech sounds. However, in the case of inner speech the resulting prediction error is minimized via 'downmodulating' or decreasing the precision on the prediction error. Inner speech is supposed to be identical to the high-precision prediction.

We thus have two predictive models of inner speech – the Standard Model and the Radical Model. Both models claim that inner speech is identical to a prediction of sensory data. Both also claim that inner speech is achieved when these predictions are taken offline. However, the models differ in what they say about how predictions are taken offline. The Standard Model claims that predictions are taken offline by suppressing motor commands at the vocal tract. In contrast, the Radical Model claims that predictions are taken offline using predictive processes themselves, namely, by decreasing the precision on prediction error. This difference in how inner speech is achieved will become theoretically important in Section 3.4. But first I want to focus on whether the empirical predictions made by these models of inner speech hold water.

3.3 Empirical Problems with Predictive Models of Inner Speech

Having explained the Standard and Radical Models, in this Section I will challenge these predictive models of inner speech on empirical grounds. There has been much empirical research that has claimed to support predictive models of inner speech. In order to assess this evidence, we need to grasp the mechanisms underlying these models and the experimental predictions they make.

3.3.1 A Prediction

On both models, there is a mechanism that determines whether there is a match or mismatch between prediction and data. If prediction and data match, then the processing of the data is *attenuated* or *decreased* relative to baseline perception of that data. In contrast, if there is a mismatch between prediction and data, the processing of the data is *not* decreased relative to baseline perception of that data. In effect, in the case of a match, the incoming signal is canceled, and so marked as uninformative, whereas in the case of a mismatch the incoming signal is enhanced, and so is marked as informative (Pickering and Clark, 2014).

This sort of mechanism is nicely illustrated by the mormyrid fish, a nocturnal fish that uses electrolocation to navigate and track prey. The fish sends an electric pulse into its environment, a copy of which is used to predict the incoming electric pulse were no creatures present in its environment (Bell, 1989). This electrosensory prediction is subtracted from the total returning signal impinging on its electrosensory receptors. The portion of the returning signal that was predicted is not further processed, while the remaining portion is enhanced and made available for the tracking of prey. Perception of the fish's environment is thus determined by mismatches between prediction and incoming signal. The same mechanism is present in online speech production. The processing of speech feedback is enhanced during online speech production when its pitch or temporal onset are altered relative to when it is unaltered (Heinks-Maldonado et al., (2005) and (2006); Behroozmand and Larson, 2011). Similarly, the processing of auditory feedback is enhanced when a tape-recorded version is heard after having spoken than when the same auditory feedback is heard during speaking (Curio et al., 2000). These perturbation and playback paradigms suggest that ordinary speech production is subserved by a mechanism wherein

incoming data is attenuated if it matches a prediction, and is not attenuated, but enhanced, if it does not match a prediction.

For these reasons, the Standard and Radical Models of inner speech make the following prediction:

> *Attenuation*: The processing of an incoming sensory signal will be attenuated when it matches inner speech either relative to when there is a mismatch between inner speech and an incoming sensory signal or relative to when there is mere perception of an incoming sensory signal.

Attenuation has been investigated using electroencephalography (EEG). EEG is a neuroimaging technique designed to measure stereotyped electrical potentials produced by the brain. These stereotyped electrical potentials are often associated with particular cognitive functions: the P300 is associated with evaluation of a stimulus, the N400 is associated with meaningful stimuli, and, importantly for our purposes, the N100 is associated with attenuated processing (Schafer and Marcus, 1973).

3.3.2 Assessing Representative Studies

I present two EEG experiments that claim to confirm *Attenuation* before casting doubt on their support for predictive models of inner speech. I then show that a third, behavioral experiment that claims to validate *Attenuation* does not in fact do so. Although there are a handful of other studies that test *Attenuation*, the experiments that follow have been widely cited and are representative of experimental paradigms used in the studies I am not discussing (see Tian and Poeppel (2013), Ylinen et al. (2015), and Jack et al. (2019)).

Whitford et al. (2017) had subjects sit in front of a ticker tape display and fixate on a line moving across the display (Figure 5). The moment the moving fixation point intersected a fixed

target, subjects were instructed to produce a specified speech sound in inner speech, e.g., /ba/. At that precise time, a speech sound was also played over a headset. The speech sound either matched or did not match the content of the speech sound produced in inner speech. For example, subjects produced /ba/ and heard /ba/ or produced /ba/ and heard /bi/.

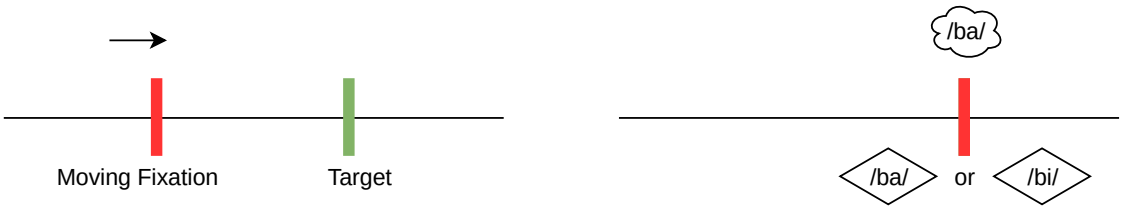

Figure 5: An illustration of the Whitford et al. (2017) ticker-tape paradigm. The cloud indicates inner speech and the parallelograms indicate the auditory probes.

The authors found that the speech sound played over the headset was attenuated (relative to baseline listening) only when there was a match between the speech sound produced in inner speech and the presented speech sound (i.e., /ba/-/ba/). The authors conclude that the result confirms *Attenuation*: "inner speech is associated with a…content-specific internal forward model" (p. 3).

The Whitford et al. experimental paradigm is an instance of a pioneering paradigm developed and used by Judith Ford in a number of papers (2001a, 2001b, and 2004). Ford et al. (2001a) tracked attenuation in two conditions. In the baseline listening condition, subjects fixated a target on a screen and were presented probes that alternated between a visually presented checkerboard, speech sounds, and broadband noise. In the inner speech condition, subjects were presented with these same probes while repeating statements in inner speech over the span of 30 seconds (e.g., "…That was stupid…That was stupid…").

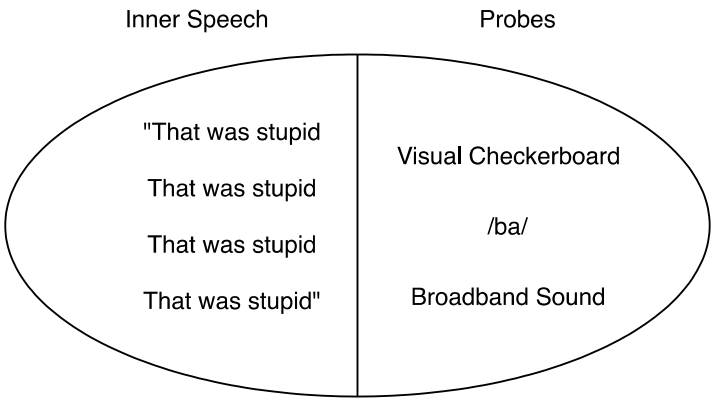

Figure 6: A comparator illustration of the inner speech condition in Ford et al. (2001a)

The authors found that the auditory probes (but not the visual checkerboard) were attenuated in the inner speech condition relative to the baseline listening condition. Ford et al. claim that their results confirm *Attenuation*: inner speech "may reflect the action of a corollary discharge [efference copy] from frontal brain structures preparing auditory cortex for speech" (p. 45).

Although both Whitford et al. and Ford et al. claim to support the association of inner speech with a prediction, their results are in tension with one another. Whitford et al. found attenuation of the incoming speech signal relative to listening only when there was a *match* in the speech sound produced in inner speech and the sound played over the headset. However, Ford et al. found attenuation relative to listening even when the sounds played over the headset – speech sounds and broadband sound – *did not match* the whole sentence produced in inner speech – "That was stupid". What gives? Since much of the evidence for predictive models of inner speech stems from such EEG studies, I will consider this puzzle in relative depth. In the following, I present a framework for making sense of the contradictory results. Once we apply this framework, we will find that neither result supports the idea that inner speech is associated with a prediction.

There are a number of ways the processing of an incoming signal can be attenuated.

One cause of attenuation is repetition. Repeated presentations of a stimulus have been shown to lead to attenuated processing of that stimulus. Relative to the first presentation, there is less activation during the second presentation of the same stimulus, an effect so robust that it is often used to localize functions to brain areas. For example, Epstein et al. (2008) show that when subjects are presented with two consecutive identical images of a campus scene, there is suppression of neural response in the parahippocampal place area for the second compared to the first stimulus. Moreover, attenuation is also observed when the same stimulus is repeatedly *imagined*. For example, Navarro and Janata (2010) show that subjects exhibit attenuation when repeatedly imagining non-vocalized musical notes. Similarly, Szpunar et al. (2014) show that when subjects repeatedly imagine a social scene, there is attenuation in select areas of the brain. These results show that repeatedly imagining a stimulus in some modality leads to a decrease in processing in that modality.

Another cause of attenuation is expectation. Controlling for repetition, researchers have shown that when a subject expects a stimulus, there is less neural activation during the presentation of that stimulus than when she does not expect it. For example, Summerfield (2008) had subjects placed in one of two conditions involving the repetition of the same face or alternation of different faces. In the first condition, the repeated faces were more probable than the alternating faces, whereas in the second condition, the alternating faces were more probable than the repeated faces. The authors found that neural suppression in the fusiform face area was a function of the *probability* of repetition or alternation in each condition and not a function of brute repetition.

Based on these sorts of results, a number of authors have concluded that that there are at least two distinct mechanisms underlying attenuation: repetition of a stimulus and expectation of a stimulus (e.g., Todorovic and Lange, 2012; Grotheer and Kovacs, 2015, 2016). Using this

insight, we can make sense of the apparent contradiction between the Whitford et al. and Ford et al. results, but in a way that fails to support predictive models of inner speech.

Recall that Ford et al. had their subjects *repeat* a sentence in inner speech for 30 seconds. As we have seen, however, the mere repetition of a sentence in inner speech is sufficient to generate attenuation in the auditory modality. As a result, we should expect that the processing of an auditory stimulus will be attenuated relative to baseline listening when the auditory modality is already being taxed by repeated inner speech. Hence, the attenuation observed by Ford et al. is, I submit, not the result of a comparison between prediction and data, but simply the effect of presenting a stimulus to a modality already taxed by repeated imagery (repetition priming). The Ford et al. study therefore does not provide evidence for predictive models of inner speech.

Whitford et al.'s attenuation results fare no better. Notice that subjects produced speech sounds in inner speech that were each composed of two speech sounds: /ba/ is composed from /b/ and /a/, while /bi/ is composed from /b/ and /i/. The auditory information contained in /ba/ and /bi/ are respectively determined, in part, by the auditory information contained in /b/ and /a/ and /b/ and /i/, respectively. It is because this dependency holds that /ba/ sounds more similar to /bi/ than /fi/ (/ba/ and /bi/ share the speech sound /b/). The dependency therefore generates a prediction: if inner speech is identical to a prediction, then we should expect more attenuation when one hears /bi/ and produces /ba/ in inner speech than (say) when one hears /fi/ and produces /ba/.[17] This is because the former pair share more auditory features than the latter pair. However, this prediction does not seem to hold. Although Whitford et al. showed that the /ba/-/ba/ match produced attenuation relative to the /ba/-/bi/ mismatch, they also showed that there is no significant

[17] This prediction of graded or continuous attenuation is also supported by evidence that shows that the less of a match there is between auditory stimulus and prediction, the less attenuation there is of the auditory stimulus (see Niziolek, et al., 2013).

difference between the level of attenuation produced by the /ba/-/bi/ mismatch and the level of attenuation produced by simply listening to /bi/, despite the fact that hearing /ba/ while producing /bi/ in inner speech involves more overlap in auditory features than simply hearing /bi/ without inner speech. This suggests that inner speech is not identical to a prediction.

If the Whitford et al. results do not show a pattern of attenuation reflective of inner speech being a prediction, then how do we account for their findings? Recall that subjects are cued to produce either /ba/ or /bi/ *before* the line starts to move across the ticker tape. For a given cue there is a 0.5 chance that the cued or non-cued speech sound (/ba/ or /bi/) will be presented. The objective probability is thus (/ba/, .5) and (/bi/, .5). Within this unpredictable environment, I suggest that the mere presence of the cue, for example, /ba/, produces a bias such that subjects' subjective probability becomes, say, (/ba/, .7) and (/bi/, .3). In the context of this subjective probability, when it turns out that /ba/ is presented over the headset, it is attenuated because it had been taken to have a higher chance of occurring. There is no attenuation in the case of a partial match, I suggest, since subjects are expecting /ba/ as a categorical event type and so the /ba/-/bi/ mismatch is treated as a complete mismatch. In short, I suggest that we are observing attenuation via expectation (as in Summerfield (2008) above).[18]

Neither the Whitford et al. nor Ford et al. study support predictive models of inner speech. Although there are other EEG studies that attempt to support *Attenuation*, their paradigms are

[18] Despite the weaknesses of the Ford et al. and Whitford et al. studies, there are ways of testing whether inner speech has predictive effects. Using the ticker-tape paradigm described by Whitford et al., I suggest that we can test whether inner speech is associated with a prediction by making mismatching speech sounds predictive of one another, while making matching speech sounds non-predictive of one another. For example, at the beginning of the ticker the subject would be cued to produce /bi/ (at time t). But, unlike the Whitford et al. paradigm, the cue to produce /bi/ is *positively correlated* with the auditory presentation of /ba/ (at t). In this context, if we observed attenuation when the subject produced /bi/ and was presented with /ba/ (at t), then this would mean that what drives attenuation is the amodal expectation that /ba/ will occur (at t).

similar to those used in these studies. In general, the results of these studies can be explained away in the way I have illustrated above.

There has also been an attempt to confirm *Attenuation* outside of the EEG paradigm. Scott et al. (2013) presented subjects with an ambiguous or hybrid speech sound between /ava/ and /aba/ while subjects concurrently produced /afa/ or /apa/ in inner speech.

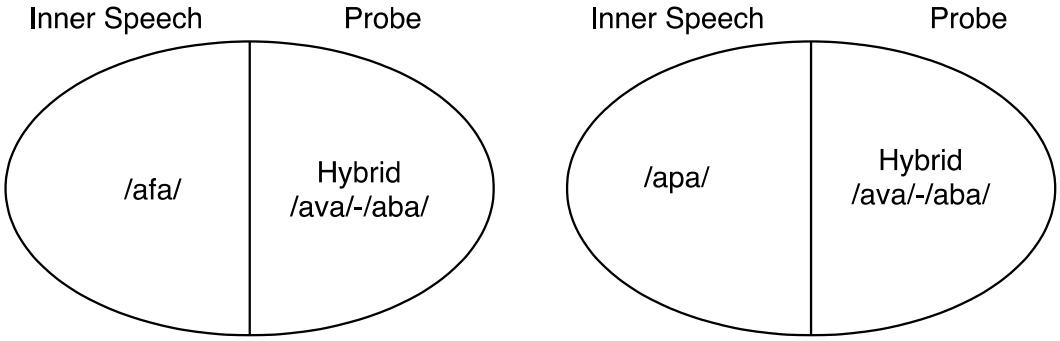

Figure 7: A comparator illustration of the inner speech conditions in Scott et al. (2013)

The authors found that subjects tended to report perceiving the hybrid speech sound as /ava/ if they produced /afa/ and as /aba/ if they produced /apa/. Their explanation is that inner speech 'captures' the hybrid speech sound: the hybrid sound is heard as /ava/ or /aba/ depending on the character of the middle consonant produced in inner speech. On the one hand, since /p/ shares more articulatory/acoustic properties with /b/ than with /f/ (/p/ and /b/ are bilabial stops), it generates a perception of /aba/, on the other hand, since /f/ shares more articulatory/acoustic properties with /v/ than with /b/ (/f/ and /v/ are labiodental fricatives), it generates a perception of /ava/. Scott et al.'s basic idea is that inner speech serves as a prediction that pulls the hybrid in its direction, thereby generating percepts that are closer to it. Scott et al. conclude that their result "provide[s] support for the claim that inner speech has rich auditory content...that is provided by corollary discharge [efference copy]" (p. 291).

However, the Scott et al. result actually directly contradicts *Attenuation*. Assume that the prediction is /afa/ and the incoming signal is a hybrid between /ava/ and /aba/. There is a mismatch between /afa/, which is labiodental, and the hybrid, which is not labiodental. Predictive models hold that a mismatch between predicted and actual speech sounds results in enhancement in the processing of the actual speech sound relative to a match. But if this is the case, then subjects should hear a true hybrid. And, given that the subjects in Scott et al. are presented with a forced choice between /ava/ and /aba/, they should report the options at chance. Nevertheless, the Scott et al. subjects report /ava/ above chance when the prediction is /afa/.

But suppose that there is some articulatory/acoustic property that /afa/ but not /aba/ shares with the hybrid. The problem is that this property shared between /afa/ and the hybrid should be attenuated, according to predictive models. As a result, when subjects concurrently produce /afa/ while being presented with the hybrid, there should be an above chance probability of reporting *not* hearing /ava/. However, this prediction is directly contradicted by the results obtained by Scott et al. Therefore, the Scott et al. results do not support predictive models of inner speech. Instead, the Scott et al. result seems best explained by simple priming effects which cause an ambiguous stimulus to be interpreted in line with an imagined prime.

I have presented a representative sample of the studies on inner speech that are used to support *Attenuation*. I have argued that these studies in fact fail to validate this core prediction of the Standard and Radical Models of inner speech. The studies at most reveal priming effects of various kinds: Ford et al. shows repetition priming (repetition of a stimulus leading to a decrease in auditory activity), Whitford et al. shows expectation-based priming (an increase in the subjective probability that /ba/ will occur leads to a decrease in auditory activity when it does occur), and

Scott et al. shows perceptual priming (because /ava/ and /afa/ sound similar, the hybrid is heard as /ava/). Priming provides a much-needed alternative interpretation of these studies.

3.3.3 Generalizing the Result: Priming vs. Prediction

I now want to show that effects due to priming are in general inconsistent with effects due to prediction. As a result, if we can show that imagery generates priming effects, then we can conclude that imagery in general – and not just inner speech – is not to be identified with a prediction.

The central difference between priming and prediction is that whereas predictions make salient data that *conflict* with them, primes make salient data that are *consistent* with them. In effect, in priming what stands out is what is *similar* to the prime, whereas in prediction what stands out is what is *dissimilar* to the prediction. Support for this difference comes from empirical evidence showing that salience tracks mismatches between predictions and data. This is already observable in the example of the mormyrid fish: salient information – information relevant to navigation and prey location – is captured by that portion of the incoming signal that fails to match the prediction. There is also more direct empirical evidence that salience tracks mismatches between predictions and data. For example, predictable acoustic stimuli do not to capture attention and so are easier to ignore than unpredictable acoustic stimuli (Southwell et al., 2017). Such results have also shaped models of attention: Itti and Baldi (2005) have developed a model of visual attention according to which those regions of a visual scene that elicit the largest prediction error are most salient (see also Kaya and Elhilali (2014) for an example from the auditory domain). There is, in short, a wealth of converging empirical evidence showing that prediction error drives salience.

Whereas predictions make salient data that conflict with them, primes make salient data that are consistent with the prime. Theeuwes and van der Burg (2013) presented subjects with an uninformative visual object that matched the color of one of two subsequent test objects. Subjects noticed the matching test object before the non-matching one. This sort of priming result has also been observed in subjects asked to interpret the movement of an ambiguous display: subjects report a direction of movement that is consistent with the direction of a concurrently executed hand movement (Ishimura, 1995). In short, whether the priming is perceptual, semantic, or affective (McNamara and Holbrook, 2003; Eich and Forgas, 2003), priming, unlike predicting, makes salient those features that are consistent with the prime.

Now, if we can show that imagery engages priming effects, then this gives us reason to think that imagery does not have predictive effects, and so should not be modeled as a prediction. Research shows that imagery is associated with a wide array of priming effects. For example, Farah (1985) found that when a subject is presented with a faint image of an "H" or "T", the subject's report of which letter she sees tends to depend on whether she is visually imagining an "H" or "T": "H" if she is imagining an "H" and "T" if she is imagining a "T". Relatedly, Pearson et al. (2008) find that the perception of a bistable image – an image with two fixed interpretations – is biased in the direction of the visual imagery one is producing. Similar findings have been shown using fMRI (Lu et al., 2017), and have been discovered to be present also in the olfactory (Tomiczek and Stevenson, 2009) and auditory (Pitt and Crowder, 1992) modalities. Hence, the literature on imagery provides overwhelming evidence that imagery implicates priming effects.

Empirical evidence on prediction, priming, and imagery therefore suggests that imagery does not have predictive effects, and so is not identical to a prediction.[19] In this Section, I have argued that studies purporting to provide evidence that inner speech is associated with a prediction at most show that it is implicated in priming. Moreover, given that effects due to priming are opposite those due to prediction, I have argued that imagery in general is not to be associated with a prediction. The end result is that empirical work tells against not only the Standard and Radical Models of inner speech, but also predictive models of imagery more generally.

3.4 A Theoretical Problem with Predictive Processing Models of Imagery

I want to turn now to a theoretical problem with identifying imagery with a prediction. Recall that the Standard and Radical Models of inner speech differed in terms of how predictions were taken offline: whereas the Standard Model posited that motor commands were suppressed at the vocal tract, the Radical Model took prediction offline by decreasing the precision on prediction error. This Section will argue that because of this feature, the Radical Model fails to account for

[19] One might respond to this challenge by invoking the tools of the predictive processing framework. One might argue that imagery is not identical to a prediction as such, as I have been assuming, but that it is identical to a prediction that has a high level of precision. The result would be that in updating the high-precision prediction in light of data, one would stick close to the prediction. For example, in imagining a bird – generating a high-precision prediction that *a bird is present* – there will be a higher chance of perceiving a bird in the environment – any updates of the prediction will take into account the high confidence one has in the prediction. In this way, the predictive processing framework would be able to provide predictive models of imagery while also accounting for the priming effects present in imagery. The issue with this predictive processing explanation is that although priming has perceptual effects, it does so by having effects on *attention*. When shown a particular prime, one's attention is directed at features consistent with the prime. The problem is that, according to the predictive processing framework, attention involves an *increase* of precision on prediction error (e.g., Feldman and Friston, 2010; Hohwy, 2013, Clark, 2013). However, as we noted in Section 3.2, according to proponents of the predictive processing framework, imagery involves a *decrease* in the precision on prediction error. There is thus a puzzle in the predictive processing model now on offer: how can a decrease in the precision of the prediction error for some prediction-data pairs (imagery) lead to an increase in the precision of prediction error for that or nearby prediction-data pairs (attention)?

imagery without undermining predictive processing accounts of perception and action in general. Although I will again focus on inner speech, the argument is generalizable to predictive processing models of other forms of imagery.

Recall that the Radical Model identifies inner speech with a high-precision prediction – *there are incoming speech sounds in my own voice*. Despite the fact that there is a difference between this prediction and the data – *there are no incoming speech sounds in my own voice* – this does not result in speech perception or speech production because, according to Wilkinson and Fernyhough, prediction error is minimized via 'downmodulating' the precision of the prediction error. This model introduces the notion of *precision of prediction error*. They claim that decreasing the precision on prediction error cancels the need to minimize prediction error by revising either the prediction (which would generate speech perception) or the incoming data (which would generate speech production).

The problem with the introduction of this precision parameter is that there is nothing to stop a subject from *always* canceling revision no matter the magnitude of the prediction error. As I explained in Section 3.2, predictive processing holds that the mind has the sole aim of minimizing prediction error. But, once one introduces an independent precision parameter whose downmodulation cancels the need to revise either prediction or data, subjects have at their disposal a standing resource to always avoid the revision of prediction and data. If subjects are always able to simply downmodulate the precision of prediction error, then they have every reason to reject the life of perception and action in favor of a life of imagination: in conversation, I need not respond to my interlocutor, but only imagine responding; I need not see the obstacle in front of me, but only imagine seeing it; I need not drink, but only imagine the satisfaction of quenched thirst. Indeed, minimizing prediction error via revision of prediction or data faces far more

potential upsets and frustrations than simply reducing prediction error via decreasing the precision of the prediction error. For example, though I will always be able to downmodulate the precision on my prediction error, I may not be able to generate predictions that explain away incoming data, and I may not be able to bring about the actions that would confirm my predictions. The life of the imagination is not only possible, but, within the predictive processing framework, it is to be favored over the lives of perception and action. In the face of this problem – I will call it the 'fantasy problem' – the predictive processing theorist owes us a reason for thinking that the subject would not always take up downmodulation of this independent parameter once it has been offered as a standing resource.

A response on behalf of the predictive processing theorist can be formulated once we note the similarity between the fantasy problem and the much-discussed 'dark room problem'. The dark room problem underscores a family of issues that show that predictive processing fails to account for motivation. According to one version of the dark room problem, the most effective means of minimizing prediction error is to remain in a perfectly predicted, unchanging environment – a so-called 'dark room' – until one dies. If all one ever has to do is minimize prediction error, then it seems we should spend our whole lives in dark rooms. Since we in fact find ourselves engaging a variety of highly unpredictable environments, according to the dark room problem, predictive processing is inadequate in its account of how we go about perceiving and acting.

Though there is a superficial similarity between the dark room problem and the fantasy problem, there is a deep and significant difference. The creature of the dark room and the creature of fantasy both minimize prediction error by entering a situation – the dark room or fantasy – where revisions of predictions or data are never required. However, the means by which the

creatures achieve this state is different: the creature of the dark room finds an area in the world that is perfectly predicted to minimize prediction error, whereas the creature of fantasy uses tools intrinsic to predictive processing to accomplish the same goal. In other words, while the creature of the dark room must put itself in relation to the world such that it receives perfectly predicted inputs, the creature of fantasy need not do so, since its strategy eliminates the need to revise prediction and data whatever its relation to the world happens to be.

This difference between the problems is key to showing how standard responses to the dark room problem do not transfer to the fantasy problem. In response to the dark room problem, predictive processing theorists often argue that evolution has shaped us such that we would in fact generate *high prediction error* if we remained in highly predictable states. What gets the creature out of the dark room, according to this response, is that it is equipped with a prediction or 'hyperprior' concerning the rate of prediction error (Hohwy, 2013) or with predictions of "change, motion, exploration, and search" (Clark, 2013). In the spirit of this response to the dark room problem, one might claim that the creature of fantasy does not always downmodulate the precision of prediction error because, following Clark, our predictions are predictions to *perceive* the world or *act* in the world: they are predictions concerning *responding* to one's friend, *seeing* the obstacle, or *taking* a drink. Or perhaps, following Hohwy, the creature escapes fantasy because she has a hyperprior concerning the rate at which she revises predictions or data: never making good on the predicted quota of revisions would lead to high prediction error.

However, these dark room-inspired responses fail to address the fantasy problem. If we pursue the reply inspired by Clark, then the creature of fantasy can simply downmodulate the precision on prediction error generated by predictions about *responding* to one's friend, *seeing* the obstacle, etc. The Hohwy-inspired proposal fails for the same reason. No doubt there will be

prediction error borne of the fact that the hyperprior (that there be some quota of prediction-data revisions) fails to be carried out. But once again, there is nothing to stop the creature of fantasy from downmodulating the precision on *this* prediction error, thereby bringing the hyperprior into the fold of its imagination. In other words, even if a creature is equipped with predictions that she not fantasize, she can always downmodulate the resulting prediction error. This is because the creature of fantasy exploits tools inherent to the predictive processing framework.

The dark room problem is solved by adding predictions that would conflict with incoming data were one to remain in the dark room. In contrast, the fantasy problem cannot be solved by the addition of new predictions. The result is that the fantasy problem is far more insidious than its dark room counterpart. By making room for imagery, supporters of predictive processing fundamentally undermine their accounts of perception and action.

3.5 Toward a Non-Predictive Model of Inner Speech

In Sections 3.3 and 3.4 I have shown that predictive models of imagery are empirically problematic and theoretically inadequate. These arguments against predictive models of imagery can be bolstered if there is an alternative, non-predictive model that does better than its predictive counterparts at capturing imagery. In the current Section, I will provide a sketch of such a model of inner speech in terms of the mechanisms of the standard framework of motor control.

According to my Simple Model, inner speech is identical to the top-most sensory representation of speech sounds (green), where the inverse model has been taken offline (gray) (Figure 8).

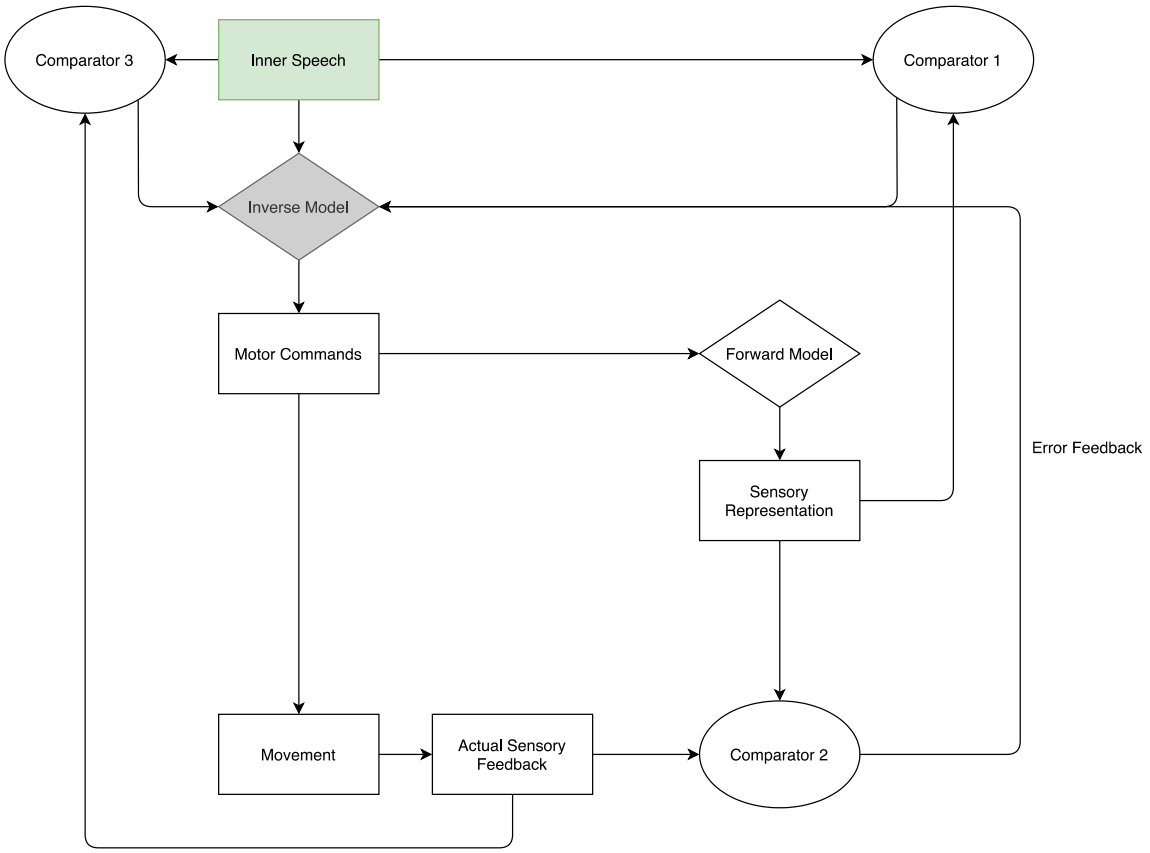

Figure 8: The Simple Model of inner speech

Central to the Simple Model is the idea that sensory representations are present *prior* to the construction of motor commands, and predictive processes more generally. This is accepted by a number of models of speech production (e.g., Guenther et al., 1998). Studies have shown that when the vocal tract is obstructed by a bite block or lip tube, speakers compensate by producing abnormal articulations, while nevertheless preserving the same speech sound region in auditory space (Abbs, 1986; Savariaux et al., 1995). This suggests that speakers select speech sounds prior to selecting motor commands and use speech sounds as a fixed target to guide the production of motor commands. The Simple Model exploits this agreed-upon architecture in claiming that inner speech is identical to the representation of speech sounds prior to the activation of the inverse model, which is taken offline.

By locating inner speech at a safe distance from predictive processes, the Simple Model has several advantages over the Standard Model. A central advantage is that the Simple Model is better able than its predictive counterparts to capture motor-dependent enhancements of inner speech.

Researchers have often studied a form of inner speech called 'silent mouthing' that engages the articulators but does not involve the actual production of sounds. A number of studies have shown that relative to 'pure inner speech', silent mouthing produces a more determinate and detailed form of inner speech. For example, as we saw above, Scott et al. (2013) found that subjects who hear a hybrid speech sound between /ava/ and /aba/ will tend to report perceiving the speech sound as /aba/ if they concurrently produce /apa/ in pure inner speech. These authors also found that this 'perceptual capture' of perception by inner speech was significantly stronger if subjects *silently mouthed* /apa/. In another widely discussed paper, Oppenheim and Dell (2008) also showed motor-dependent enhancements of inner speech. The authors studied the phonemic similarity effect, which is the substitution of a correct phoneme with an incorrect phoneme that is close to it in auditory/articulatory space (e.g., /pat/ to /bat/ is more likely than /pat/ to /fat/, since /b/ is closer to /p/ than /f/ is to /p/). The authors found that errors on tongue twisters in pure inner speech do not implicate the phonemic similarity effect, while silent mouthing does show such an effect. In effect, silent mouthing faithfully represents auditory/articulatory space, and so generates the phonemic similarity effect, while pure inner speech does not. These experiments suggest that engaging motor representations generates more detailed inner speech.

On predictive models, motor-dependent enhancement of inner speech is the result of more detailed motor commands resulting in more detailed predictions. The added detail in inner speech is the product of added detail in the motor command itself. However, this gets things backward:

at least within the standard motor control architecture, a precondition on generating a precise motor command is generating a precise auditory representation of speech sounds. If the auditory representation of speech sounds is itself indeterminate, then the motor commands issued by the inverse model will also be indeterminate. This suggests that, within the standard framework of motor control, motor-dependent enhancements associated with inner speech are not an *effect* of selecting particular motor commands, as predictive models must assume, but rather part of the *cause* of selecting detailed motor commands. Because the Simple Model views inner speech as the input to an inverse model which generates motor commands, it can accommodate the idea that silent mouthing generates detailed inner speech as a cause and not an effect of generating detailed motor commands.

Although I have at most sketched the Simple Model, it is already apparent that it has clear advantages over predictive models that have been criticized in Sections 3.3 and 3.4. There are of course a number of other states of imagery aside from inner speech that have received predictive treatments, including motor imagery (Jeannerod, 2006) and visual imagery (Grush, 2004). However, models of motor and visual imagery analogous to the Simple Model of inner speech may not be available to rescue these forms of imagery from predictive treatment. Much will depend on the details of the type of imagery under consideration, but I hope to at least have shown that the development of alternative, non-predictive models is needed and promising.

3.6 Undermining the Initial Motivation for Predictive Models of Imagery

I noted in the Introduction that part of the attraction of identifying imagery with a prediction is the idea that predictions are images. According to predictive processing theorists, a prediction

generates sensory data "for itself" (Clark), generates "its own sensory inputs top-down" (Kirchhoff), or generates "fictive (surrogate) sensory signals" (Seth). Indeed, a similar idea is present in the standard framework of motor control which pre-dates predictive processing models: predictions are there thought of as sensory representations of incoming data. Since predictive models view predictions as images, the identification of imagery with prediction is supposed to come for free with the adoption of such models – in Clark's works, imagery is part of a "package deal".

I want to close by showing that a properly trimmed-down conception of prediction fails to provide even incipient motivation for predictive models of imagery. Predictions are, at base, probability distributions over events. This notion of a prediction is already at home in the predictive processing framework. Probability distributions have all the statistical features that allow for predictive processing to get off the ground: probability distributions allow for the computation of error, can be updated, and have precision. This trimmed-down notion of a prediction is also compatible with the standard motor control framework. In particular, the representation of sensory consequences of executing a motor command can be modeled as a probability distribution over sensory information given some set of motor commands.

But there is nothing in the notion of a probability distribution over states of affairs that requires that it 'generate' 'virtual' or 'fictive' versions of real sensory data. For example, representing the probability of /bi/ as being .7 (or 1, etc.) does not on its own require the generation of 'virtual' or 'fictive' sensory signals related to hearing /bi/. Clark and others sometimes claim that predictions are images because predictive models are generative. The idea is that because predictions are part of a *generative model*, predictions *generate* 'virtual' or 'fictive' sensory signals for themselves. But, again, there is nothing about the notion of a generative model that suggests

that predictions need to be thought of as images. Generative models, at base, provide the probability of some data given some set of causes (Friston, 2005). But providing the probability of some data given some set of causes does not require bringing about a 'virtual' version of the data.

This Chapter has argued that identifying imagery with prediction is problematic on empirical (Section 3.3) and theoretical (Section 3.4) grounds. On the empirical front, I have shown that whereas imagery produces priming effects, predictions do not. And, on the theoretical front, I have shown that if predictions are images, then the most natural model of imagery – one that takes those predictions offline – undermines predictive processing models of perception and action. Although the Chapter has focused on predictive models of imagery, it ultimately suggests that we need to think more carefully about prediction, one of the fundamental building blocks of predictive models of the mind.

4.0 Of Substrates and Scientific Understanding: Explaining Hallucination

4.1 Introduction

To provide a broad and comprehensive understanding of a target system, science must often draw on a number of different explanatory levels. Explanations in psychiatry are no exception (Kendler, 2012). Take the case of visual hallucination. Suppose we know that visual hallucination is caused by excess dopamine absorption at the striatum; that the right interior insula is strongly implicated in hallucination; and that hallucination is tightly correlated with childhood abuse and trauma. That is, imagine that we have complete accounts of visual hallucination in terms of molecular, neural, and social factors. Do we thereby have the sort of broad and comprehensive understanding sought by a science of mental illness? Are molecular, neural, and social explanations sufficient for a comprehensive scientific grasp of visual hallucination? Many would say not. According to these authors, at least one crucial level of explanation would be missing – a *psychological* or *cognitive* level of explanation.

This intuition is nicely summarized in a paper by Patrick Fletcher and Christopher Frith:

> it is [difficult] to understand how a brain disorder can create new and compelling experiences. Such explanations require one to forge a link between brain activity and the subjective experience of a mind. Explanations such as 'hallucinations are caused by overactive dopamine receptors' are unsatisfying because they leave an explanatory gap between the mental and the physical. How can dopamine cause a voice or a belief? (2009, p. 49)

According to Fletcher and Frith, explanations in terms of molecular and neural factors alone are unsatisfying because they leave an 'explanatory gap between the mental and the physical'. To be clear, for Fletcher and Frith the 'explanatory gap' is not the perennial gap enshrined within the mind-body problem (Levine, 1983). Their appeal to hierarchical Bayesian models in closing the

gap shows that it is far more mundane: the gap merely arises from the absence of a psychological level of explanation (see also Frith (1992)).

John Campbell disagrees with this view of psychological explanation in psychiatry. Campbell argues that explanations in psychiatry do not, *contra* Fletcher and Frith, require appealing to a psychological level of explanation. But even Campbell seems willing to admit that psychological explanation is distinctive:

> there is some broader kind of explanation and understanding that we have when the functional architecture of the system is made explicit by a wiring diagram…Similarly, it may be strictly true that "alien thoughts are caused by the firing of dopamine neurons," even if a broader understanding is provided by a functional characterization of the psychological system that generates alien thoughts. (2013, p. 940)

If we only appeal to explanations at the molecular and neural levels, according to Campbell, we miss out on a 'broader kind of explanation and understanding' of mental illness. This distinctive sort of explanation is secured by appeal to a 'functional characterization of the psychological system', according to Campbell. That authors treat psychological explanation as distinctive should be no surprise: where molecular and systems neuroscience along with sociology provide explanations at molecular, neural, and social levels, the unique domain of cognitive science is thought to be explanation at a psychological level (Fodor, 1968).

I will understand psychological explanation as explanation in terms of psychological states and their transformations. For example, a standard psychological explanation of speech production appeals to a hierarchical transformation from semantic representations into lexical representations and then into motor representations (e.g., Levelt, 1993; see also Section 2.2 of Chapter 2). Of course, these representations and their transformations are implemented by neural and molecular factors – for example, an assault to Broca's area will result in speech production deficits. Yet, semantic, lexical, and motor representations are supposed to 'abstract' from these lower-level factors.

I will assume the dominant view that psychological explanation is distinctive: without it we fail to achieve a broad and comprehensive understanding of mental illness. The target of this Chapter is an increasingly popular pattern of psychological explanation, which I will label "substrate-based explanation". Substrate-based explanations seek to explain a pathological mental state in terms of a normal mental state kind whose phenomenological or functional profile is similar to that of the target state.[20] The mark of substrate-based explanation is the similarity that the normal mental state kind or 'substrate' bears to its target state.

Substrate-based explanations are frequently offered to explain the positive symptoms of schizophrenia, including hallucination, delusion, and thought insertion. Some authors have argued that auditory verbal hallucination is to be explained in terms of inner speech (Moseley and Wilkinson, 2014), while others claim it is to be explained in terms of auditory imagery (Wu, 2012). Delusion has been explained in terms of belief (Bortolotti, 2009), imagination (Currie and Ravenscroft, 2002), or some mixture therein (Egan, 2008). Finally, although most authors have assumed that the substrate of thought insertion is thought, some have suggested it is inner speech (Langland-Hassan, 2008). These explanations turn on the extent to which a given normal mental state kind – inner speech, auditory imagery, belief, imagination, etc. – is phenomenologically or functionally similar to the pathological mental state kind that underlies the positive symptom: the greater the match, the better the explanation of the pathological state in question.

The focus of this Chapter will be on substrate-based explanations of auditory verbal hallucination (AVH). AVH is the experience of perceiving a voice in a situation in which there is no actual voice to be heard (American Psychiatric Association, 2013). With an incidence of approximately 60-80% in people with schizophrenia, it is by far the most prevalent form of

[20] A mental state kind counts as normal to the extent that the kind is prevalent within the normal or typical population.

hallucination (Lim, et al., 2016). Ultimately, I will argue that the problem with substrate-based explanations of AVH is that they confuse phenomenological and scientific understanding of mental illness. Because substrate-based explanations take the form of a psychological explanation, they appear to provide scientific understanding of AVH. But, because they appeal to phenomenological similarity, they at most provide a phenomenological understanding of AVH. Going back to the example of visual hallucination: something is indeed lacking if we only appeal to molecular, neural, and social explanations, but if we fill in the gap with a substrate-based explanation we risk having achieved only an illusion of scientific understanding.

I open, in Section 4.2, by illustrating substrate-based explanations with examples from the philosophical and psychological literatures on hallucination. I then go on, in Section 4.3, to identify three assumptions that constitute substrate-based explanations of auditory verbal hallucination. The first assumption is that the normal mental state kind and AVH match both in phenomenology and linguistic content. The second is that this phenomenology and linguistic content are preserved by whatever process transforms the normal mental state kind into an AVH. And the third is that AVH possesses an abnormal phenomenology in addition to whatever phenomenology it shares with the normal mental state. Sections 4.4-4.6 are devoted to rejecting these assumptions: in Section 4.4, I show that there is likely no normal mental state kind that matches the phenomenology and linguistic content of AVH; in Section 4.5, I show that processes that range over normal mental state kinds are not sufficient to generate AVH; and in Section 4.6, I argue that AVH does not possess an abnormal phenomenology. Having rejected the framework of substrate-based explanation, in Section 4.7 I consider the objection that I have misconstrued AVH. I then go on, in Section 4.8, to sketch alternative psychological explanations. Finally, in Section 4.9, I provide a diagnosis of both the attraction and the error in substrate-based explanation

of AVH in terms of Karl Jaspers's distinction between phenomenological and scientific approaches to mental illness.

4.2 Substrate-Based Explanation

In the current Section, I illustrate substrate-based explanations both in terms of the philosophical and psychological literature and in terms of a recent debate over how to explain AVH.

4.2.1 Illustrating Substrate-Based Explanation

That hallucination is to be explained in terms of phenomenologically similar mental states has been a constant across much of the philosophical and psychological literature. Consider the following philosophical passages:

> when the external object of perception has departed, the impressions it has made persist, and are themselves objects of perception, and let us assume, besides, we are easily deceived respecting the operations of sense-perception when we are excited by emotion…so that, with but little resemblance to go upon…the less similarity is required to give rise to these illusory impressions. (Aristotle, 1984)

> But, except the mind be disordered by disease or madness, they [ideas] never can arrive at such a pitch of vivacity, as to render these perceptions altogether undistinguishable. (Hume, 1993)

> The assumption is that hallucinations are – or can be – produced by activating the same physiological resources as are involved in perception…. In the case of hallucination the product is some *private image* and so it must be so in the other case. (italics added) (Robinson, 2003)

Each of these quotes reveals a common explanatory strategy: a normal mental state kind is taken to match the phenomenology of a hallucinatory experience. In the case of Aristotle, an impression is taken to be a real percept because, in part, the impression bears some resemblance to a real percept. Because this impression is active in an abnormal context – when one is excited by emotion

– one treats the impression as a real percept. For Hume, an item that normally has a faint resemblance to a perception – an idea – reaches such a 'pitch of vivacity' in a 'disordered mind' that it is taken to be a perception. Finally, according to Howard Robinson, a private image with no causal connection to the world is taken by its subject to match the phenomenology involved in having a veridical perceptual experience. What makes these cases of substrate-based explanation is that hallucination is being explained in terms of a normal mental state kind – an impression, idea, or private image – that is phenomenologically similar to the resulting hallucinatory state.

Appeals to phenomenologically-matching substrates are also prevalent in the contemporary psychological literature on hallucination. Consider the following sample:

> Hallucinations…arise from a conceptually similar and more direct process: the abnormal salience of the internal representations of percepts and memories. (Kapur, 2003)
>
> Our model proposes that auditory hallucinations consist of the activation of auditory mental events that include memories and other currently active mental associations. (Waters, et al., 2006)
>
> It is generally acknowledged that hallucinations in schizophrenia are a result of the erroneous attribution of internally generated information to an external source. (Aleman, et al., 2003)

The same explanatory strategy is present in the psychological as in the philosophical literature: a normal mental state kind that matches the phenomenology of AVH – in this case, auditory imagery, auditory memory, or some other auditory mental event – is taken to explain AVH. As these vignettes illustrate, both philosophers and psychologists share an allegiance to substrate-based explanation.

4.2.2 A Recent Debate over Substrates

In order to get clearer on what substrate-based explanation consists in, I will now turn to a recent debate over whether *inner speech* or *auditory imagery* is the substrate of AVH. In what follows, I will understand inner speech as the sort of imagery implicated in the experience of a

"little voice in one's head" or "thinking in words" (Alderson-Day and Fernyhough, 2015). Auditory imagery is any imagery that depends on auditory perception. Thus, whereas imagery of chirping birds would count as auditory imagery, it would not count as inner speech. Conversely, while the imagery involved in inner speech is often auditory, there also seem to be types of inner speech that are non-auditory (Gauker, 2018).[21]

By far the most popular view is that the substrate of AVH is inner speech. In order to understand this position, it is important to understand the mechanisms thought to be involved in online speech production. Online speech production is standardly thought to involve both feedforward and lateral processes. The feedforward process involves an intention to speak that generates speech motor commands, which are then executed by the vocal tract in the form of audible speech. Along with this feedforward process, the system also predicts the auditory consequences of executing speech motor commands. This prediction is supposed to serve as a check on speech production: if there is a mismatch between the prediction and the actual auditory feedback, the system computes the difference as error, and thereby corrects the speech by sending a corrective motor command to the vocal tract (e.g., Pickering and Garrod, 2013).

As in the case of online speech production, in the case of inner speech, one produces an intention, which generates speech motor commands, but, unlike speech production, one *suppresses* those commands at the vocal tract (Figure 9, grey). Despite this suppression, the lateral process is still engaged and a prediction of auditory feedback is generated. Inner speech is identified with this prediction of auditory feedback (Figure 9, green): the subject hears herself 'in her head' but

[21] From the perspective of Chapter 2, inner speech theorists assume that the content of inner speech is exhausted by the 'Information' slot in Communicate (Voice, Information). Though this conflicts with my account of inner speech, in the present Chapter I will use 'inner speech' to mean what inner speech theorists intend by the term. Nothing of substance turns on this terminological point.

does not produce outer speech (Swiney and Sousa (2014), Carruthers (2018), Loevenbruck (2018)).

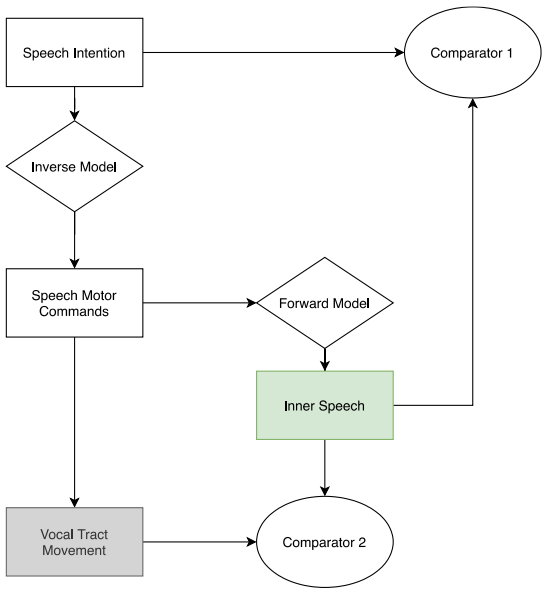

Figure 9: A model of inner speech in terms of simplified speech control mechanisms

Building on this model of inner speech, many authors argue that in auditory verbal hallucination, the prediction of auditory consequence fails to match the original speech intention (Figure 10, Comparator 1) (Swiney and Sousa, 2014). For example, one may generate the intention to say "The book", but the prediction may turn out to be "Send backup". According to these authors, this mismatch gives rise to "an unusual feeling of agency…[that] inner speech is outside of intentional control" (Swiney and Sousa, 2014). Because schizophrenics have deficits in attribution, they misattribute this inner speech to some external agent. Thus, in the example, someone with schizophrenia may experience a voice saying, "Send backup". On this view, AVH amounts to the experience of inner speech as 'alien' or otherwise attributed to some external agent.[22]

[22] This is not the only model of AVH that implicates inner speech. According to an earlier model, AVH is generated when inner speech fails to be *predicted* (Seal et al., 2004). On a more recent model, inspired by predictive processing,

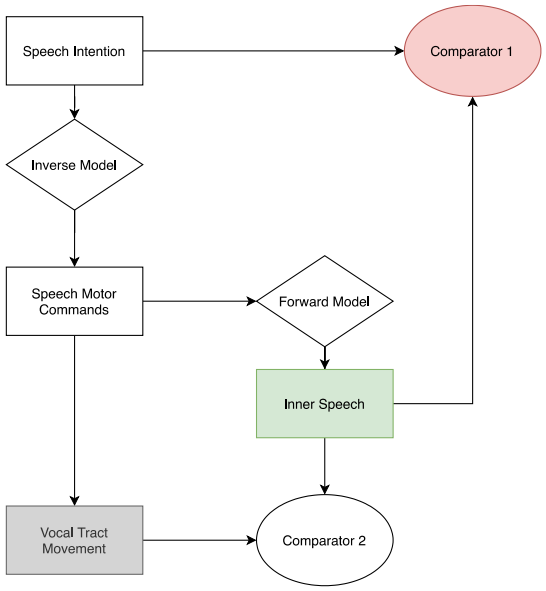

Figure 10: A model of auditory verbal hallucination in terms of simplified speech control mechanisms

However, inner speech models of AVH have come under attack on phenomenological grounds (Wu, 2012; Cho and Wu, 2013). Wu cites phenomenological reports suggesting that while AVH is often auditory in character, inner speech is "more often abstracted from an auditory format" (Wu, 2012). On the basis of this difference in phenomenology, Wu claims that in order for the predictive mechanism proposed by inner speech theorists to be viable it must be equipped with "an additional mechanism to effect the transformation" from a non-auditory state (inner speech) to a state with auditory phenomenology (AVH) (p. 94). However, once this transformation has been implemented inner speech becomes otiose in an explanation of AVH: the explanation is better off appealing only to an "auditory experience of another voice" (p. 94). In short, according

AVH is generated when the prediction error generated by producing a state of inner speech fails to be suppressed (Wilkinson and Fernyhough, 2017). I do not consider this predictive processing model of AVH in the current Chapter in part because I argued previously (in Chapter 3) that predictive processing models fail to account for inner speech. The criticisms I make in the current Chapter will also apply to both the earlier and predictive processing models of AVH.

to Wu, inner speech cannot explain AVH because inner speech does not possess phenomenology that matches that of AVH.

Cho and Wu (2013) suggest that inner speech theorists should claim that the substrate of AVH is *auditory imagery*: "on grounds of simplicity, we suggest that self-monitoring [inner speech] accounts should endorse auditory *imagination* of another person's voice as the substrate of AVH" (p. 2). On this view, AVH occurs when auditory imagery is both involuntarily and persistently activated. According to Wu, the automatic and persistent activation of auditory imagery causes the sort of "disruption and disturbance" indicative of AVH (Wu, 2012).

Inner speech theorists have responded to imagery theorists by arguing that, although inner speech is sometimes non-auditory in character, it is often auditory (Moseley and Wilkinson, 2013; see also Maiese, 2018). This allows them to evade Wu's objection, since there is no need to transform inner speech into a state whose phenomenological properties match those of AVH. Subsequent debate between the two camps has centered on the extent to which inner speech or auditory imagery match the phenomenological properties of AVH (Cho and Wu, 2014).

4.3 Deriving Conditions on Substrate-Based Explanations of AVH

Underlying the debate between inner speech and imagery theorists is a shared commitment to three interlocking assumptions – the *matching condition*, the *preservation condition*, and the *abnormal phenomenology condition* – which I will call into question in succeeding Sections (Sections 4.4-4.6). These three assumptions together constitute substrate-based explanation.

First, the debate rests on the idea that it is objectionable to explain AVH by positing some transformational process that changes the phenomenology and linguistic content of a normal

mental state kind to match that of AVH. The first assumption, then, is that the AVH and the normal mental state kind match in both linguistic content and phenomenology. AVH and the normal mental state kind must share the same *total linguistic content* on pain of positing an objectionable transformational process. But AVH and the normal mental state kind cannot share the same *total phenomenology* since the normal mental state kind would thereby automatically count as an AVH. To avoid this implication, substrate theorists need to distinguish two parts of the phenomenology of an AVH: one part that is identical to the total or partial phenomenology of the normal mental state kind – call this the 'standard phenomenology' – and another part that is not a part of the normal mental state kind, but is rather unique to the AVH – call this the 'abnormal phenomenology'. These terminological decisions allow us to precisely formulate the *matching condition*: for any explanation of some AVH *A*, where a normal mental state kind *N* is part of the explanans, *N* possesses all and only the linguistic contents of *A* and *N* possesses all and only the standard phenomenological properties possessed by *A*, but not its abnormal phenomenology.

The second assumption of the debate follows from the first: the aberrant process that operates over the normal mental state kind to generate an AVH must preserve the linguistic content and standard phenomenology of the normal mental state kind. If the transformation did not preserve these aspects of the normal mental state kind, then there would not be a match between the normal mental state kind and the AVH. This condition is supposed to be satisfied by the mechanisms offered by both inner speech and imagery theorists. In the case of inner speech theorists, the mismatch at the comparator brings about the feeling of other-authorship of inner speech but does not change the standard phenomenology and linguistic content of the inner speech. For an imagery theorist, the aberrant process – persistent automaticity – does not change the standard phenomenology or linguistic content of the auditory imagery. The *preservation condition*

therefore states that for any explanation of some AVH *A* in terms of some normal mental state kind *N*, there is an abnormal process *P* over *N* that preserves all and only the standard phenomenology and linguistic content of *N*.

The third assumption of the debate is that AVH possesses an abnormal phenomenology. This phenomenology is what is supposed to distinguish the AVH from the normal mental state kind. According to Wu and Cho, this abnormal phenomenology manifests in terms of a feeling of automaticity or involuntariness, while for inner speech theorists, it manifests as a feeling of other-authorship or alienness. While preserving standard phenomenology and linguistic content, the aberrant process also bestows abnormal phenomenology onto the AVH. Hence, the *abnormal phenomenology condition* states that for any explanation of some AVH *A* the explanans must explain the presence of some abnormal phenomenology P* in *A*.

In what follows, I will not be concerned with adjudicating the debate between inner speech and imagery theorists. Rather, I will target the pattern of explanation – substrate-based explanation of AVH – that serves as common ground for the debate.

4.4 Against the Matching Condition

I will now argue against the matching condition by showing that even the best candidates for matching the standard phenomenology or linguistic content of AVH fail to provide such a match. This argument will proceed in two steps, addressing two types of AVH: auditory and non-auditory (e.g., see Moritz and Laroi (2008) and Jones (2010)).

4.4.1 Auditory Imagery

The paradigmatic type of AVH is auditory. Consider the following description: "Most of the time I can hear it like it was just someone standing next to me. It's a different feeling than when you think words inside of your head…" (Cox, 2018). In light of such typical reports, I will assume (with imagery theorists) that auditory imagery is the best candidate to match the auditory type of AVH. However, there is converging evidence that auditory imagery fails to match the life-like auditory features present in AVH.

First, whereas auditory AVH is often reported to possess a particular level of loudness (Vercammen, et al., 2010), auditory imagery in general does not represent loudness. Pitt and Crowder (1992) presented subjects with two consecutive tones. After the tones were played, subjects had to judge whether the two tones matched in pitch. The authors found that reaction times were faster when the tones also matched in loudness. However, this priming effect was not observed in a corresponding imagery condition. In the imagery task, subjects imagined a reference tone either softly or loudly and then heard a subsequent tone.[23] Unlike in the perception condition, the authors did not observe faster reaction times when the tones matched in loudness. The authors concluded that loudness is not typically represented in auditory imagery. Furthermore, it has been shown that the life-likeness of changes in loudness in auditory imagery increases with musical experience (Bishop et al., 2013). Given that some voice hearers have very limited musical experience, this suggests that auditory imagery cannot account for the life-like quality of changes in loudness in AVH. The takeaway from such research is that the matching condition seems to

[23] Subjects underwent a training phase where they were played a reference tone at a loudness labeled "soft" or "loud" and practiced imagining the same level of loudness. The subjects were able to correct themselves using feedback.

fail with respect to loudness: although AVH involves the representation of loudness, auditory imagery typically does not.

Second, given that AVH is like hearing a real voice, the matching condition requires that the tempo of AVH match the tempo of the actual voice being represented. In other words, if I am representing my neighbor's voice in an AVH, the representation of the tempo of my neighbor's voice should match that of normal experiences of my neighbor's voice. However, research shows that auditory imagery is often degraded in its tempo. Janata and Paroo (2006) presented subjects with the first three notes of an ascending scale and had subjects imagine the rest of the notes at the same tempo. At a later time during the imagery task, a probe note was played. Subjects had to judge whether it was played when the last note of the scale would have been played. The authors found that subjects reliably failed to match the timing of the probe and the last note. This suggests that the tempo of auditory imagery is greatly degraded relative to the life-like tempo that is represented in AVH. This finding is reinforced by Halpern (1988), who found that subjects often 'level out' differences in tempo in imagery: subjects slowed the tempo of a fast song and speeded the tempo of a slow song. These studies suggest that the matching condition fails regarding the tempo of auditory imagery.

Third, AVH is often reported to involve the presence of a life-like voice, such that subjects represent the precise vocal timbre of some particular agent or other (Junginger and Frame, 1985). However, evidence shows that auditory imagery fails to represent crucial aspects of timbre. Using the same paradigm as their loudness experiment, Pitt and Crowder (1992) concluded that auditory imagery fails to represent dynamic timbre (whether sound onsets are gradual or abrupt). In addition, on the assumption that auditory imagery is conscious, it is plausible that subjects would at least report being able to represent timbre. However, even musically experienced subjects often

report being unable to represent the timbre of an instrument, or at least report that it is "non-veridical" or "patchy" (Bailes, 2007).

Taken together, these studies suggest that the matching condition fails for auditory imagery: auditory imagery often does not represent phenomenological features that are present in AVH, and when it does, it is not life-like.

An imagery theorist may object on two grounds.

First, one may object that I have shown only that *for the most part* auditory imagery fails to match AVH in loudness, tempo, and timbre. But there might be other tokens of auditory imagery – those that do match AVH in loudness, tempo, timbre, and so on – that would subserve AVH. In response, I am doubtful that there are such tokens of auditory imagery. These would have to be tokens that are not only life-like in terms of loudness, but also in terms of tempo, timbre, and so on for all the standard phenomenological properties of an AVH. But even if such a subset of auditory imagery exists, the substrate theorist is put in the awkward position of needing to explain why it is that these particular tokens of auditory imagery engage the aberrant processes that lead to an AVH as opposed to auditory imagery in general. Without an explanation of this difference between mismatching and matching auditory imagery, identifying the substrate of AVH with a life-like subset of auditory images seems *ad hoc*.

Second, one may object that I have cited studies that employ healthy subjects. According to the objection, imagery in those suffering AVH might be more like-like than in healthy subjects. However, it has been shown that those who are hallucination-prone often have the same imagery relative to non-prone subjects (Aleman and Haan, 1998). In particular, although subjective questionnaire reports have suggested that those suffering AVH have more acute and intense imagery (Mintz and Alpert, 1972), objective measures show that subjects have the same imagery

abilities. These objective measures include paradigms in which subjects judge the oddball from imagined sounds or judge stress on the basis of imagined lyrics, among other behavioral tasks (Aleman, et al. 2000). The fact that schizophrenic subjects regularly do not differ from controls on such measures suggests that the above findings are also applicable to schizophrenic subjects.[24]

4.4.2 Inner Speech

The second important subtype of AVH is non-auditory. Consider the following description: "I did not hear the voices aurally. They were much more intimate than that, and inescapable. It's hard to describe how I could 'hear' a voice that wasn't auditory…" (Woods et al., 2015).[25] The best candidate substrate for the non-auditory type of AVH would presumably be some form of abstract, non-auditory inner speech (see, e.g., Gauker, 2018). In arguing that such inner speech fails to fulfill the matching condition, I will target a difference in the linguistic contents of inner speech and AVH.

Tovar et al. (2018) recently transcribed the verbal content of AVH as subjects experienced it and found that the first-person was present in only 12 percent of reports, while the second- and third-person were present in 48 and 44 percent of reports, respectively – a pattern that is not present in the inner speech of people with AVH (Langdon et al., 2009). In addition, Hoffman et al. (2007)

[24] Additionally, imagery theorists may object that the substrate of AVH is *bottom-up* auditory imagery, but that my arguments have only targeted *top-down* auditory imagery. However, as discussed in the literature on imagery, imagery is defined as top-down. There are two things that the imagery theorist might have in mind by 'bottom-up imagery': involuntary imagery or perceptual representation. If the former, then the matching condition seems to fail for bottom-up imagery, since it is not clear how involuntariness would change the auditory features of imagery. If the latter, then, though the matching condition is satisfied, the imagery theorist would fail to provide an explanation of AVH (since AVH just is perceptual representation in the absence of an object).

[25] Although paradoxical-sounding, I will follow the psychological literature in averring the existence of auditory verbal hallucinations that are non-auditory, but nevertheless verbal in character. If the reader prefers, these can be typed as hallucinations that are non-auditory but communicative.

found that "46 percent of respondents reported that verbal content of voices [in AVH] was distinct from verbal thought [inner speech] either most of the time or all the time" (p.1170).[26] Overall, then, the phenomenological literature on inner speech suggests that the matching condition fails for inner speech: although inner speech is often in the first-person, AVH are not. (See Wu (2012) for a similar line of reasoning against inner speech theorists.)

4.4.3 An Objection

I have so far argued that the matching condition fails for auditory imagery and inner speech: auditory imagery does not match the standard phenomenology of AVH, while inner speech does not match the linguistic content of AVH. However, the substrate theorist might object that the preceding argument makes the implicit assumption that there must be a *maximally determinate* match between the normal mental state kind and AVH. These theorists might claim, instead, that there at most needs to be a match between the *determinables* of the substrate and AVH. For example, according to the objection, the substrate and the AVH need to match in terms of the determinables *auditory* and *verbal*, but can differ in the determinates that fall under these. Thus, the substrate can have an *inaccurate tempo* while AVH an *accurate tempo*, the substrate can be *first-personal* while the AVH *third-personal*, as long as both have *tempo* and *perspective*. The arguments above do not tell against this weakened matching condition.

[26] In contrast to these studies, however, Rosen et al. (2018) has recently found a correlation between subtypes of inner speech and AVH. In one finding, the authors found that psychotic patients had more dialogic inner speech relative to healthy controls. But, in light of the above mismatch between the contents of inner speech and AVH, these correlations should not be taken to suggest that dialogic inner speech is a *substrate* of AVH, as Rosen et al. claim. Rather, dialogic inner speech, and inner speech involving the voice of others, more generally, may be a way that subjects *cope* with their voices, e.g., by replaying past AVH episodes.

However, this weakening of the matching condition is not available to either inner speech or imagery theorists. This is because the aberrant processes that they posit do not involve any transformation of their respective substrates. For inner speech theorists, generating the prediction of incoming speech sounds is sufficient to produce inner speech. When there is a mismatch between the original intention and the prediction, this does not transform the linguistic content of the prediction in any way. The linguistic content of the AVH will just be the linguistic content of the prediction absent a mismatch. Moreover, for auditory imagery theorists, automatic and persistent auditory imagery does not transform the phenomenology of the imagery. Indeed, repetitive activation of auditory imagery presupposes that the *same* auditory imagery is being repeatedly activated. Now, if a model of AVH involves no transformation of the substrate, as in the inner speech and auditory imagery models, then the model requires that the substrate and AVH match in the *maximally determinate* standard phenomenology and linguistic content. Therefore, the explanations put forward by both inner speech and imagery theorists fail to satisfy the matching condition, regardless of whether it is interpreted in terms of determinates or determinables.

4.5 Against the Preservation Condition

I have so far argued that even the best candidate normal mental state kinds fail to satisfy the matching condition. But I cannot rule out that there is *some* normal mental state kind that satisfies it. Moreover, Wu is hesitant to pin down AVH to any single substrate, which suggests that we should not assess the matching condition in terms of whether some specific substrate matches AVH. As Wu states:

> in speaking of auditory experience, I am intentionally vague about whether the state is like auditory perception, auditory imagery (imagination), or auditory memory…. In the end, the account I favor can

> encompass all or a subset of these; what it emphasizes is that the automaticity of such auditory states as a basis of AVH. (p. 97, 2012)[27]

Although the debate between inner speech and imagery theorists focuses on the substrate of AVH, the debate may turn ultimately on the *transformations* that underlie the generation of AVH. For this reason, I will now target the preservation condition. According to the preservation condition, a process over some substrate preserves the standard phenomenology of the substrate, while adding an abnormal phenomenology.

Both inner speech and imagery theorists attempt to satisfy the preservation condition by appealing to the *unintended activation* of either inner speech or auditory imagery. One mark of inner speech and auditory imagery is that they are often under our control: in the normal case, I am in control of my inner speech in just the way I am in control of what I think, and I am also in control of the auditory images that I call up to consciousness. This control over inner speech and auditory imagery contrasts with AVH: a sufferer is generally not in control of the voices she hears. Unintended activation of inner speech and auditory imagery is supposed to account for AVH by eliminating the element of control that is normally present in these states. The result is supposed to be that unintended inner speech and auditory imagery emulate the perceptual representation of voices in their absence.

Unintended activation also fulfills the preservation condition. The standard phenomenology and linguistic content of a mental state is often preserved across intentional and

[27] Two notes about the plurality of substrates that Wu refers to here. First, given that Wu explains AVH in terms of the process of automatic and repetitive activation, the citation of 'auditory perception' should strike us as strange. This is because auditory perception is automatic – not controlled by intentions – and presumably only a single activation of an auditory perception would make for an AVH. Second, it is difficult to see how auditory memory is distinct from auditory imagination for the purposes of explaining AVH. Presumably AVH does not involve the feeling that one automatically and repeatedly *remembers* someone talking to them. If auditory memory also involves auditory imagery, then it seems likely that what explains AVH for Wu is simply the auditory imagery component. This suggests that despite appealing to a plurality of substrates of AVH, Wu actually has in mind states of imagery.

unintentional activations. No doubt the phenomenology of intentionally representing chirping birds differs dramatically from the phenomenology of unintentionally representing chirping birds: the former feels uneventful and controlled, while the latter feels jarring and abrupt, for example. Despite these phenomenological differences across both situations, however, there remains a stable phenomenological element associated with *the sound of chirping birds*. Unintentional activation is implemented in different ways by inner speech and imagery theorists. In the case of inner speech theorists, unintended activation is realized by a mismatch between intended speech and predicted speech, while in the case of imagery theorists, unintended activation is realized when there is no intention that controls or modulates auditory imagery. However, as I shall now argue, these processes are not sufficient for generating AVH.

Take the case of automatic and persistent activation, appealed to by Wu. The activation of an auditory state is automatic, according to Wu, in the sense that there is no feature of the auditory state that is the result of top-down modulation, that is, no feature of the auditory state is the result of an intention to produce that state. According to Wu, however, automaticity of auditory states is itself insufficient to generate AVH, since cases of having a tune pop into your head are also automatic. For this reason, Wu adds that the auditory state must also be persistent in the sense that it is activated repeatedly.

The problem for Wu is that obsessive-compulsive disorder (OCD) also often involves the automatic and repetitive activation of visual and auditory imagery (Lipton et al., 2010). For example, Brown (2006) describes a patient with a perpetual music track running through their head throughout the day. Though the perpetual track is disruptive and disturbing, it is not reported to be an AVH. Thus, the process posited by Wu to account for AVH is insufficient to generate it.

Rather, the process at best accounts for the repetitive and uncontrollable imagery we observe in OCD.

According to inner speech theorists, unintended activation is realized by a mismatch between an intention and prediction at a comparator. On this model of AVH, a match at the relevant comparator initiates a feeling of agency, whereas a mismatch does not (Swiney and Sousa, 2014) (see Figure 2). The problem is that the same mismatch is also implicated in a far more mundane phenomenon, *covert repair*, whereby subjects catch themselves before making a speech error. On standard speech control models, covert repair is thought to be subserved by a mismatch between intended speech and predicted speech prior to speaking (Postma and Kolk, 1993). The problem for the inner speech theorist is that the mismatch between the intention and the prediction does not seem to generate a feeling of loss of agency of the robust sort posited in explanations of AVH. In fact, it is quite the opposite: a condition on engaging in covert repair is that the speech that is predicted but not intended is recognized as a mistake made by *oneself*. As a result, even if schizophrenic subjects tend to misattribute self-produced information, we should not expect a mismatch between prediction and intention to lead schizophrenics to misattribute inner speech. Thus, the mismatch posited by inner speech theorists is insufficient to generate AVH.

As we have seen, the problem with appealing to the process of unintended activation is that it at best explains *mundane* pathologies of the mind. In the case of imagery theorists, it explains OCD, while in the case of inner speech theorists, it explains covert repair, but in neither case does it explain AVH. Why does appealing to unintended activation at best explain these mundane pathologies? This has to do with the adoption of the preservation condition: that the process that is supposed to generate AVH is also supposed to preserve the standard phenomenology and linguistic content of the substrate. Thus, if a state of imagery serves as input to a process subject

to the preservation condition, then the output of the process will also be *a state of imagery* – only degraded or altered in various ways. It is not surprising, then, that breakdowns in auditory imagery and inner speech lead to abnormalities relating to imagery, covert repair and OCD, and fail to lead to AVH. This suggests, more generally, that insofar as some process is subject to the preservation condition it will at best be able to account for deficits associated with the kind of mental state that serves as input. In turn, this puts pressure on the idea that normal mental state kinds serve as input to such a process, as is assumed by substrate-based theorists.

4.6 Against the Abnormal Phenomenology Condition

I have so far argued against two planks of substrate-based explanations of AVH, the matching condition and the preservation condition. First, I argued that even the best candidate substrates fail to match the standard phenomenology and total linguistic content of AVH. Second, I argued that the process of unintended activation fails to generate AVH. However, these criticisms have targeted only dominant instantiations of the matching and preservation conditions, leaving open other possible substrates and other possible aberrant processes that might explain AVH.

The third and final assumption of substrate-based explanation is that AVH possesses an *abnormal phenomenology*. The abnormal phenomenology condition brings in its wake general versions of the matching and preservation conditions. So, if we reject the abnormal phenomenology condition, we undermine the matching and preservation conditions themselves, and not just the dominant instantiations of the conditions I have examined above. How does the abnormal phenomenology condition imply general versions of the matching condition and preservation condition? Abnormal phenomenology just is phenomenology that is not normally

possessed by some mental state kind. But this means that the mental state kind would be normal if it did not possess the abnormal phenomenology. This is part of what is required by the matching condition. Moreover, if the normal mental state kind is to come to possess an abnormal phenomenology, then there must be some process that brings that about. But that is just to accept part of the preservation condition: that there is a process that adds abnormal phenomenology to a normal mental state kind. Therefore, by attacking the abnormal phenomenology condition we thereby undermine general versions of the matching and preservation conditions.

Most theorists believe that the abnormal phenomenology of AVH is *the feeling of a lack of control*. This putative abnormal phenomenology goes by a number of different descriptions, including 'alienness', 'extraneity', 'other-authorship', 'automaticity', 'spontaneity', and the like. The following quotes express this dominant view:

> [the] proposal is then made that top-down factors…lead to unintended inner speech being experienced as *other-generated* (italics added) (Jones and Fernyhough, 2007)

> the mismatch in the second comparator between the desired state and the predicted state would mean that an *unusual feeling of agency* would accompany the associated inner speech, potentially a *feeling that inner speech is outside of intentional control* (italics added) (Swiney and Sousa, 2014)

> …all accounts of AVH must explain (A) the *spontaneity of AVH episodes* (they often just happen)…Both accounts can deal with the spontaneity: in self-monitoring, it is explained by spontaneous failure of self-monitoring so that AVH *feels spontaneous*; in the spontaneous activity account, it is the actual spontaneous activity of auditory areas (italics added) (Cho and Wu, 2013)

Imagery theorists capture the feeling of a lack of control in terms of the feeling of passivity and automaticity, while inner speech theorists capture it in terms of a feeling of alienness or other-authorship (see also Aggernaes (1972), Nayani and David (1996), Garrett and Silva (2003), and Hoffmann et al. (2008)).

In what follows I challenge the idea that AVH possesses a feeling of a lack of control as part of its abnormal phenomenology. Instead, I argue that a feeling of a lack of control is part of the normal phenomenology of AVH. To do this, I first contrast subjective reports of OCD and

anarchic hand syndrome, both of which cite a feeling of a lack of control, with reports of AVH, which do not cite such a feeling. Consider the following examples of OCD and anarchic hand syndrome, respectively:

> I have constant intrusive thoughts in any situation, and in no particular situation, about cannibalising, slaughtering or eating my animals…Being close to people and animals makes me feel extremely uneasy, and I often observe phantom sensations of hunger and excess salivating... (Anonymous, reddit.com r/OCD, 2020)

> A 77-year-old woman with chronic atrial fibrillation had her anticoagulation stopped temporarily for spine surgery….Two days later, while watching television, she noted her left hand flinging across her visual field. Her left hand stroked her face and hair without her will. She got terrified. Her attempts to control the left hand with the right hand were unsuccessful. (Panikkath et al., 2014, p. 219)

The sufferer of 'harm OCD' does not want to think thoughts about cannibalizing her animals – she cites the unease of being in close quarters with them – but has those thoughts anyway. Similarly, the patient with anarchic hand syndrome reports her inability to control the movements of her left hand. A feeling of a lack of control is thus cited as part of the abnormal phenomenology of these conditions. What is felt to go wrong has to do with the inability to control thoughts or movements, respectively.

Contrast these reports with those of AVH, where we do not find reports of a lack of control. The following are representative samples taken from a schizophrenia-related forum:

> I only get hallucinations now when under a lot of stress but when going through one of those times the most common thing I'd hear was screaming, like blood curdling screams. Those were outside my head. I'd also hear whispering right in my ear, someone calling my name…

> To me the voices always sound like someone is saying something to me just outside the room i'm [*sic*] in or having a conversation just outside. Usually it is 3 voices. One constantly says he wants to kill me or hurt me in some way [*sic*]. The other two can go from trying to talk him down or to saying they want to kill me too.

> I only had auditory hallucinations twice…The other time I was on the street and it was a guy calling my name, I thought it were some random dudes on [*sic*] a coffee shop and ran away from them scared shitless. (Anonymous, forum.schizophrenia.com; 2016)

Unlike the reports of OCD and anarchic hand syndrome, these reports of AVH do not cite a feeling of a lack of control. In the first case, we have a report of hearing screaming, in the second a report of hearing people talking outside the room, and in the third a report of hearing someone calling the subject's name.

What determines whether or not the feeling of a lack of control is cited as part of the abnormal phenomenology of these conditions? Take first the case of OCD and anarchic hand syndrome. A feeling of a lack of control shows up as abnormal in these conditions precisely because thoughts, on the one hand, and limb movements, on the other, are *normally* the sorts of things that are under some kind of control. My hand movements are normally under my control, so when my hand moves in the absence of that control, that lack of control is experienced as abnormal. Although the sense in which we are in control of our thoughts is perhaps more difficult to pinpoint, when we experience thoughts outside of our control, we experience it as abnormal because, in the relevant sense, thoughts are normally under our control.

Why, then, do reports of AVH not cite a feeling of a lack of control? The above reports of AVH are reports of the experience of *speech* or *voice perception* – hearing people talking outside the room, and so on. If we take them at face value, they suggest that AVH is an instance of the types *experience of speech perception* or *experience of voice perception*. However, unlike thought and movement, speech and voice perception normally involve a feeling of a lack of control: I cannot control *what* I hear you say nor can I control my *hearing* what you say. (At most I might manipulate your responses via conversational scheming (in the first instance) and might cup my ears (in the second instance), but this is not the sense of control at issue here.) Thus, if AVH is a species of the experience of speech or voice perception, then lack of control enters into AVH as a *normal* phenomenological factor, not as an extra, *abnormal* phenomenological factor. Indeed, if AVH is a species of speech or voice perception of the sort illustrated in the above reports, then it seems that AVH does not possess an *intrinsic* abnormal phenomenology in general. But surely there is *something* that is felt as abnormal about AVH by those who suffer from it. I suggest that what is experienced as abnormal is simply that it involves hearing a voice *where none is present* –

that is the frightening bit. According to this alternative, there is no deeper phenomenal fact about the felt abnormality of AVH.

Given that the abnormal phenomenology condition brings in its wake general versions of the matching and preservation conditions, rejecting it undermines the latter two conditions as well. If AVH does not possess an intrinsic abnormal phenomenology, then AVH is not to be explained in terms of a mental state that would be normal save for the abnormal phenomenology (matching condition). If AVH does not possess an intrinsic abnormal phenomenology, then there is no aberrant process that adds an abnormal phenomenology to an otherwise normal mental state kind (preservation condition). Since the abnormal phenomenology condition is not satisfied, the general framework of substrate-based explanation of AVH seems to be misguided.

4.7 An Objection from Inner Voices

I have challenged substrate-based explanations of AVH by arguing against the matching, preservation, and abnormal phenomenology conditions. One way to defend substrate-based explanation would be to argue that I have failed to attend to a particular subset of AVH, so-called 'inner space' or 'in the head' AVH (e.g., Jaspers, 1963). These are instances of AVH wherein voices are represented as located in one's head or within one's 'inner space' as opposed to being represented in external space. In addition, these 'in the head' AVH tend to be less life-like than its external space counterpart (see Berrios and Dening, 1996).[28]

[28] To be clear, 'in the head' AVH is not thought insertion. In the latter, another person's thoughts are felt to be in one's own head. In contrast, 'in the head' AVH does not involve this experience of 'insertion'.

One might claim that substrate-based explanation is supposed to explain 'in the head' AVH, but not its external space counterpart. Given that 'in the head' AVH is phenomenologically less life-like than its external counterpart, inner speech and auditory imagery may phenomenologically match 'in the head' AVH. Moreover, versions of the preservation and abnormal phenomenology conditions would be satisfied if 'in the head' hallucinations are marked by a process that generates a feeling of a loss of control. According to this objection, 'in the head' AVH is inner speech or auditory imagery that has been 'tweaked'.[29]

This defense of substrate-based explanation is problematic for two reasons.

First, substrate-based explanations of AVH are often intended as explanations of AVH in general. Thus, if we have been successful in showing that these explanations do not apply to external space AVH, then we have dramatically narrowed the scope of substrate-based explanation.

Second, there is empirical evidence suggesting that substrate-based explanations fail to account for AVH in general, including the 'in the head' variety. Hoffmann et al. (2008) found that the primary way that sufferers were able to distinguish their own verbal thoughts from AVH was in terms of the presence of *a non-self speaking voice*: the non-self speaking voice is present in the case of AVH, but is not present in the case of verbal thought. The problem is that inner speech and auditory imagery can both involve a non-self speaking voice. As Moseley and Wilkinson (2014) claim in their response to Cho and Wu (2013), we often engage in dialogic inner speech where one represents the voice of another. And, it is routine that our auditory imagery of voices involves the representation of the voices of others. For this reason, theories that implicate inner speech or auditory imagery as substrates of AVH would predict, contrary to Hoffmann et al.'s

[29] I thank Peter Langland-Hassan for this objection.

finding, that subjects should not be able to distinguish AVH from verbal thought if verbal thought is subserved by inner speech or auditory imagery. Although Hoffmann et al. do not distinguish between 'in the head' and 'outside the head' AVH, this evidence suggests that inner speech and auditory imagery are not substrates of either. 'In the head' AVH needs to be accounted for outside the framework of substrate-based explanation.

4.8 Sketching Alternatives to Substrate-Based Explanations of AVH

I have challenged the framework of substrate-based explanation by undermining three assumptions that constitute the framework: the matching, preservation, and abnormal phenomenology conditions. At the outset I noted that substrate-based explanations are only a subset of the psychological explanations of psychiatric illness. For this reason, undermining the planks of substrate-based explanation leaves untouched a number of psychological explanations of AVH that are not based on substrates. I will provide a sketch of two of these alternatives.

The first psychological explanation appeals to perceptual salience. Perceptual salience is the phenomenon whereby we find our attention drawn to certain things in our perceptual environment. Salience is driven both by bottom-up and top-down processes. In terms of bottom-up processes, perceptual salience is largely driven by *contrast* within the dimensions of orientation, color, intensity, and the like: areas of a scene with the most contrast within some dimension are marked out as most salient (Itti and Koch, 2001). But perceptual salience is also the product of top-down factors, such as expectation and desire. Thus, while the white-on-black of a waiter's uniform may normally draw one's attention, if one is hungry what will stand out is the plate of food being served by the waiter.

I think we can use this idea of perceptual salience to account for at least some forms of hallucination, particularly so-called "hyper-vigilance" or "environmentally-based" hallucinations. In these hallucinations, one's perceived environment partially constitutes one's hallucination. Take the following example:

> The…hallucinated voice was perceived solely when Mr. A simultaneously heard real engine sounds from motor vehicles. The engine sounds and the voice were perceived as "in parallel." The "engine voice" spoke to him in the second person, uttering frightening statements such as, "I've got hell for you." The timbre of this voice was mechanistic, like the accompanying engine sounds, and lacked human characteristics such as gender or accent. (Hunter and Woodruff, 2004).

What is happening in the case of the 'engine voice', I suggest, is that a part of the cognitive machinery underlying perceptual salience, 'salience maps,' are being fundamentally warped by top-down processes. On this sort of explanation, Mr. A has a background paranoid delusion that someone is out to get him, which drives the salience map to select as salient those acoustical parts of the actual engine sound that would fit the theme of the paranoid delusion. The result is that certain parts of the actual engine sound are foregrounded while others backgrounded, thereby producing an experience of the statement, "I've got hell for you", as composed from engine sound acoustics.

This sketch of a cognitive explanation of AVH is not substrate-based. In the first instance, the paranoid delusion that drives top-down processing is not the sort of thing that could phenomenologically match the statement, "I've got hell for you", coming from the engine. Similarly, what it is like to hear an engine sound on its own is not what it is like to hear an *utterance* embedded within an engine sound, so there is no phenomenological match between the actual engine sound and the utterance made by the engine voice. Given that the matching condition fails, explanations of AVH that appeal to salience maps are generally not substrate-based explanations.

Of course, most AVH is not environmentally-based, but is instead unconnected to the surrounding environment. Hunter et al. (2006) have proposed that the basis of such AVH could be

spontaneous baseline auditory activity. Baseline auditory activity is the intermittent activity observed in auditory cortex during complete silence. According to Hunter et al., while baseline auditory activity does not produce an auditory experience in healthy subjects, it becomes amplified, distorted, and auditorily experienced in schizophrenic subjects (p. 193). According to this explanation, AVH is explained by a particular psychological transformation over a baseline auditory state. What is important for our purposes is that this form of explanation is not based on substrates. Given that baseline auditory activity has no phenomenology – it occurs during complete silence – it is not the sort of thing that could satisfy the matching condition.

Although I have only provided sketches of alternative styles of psychological explanation of AVH, I hope to have given the reader a sense that these alternatives are promising.

4.9 Toward a Diagnosis: Scientific and Phenomenological Understanding

If substrate-based explanations of AVH are fundamentally misguided, why have many theorists within both philosophy and psychology been attracted to them? I will now use Karl Jaspers's distinction between 'explanation' and 'understanding' to diagnose both the attraction and the error of such explanations of AVH.

Jaspers identified two approaches to mental illness. One approach, which he termed 'explanation', involves providing an explanation of mental illness using empirical methods of the natural sciences. Another approach, termed 'understanding', involves providing a description of what it is like to undergo some mental illness. These two approaches provide different sorts of correlative *understanding* of mental illness: statements that provide a naturalistic explanation of mental illness provide us with *scientific understanding* of mental illness, while statements that

describe what it is like to undergo some mental illness provide *phenomenological understanding* of mental illness.[30] What I want to suggest is that substrate-based explanations conflate the explanatory and phenomenological approaches, and thereby confuse mere phenomenological understanding of AVH with scientific understanding of AVH.

Jaspers is popularly associated with the idea that at least some primary delusions are phenomenologically 'un-understandable': while theorists who have not experienced delusion are able to provide an explanation of primary delusions, they are unable to grasp the phenomenology of delusion. However, in recent literature, a number of philosophers have pushed back against the limits that Jaspers set on phenomenological understanding (see, e.g., Pickard (2010), Ratcliffe (2013), and Kendler and Campbell (2014)). For example, Kendler and Campbell (2014) claim that anarchic hand syndrome can be made phenomenologically understandable via the following role-play:

> Sit in a chair with your right elbow resting on the armrest. Bring your arm from horizontal to a 45 degree angle and then put it down again. Note your subjective sense of that movement. Then have your friend lift your arm in exactly the same way and return it. The subjective feeling is quite different. (p. 4)

Such simulations are also used in clinical psychiatry to both empathetic and therapeutic ends. It has been shown that clinicians are able to develop empathy for voice-hearers by wearing headphones that generate utterances of the sort heard by those suffering from AVH (e.g., Bunn and Terpstra, 2009). Moreover, a number of recent therapeutic protocols have used virtual reality-type technology to simulate voices. For example, in the AVATAR protocol, patients create computer simulations of what their AVH voices sound like and what their possessors might look

[30] There is much recent work in the philosophy of science on the nature of scientific understanding (see, e.g., Lipton (2009), Potochnik (2017), Wilkenfeld (2019)). The discussion of scientific understanding in this Section focuses only on its relation to phenomenological understanding. As a result, to my knowledge, my discussion will be consistent with much of the literature on scientific understanding, and can be taken as complementary to it. At most, I am assuming, along with much of this literature, that it is possible for scientific explanations to provide scientific understanding.

like. Patients then engage in a dialogue with their avatar, whose statements are controlled by a therapist, in an attempt to take power back from it. The therapy has led to a decrease in overall voice-hearing severity (Craig, et al., 2017). The improvement in severity suggests that the simulated voices in fact reflect the phenomenology of the AVH experienced by the patient. Simulations therefore seem to provide a phenomenological grasp of AVH.

I suggest that much of the attraction of substrate-based explanations of AVH is that they provide the theorist with just this sort of simulation-based, phenomenological understanding of AVH. In the context of substrate-based explanations, AVH is understood as just like what would happen to *me* if *my* auditory imagery were automatic and persistent (imagery theorists) or as just like what would happen to *me* if *my* inner speech were to not feel self-produced (inner speech theorists). These explanations put AVH within phenomenological reach by providing theorists with a basis for simulating, for themselves, what it would be like to undergo an AVH. In just the way that Kendler and Campbell's instructions provide a way in which I can simulate what it is like to undergo anarchic hand syndrome, substrate-based explanations are attractive because they provide a recipe for a phenomenological grasp of AVH.

The problem arises, however, when that which provides a phenomenological understanding of AVH is projected into *explanations* of AVH. Experiences and simulations that allow one to grasp what it is like to undergo some mental state give one no reason to suppose that that these experiences also feature in an explanation of that mental state. Indeed, if this were the case, then the kind of experiences had when clinicians listen to AVH simulations would be part of the explanation of AVH itself! Substrate-based explanations of AVH are a product of inappropriately projecting a means of acquiring a phenomenological understanding of AVH – simulations – into psychological explanations of AVH. Thus, it is one thing for a theorist to note

that there are phenomenological overlaps between inner speech and AVH – both involve voices – but it is quite another for the theorist to then assume that because of this inner speech is a viable substrate of AVH (inner speech theorists). Similarly, it is one thing for a theorist to argue that auditory imagery is a better phenomenological match with AVH than inner speech, but it is quite another to go on to claim that because this is so it is the substrate of AVH (imagery theorists). Jaspers was keen on warning against this kind of slide between phenomenological and scientific approaches to mental illness:

> Empathetic psychology [phenomenological description] has totally different tasks from performance-oriented psychology (objective psychology), which developed out of physiology. The two do not impinge on one another and neither has the right to criticise the other since both pursue entirely different aims….The mistake first arises when they replace one another and mistakenly want to transpose something from one field to the other. (Jaspers, 2007; p. 177)

In the context of AVH, the 'transposition' involves reading a phenomenological description of AVH into a psychological explanation of AVH. Moreover, given that psychological explanations are supposed to underwrite scientific understanding, the result is that, in grasping a substrate-based explanation, one at most appears to have a scientific understanding of AVH. Phenomenological understanding of AVH perniciously disguises itself as scientific understanding.

Of course, I have not shown that substrate-based explanations *in general* do not provide scientific understanding of their targets. Much of the reasoning in this Chapter is focused on features that are unique to AVH, and so it may be that substrate-based explanations of delusion or other mental illnesses do provide scientific understanding. However, the argument of the Chapter should make us at least cautious in adopting the sort of phenomenological-explanatory strategy indicative of substrate-based explanations. At the outset I registered my agreement with the idea that psychological explanations provide us with a distinctive understanding of mental illness. The warning of the Chapter is that there is a subset of psychological explanations – substrate-based

explanations – that may result in mistaking a phenomenological understanding for the sought-after scientific understanding of mental illness.

5.0 Conclusion

This dissertation has targeted a picture that dominates current research on inner speech: that inner speech is essentially a speech phenomenon. This picture showed up in Chapter 2 as the assumption that the content of inner speech is to be derived solely from the speech processing hierarchy. In Chapter 3, the theme is exhibited in the idea that inner speech is a prediction whose function it is to guide speech production. Finally, in Chapter 4 the view showed up as the tendency to leverage the putative position that inner speech has in speech production to provide an account of auditory verbal hallucination. The idea that inner speech is essentially a speech phenomenon dominates discussion of the content of inner speech, the role of inner speech within functional architecture, and the explanatory significance of inner speech.

I have tried to dislodge the dominant picture in each of its expressions. In Chapter 2, I argued that the content of inner speech is not speech-based, but vocalic. The result is that inner speech is a vocal phenomenon, not a speech phenomenon. In Chapter 3, I argued that inner speech is not a prediction that guides speech production. I replaced this view with one on which inner speech – or at least one component of inner speech – is a goal state that is not involved in the active guidance of speech production. In Chapter 4, we found that inner speech models of AVH fail to provide explanations of AVH, but at most provide a phenomenological understanding of what it is like to undergo AVH. The result of the dissertation is that inner speech is liberated from the entrenched presupposition that inner speech is essentially a speech phenomenon.

The process of so liberating inner speech has given rise to a number of themes that unite the dissertation. One theme concerns the relationship between the philosophical and empirical domains. The dissertation has illustrated the way that empirical work can shed light on

philosophical questions about inner speech. This theme was exhibited in Chapter 2, where we appealed to empirical research on vocal processing to develop a pluralistic view of inner speech. At the same time, the dissertation has also illustrated the way that empirical work harbors philosophical prejudices. One of these prejudices is the endorsement of the dominant picture mentioned above. Another is exhibited in Chapter 4, where we found that substrate-based explanations of AVH (often appealed to in empirical research) are grounded in a confusion between phenomenological and scientific understanding.

A second theme of the dissertation is a nuanced skepticism about using phenomenological considerations in theorizing about the mental. Although I have relied on phenomenological reports to infer the existence of different forms of inner speech (see Chapter 2), I have criticized ways in which psychologists and philosophers have illicitly used phenomenology to guide explanation. In Chapter 4 we found that substrate-based explanations of AVH are the product of transposing phenomenological considerations onto explanatory ones. Recognizing that inner speech is phenomenologically similar to AVH, theorists mistakenly infer that it is the substrate of AVH. Thus, although phenomenology may sometimes be a reliable guide to mental state kinds, it is not on its own a reliable guide to explanation of the mental.

There are several future projects that extend these themes while also further challenging the dominant picture and filling out alternatives.

Chapter 2 argued for vocalism, according to which inner speech is characterized in terms of vocal content – contents of the form Communicate (Voice, Information). One future application of vocalism concerns the metacognitive character of inner speech. On a traditional picture, thoughts (propositions) themselves cannot be known or thought about unless they are transformed into a linguistic or auditory register (e.g., a format or content type) (Ryle (2009); Carruthers (2011);

Frankish (2004); Bermudez (2018); Bar-On and Ochs (2018)). On this picture, the metacognitive character of inner speech is that it involves a transformation of propositions into a linguistic or auditory register. However, if inner speech is not to be accounted for in terms of transformations within the speech processing hierarchy, as I argued in Chapter 2, then this suggests that the metacognitive character of inner speech may not lie in transforming propositions into levels lower down the hierarchy. Vocalism provides for an alternative, more deflationary approach, on which inner speech is metacognitive just because when I engage in inner speech, I represent *my own voice* communicating some information. The link between inner speech and metacognition does not consist in code transformation – the transformation of a proposition into a linguistic or auditory register – but in the mere fact that I represent my own voice. In the future, I plan to develop the contrast between these two pictures and argue for the deflationary view.

Chapter 3 argued that inner speech, and imagery more generally, should not be identified with a prediction. This raises the question of how we might experimentally decide whether states of imagery are identical to states of prediction. Existing experiments have relied on the match/mismatch paradigm. According to the match/mismatch paradigm, if a representation correlates with certain electrophysiological effects when it matches versus mismatches a presented stimulus, then we can infer that the representation is a prediction. However, a number of implementations of this paradigm fail to control for possible confounds. Some implementations fail to control for repetition suppression (Ford et al., 2001a), while others fail to control for electrophysiological effects due to amodal expectations (Whitford et al., 2017). This suggests the need for an alternative implementation of the match/mismatch paradigm, which I offered in a footnote in Chapter 3 and which I hope to develop further in future work. In general, Chapter 3

suggests we need to think more carefully about experimental paradigms that dissociate effects due to imagery and effects due to other processes.

Chapter 3 also has bearing on the connection between imagery and skill. Imagining *x*-ing (serving a tennis ball, shooting a basketball, lifting weights, etc.) results in better performance in *x*-ing (e.g., Green, 1992). Although prediction-based models of imagery cannot accommodate the role of imagination in skill, a view such as that of the Simple Model presented in Chapter 3 can shed light on how imagination can lead to better online performance. Prediction-based models seem to hold the promise of explaining the link between imagination and improved performance since the prediction error generated by mismatches between predicted and actual sensory feedback are used to refine motor commands. The problem is that when taken offline there is no actual sensory feedback, and, as a result, there is no prediction error signal available to fine tune motor commands.

In contrast, a view in line with the Simple Model of inner speech suggests a way of accounting for the link between imagery and skill. According to the Simple Model, inner speech is the input to an internal model that generates speech motor commands. As such, inner speech has the character of a *goal state*, representing a sensory endpoint that motor commands strive to bring about. We can generalize the Simple Model to account for how imagination results in improved performance. If states of imagery are goal states, then, (say) in kinesthetically imagining how one's wrist turns when one serves a tennis ball, one is settling on a more precise, finely-tuned *goal representation*. Imagining, in this context, is just a way of figuring out what it is exactly one is to be doing. On the picture of imagery as prediction, it is assumed that the subject *already* knows what is to be done: imagery just allows her to refine lower-level, implementation-oriented

processes. In contrast, on my goal-based picture, imagery involves figuring out what one is to do in the first place.

Chapter 4 argued that substrate-based explanations of auditory verbal hallucination trade on a confusion between phenomenological and scientific understanding of AVH. Substrate-based explanations are explanations wherein a pathological mental state is explained in terms of a normal mental state kind that is similar to the phenomenological or functional profile of the target state.

In one strand of future work, I hope to examine the scope of substrate-based explanation. Although substrate-based explanations are not viable for certain pathological states, including eating disorders, mood disorders, and personality disorders, they are routinely applied to hallucination, delusion, and thought-insertion. In examining the scope of substrate-based explanations, I plan to ask what it is about these latter symptoms that make substrate-based explanations of them so tempting. In another strand of work, I also intend to look at substrate-based explanations in the literature on delusion, where we find argumentative moves similar to those present in the AVH literature. Some theorists have argued that delusions are beliefs (Bortolotti, 2009), states of imagination (Currie and Ravenscroft, 2002), or states between imagination and belief (Egan, 2009). Similar to the case of AVH, I suggest that the normal mental state kinds invoked in the literature on delusion are at most attempts to gain a sympathetic understanding of delusion in terms of more familiar states, but do not provide us with an explication of the nature of delusion.

By thinking of inner speech as essentially a speech phenomenon, the literature on the topic has been able to entertain only a limited range of ways that inner speech intersects with topics in other domains. However, once we dislodge the dominant picture, we are able to investigate a

larger, more diverse swath of views and topics than the ones that currently have sway over discussions of inner speech.

Bibliography

Abbs, James H. 1986. “Invariance and Variability in Speech Production: A Distinction between Linguistic Intent and Its Neuromotor Implementation.” In *Invariance and Variability in Speech Processes*, 202–25. Hillsdale, NJ, US: Lawrence Erlbaum Associates, Inc.

Aggernæs, A. 1972. “The Experienced Reality of Hallucinations and Other Psychological Phenomena.” *Acta Psychiatrica Scandinavica* 48 (3): 220–38.

Alderson-Day, Ben, and Charles Fernyhough. 2015. “Inner Speech: Development, Cognitive Functions, Phenomenology, and Neurobiology.” *Psychological Bulletin* 141 (5): 931–65.

Aleman, A., and E. H. de Haan. 1998. “On Redefining Hallucination.” *The American Journal of Orthopsychiatry* 68 (4): 656–59.

Aleman, André, Koen B. E. Böcker, Ron Hijman, Edward H. F. de Haan, and René S. Kahn. 2003. “Cognitive Basis of Hallucinations in Schizophrenia: Role of Top-down Information Processing.” *Schizophrenia Research* 64 (2–3): 175–85.

Alexander, Jessica D., and Lynne C. Nygaard. 2008. “Reading Voices and Hearing Text: Talker-Specific Auditory Imagery in Reading.” *Journal of Experimental Psychology. Human Perception and Performance* 34 (2): 446–59.

American Psychiatric Association. 2013. *Diagnostic and Statistical Manual of Mental Disorders, 5th Edition: DSM-5*. 5 edition. Washington, D.C: American Psychiatric Publishing.

Anonymous. 2016. “People with Auditory Hallucinations Question.” Thread. *Schizophrenia.Com*. https://forum.schizophrenia.com/t/people-with-auditory-hallucinations-question/53126.

———. 2020. “19yo m Suffering from Cannibalism and Murder Based Intrusive Thoughts.” Thread. *Reddit*.

https://www.reddit.com/r/OCD/comments/eqjjga/19yo_m_suffering_from_cannibalism_and_murder/.

Anscombe, G. E. M. 1957. *Intention*. 2 edition. Cambridge, Mass: Harvard University Press.

Aristotle. 1984. *Complete Works of Aristotle, Vol. 1*. Edited by Jonathan Barnes. Bollingen Series edition. Princeton, NJ: Princeton University Press.

Awwad Shiekh Hasan, Bashar, Mitchell Valdes-Sosa, Joachim Gross, and Pascal Belin. 2016. "'Hearing Faces and Seeing Voices': Amodal Coding of Person Identity in the Human Brain." *Scientific Reports* 6 (1): 37494.

Bailes, Freya. 2007. "The Prevalence and Nature of Imagined Music in the Everyday Lives of Music Students." *Psychology of Music - PSYCHOL MUSIC* 35 (August): 555–70.

Bar-On, Dorit, and Jordan Ochs. 2018. "The Role of Inner Speech in Self-Knowledge: Against Neo-Rylean Views." *Teorema* 37 (1): 5–22.

Behroozmand, Roozbeh, and Charles R. Larson. 2011. "Error-Dependent Modulation of Speech-Induced Auditory Suppression for Pitch-Shifted Voice Feedback." *BMC Neuroscience* 12 (1): 54.

Belin, Pascal. 2019. "The 'Vocal Brain': Core and Extended Cerebral Networks for Voice Processing." In *The Oxford Handbook of Voice Perception*. Oxford University Press.

Belin, Pascal, Shirley Fecteau, and Catherine Bédard. 2004. "Thinking the Voice: Neural Correlates of Voice Perception." *Trends in Cognitive Sciences* 8 (3): 129–35.

Belin, Pascal, Robert J. Zatorre, Philippe Lafaille, Pierre Ahad, and Bruce Pike. 2000. "Voice-Selective Areas in Human Auditory Cortex." *Nature* 403 (6767): 309–12.

Bell, C. C. 1989. "Sensory Coding and Corollary Discharge Effects in Mormyrid Electric Fish." *Journal of Experimental Biology* 146 (1): 229–53.

Bermudez, Jose Luis. 2018. “Inner Speech, Determinacy, and Thinking Conciously about Thoughts.” In *Inner Speech: New Voices*. Oxford University Press.

Berrios, G. E., and T. R. Dening. 1996. “Pseudohallucinations: A Conceptual History.” *Psychological Medicine* 26 (4): 753–63.

Bishop, D. V. M., and J. Robson. 1989. “Accurate Non-Word Spelling despite Congenital Inability to Speak: Phoneme—Grapheme Conversion Does Not Require Subvocal Articulation.” *British Journal of Psychology* 80 (1): 1–13.

Bishop, Laura, Freya Bailes, and Roger T. Dean. 2013. “Musical Imagery and the Planning of Dynamics and Articulation During Performance.” *Music Perception: An Interdisciplinary Journal* 31 (2): 97–117.

Bock, Kathryn, and Willem Levelt. 1994. “Language Production: Grammatical Encoding.” In *Handbook of Psycholinguistics*, 945–84. San Diego, CA, US: Academic Press.

Bortolotti, Lisa. 2010. *Delusions and Other Irrational Beliefs*. 1 edition. Oxford ; New York: Oxford University Press, USA.

Brown, Steven. 2006. “The Perpetual Music Track.” *Journal of Conciousness Studies* 13 (6): 25–44.

Bunn, William, and Jan Terpstra. 2009. “Cultivating Empathy for the Mentally Ill Using Simulated Auditory Hallucinations.” *Academic Psychiatry: The Journal of the American Association of Directors of Psychiatric Residency Training and the Association for Academic Psychiatry* 33 (6): 457–60.

Byrne, Alex. 2011. “Knowing That I Am Thinking.” In *Self-Knowledge*. Oxford: Oxford University Press.

Campbell, John. 2013. “Causation and Mechanisms in Psychiatry.” In *The Oxford Handbook of Philosophy and Psychiatry*, edited by K.W.M Fulford, Martin Davies, R.G.T Gipps, George Graham, G Stanghellini, and Tim Thornton. Oxford: Oxford University Press.

Carruthers, Peter. 2011. *The Opacity of Mind: An Integrative Theory of Self-Knowledge*. 1 edition. Oxford ; New York: Oxford University Press.

———. 2018. “The Causes and Contents of Inner Speech.” In *Inner Speech: New Voices*, edited by Peter Langland-Hassan and Agustin Vicente. Oxford University Press.

Cassidy, Clifford M., Peter D. Balsam, Jodi J. Weinstein, Rachel J. Rosengard, Mark Slifstein, Nathaniel D. Daw, Anissa Abi-Dargham, and Guillermo Horga. 2018. “A Perceptual Inference Mechanism for Hallucinations Linked to Striatal Dopamine.” *Current Biology: CB* 28 (4): 503-514.e4.

Cho, Raymond, and Wayne Wu. 2013. “Mechanisms of Auditory Verbal Hallucination in Schizophrenia.” *Frontiers in Psychiatry* 4 (November).

———. 2014. “Is Inner Speech the Basis of Auditory Verbal Hallucination in Schizophrenia?” *Frontiers in Psychiatry* 5 (July).

Clark, Andy. 2013. “The Many Faces of Precision (Replies to Commentaries on ‘Whatever next? Neural Prediction, Situated Agents, and the Future of Cognitive Science’).” *Frontiers in Psychology* 4.

———. 2019. *Surfing Uncertainty: Prediction, Action, and the Embodied Mind*. Reprint edition. Oxford University Press.

Cooper, J.M., and D.S. Hutchinson, eds. 1997. *Plato: Complete Works*. Indianapolis: Hackett Publishing Company, Inc.

Cox, Trevor. 2018. *Now You're Talking: Human Conversation from the Neanderthals to Artificial Intelligence*. Berkeley, California: Counterpoint.

Craig, Tom KJ, Mar Rus-Calafell, Thomas Ward, Julian P. Leff, Mark Huckvale, Elizabeth Howarth, Richard Emsley, and Philippa A. Garety. 2018. "AVATAR Therapy for Auditory Verbal Hallucinations in People with Psychosis: A Single-Blind, Randomised Controlled Trial." *The Lancet Psychiatry* 5 (1): 31–40.

Curio, Gabriel, Georg Neuloh, Jussi Numminen, Veikko Jousmäki, and Riitta Hari. 2000. "Speaking Modifies Voice-Evoked Activity in the Human Auditory Cortex." *Human Brain Mapping* 9 (4): 183–91.

Currie, Gregory, and Ian Ravenscroft. 2002. *Recreative Minds: Imagination in Philosophy and Psychology*. Oxford University Press.

Dehaene, Stanislas. 2008. "Conscious and Nonconscious Processes: Distinct Forms of Evidence Accumulation?" In *Better than Conscious? Decision Making, the Human Mind, and Implications for Institutions*, 21–49. Strüngmann Forum Reports. Cambridge, MA, US: MIT Press.

Dell, Gary S. 1986. "A Spreading-Activation Theory of Retrieval in Sentence Production." *Psychological Review* 93 (3): 283–321.

Dretske, Fred I. 1999. *Knowledge and the Flow of Information*. New edition edition. Stanford, CA: Center for the Study of Language and Inf.

Egan, Andy. 2008. "Imagination, Delusion, and Self-Deception." In *Delusion and Self-Deception: Affective and Motivational Influences on Belief Formation*, edited by Tim Bayne and Jordi Fernandez. Macquarie Monographs in Cognitive Science. Psychology Press.

Eich, Eric, and Joseph P. Forgas. 2003. “Mood, Cognition, and Memory.” In *Handbook of Psychology,* 61–83. American Cancer Society.

Epstein, Russell A., Whitney E. Parker, and Alana M. Feiler. 2008. “Two Kinds of FMRI Repetition Suppression? Evidence for Dissociable Neural Mechanisms.” *Journal of Neurophysiology* 99 (6): 2877–86.

Farah, Martha J. 1985. “Psychophysical Evidence for a Shared Representational Medium for Mental Images and Percepts.” *Journal of Experimental Psychology: General* 114 (1): 91–103.

Feldman, Harriet, and Karl J. Friston. 2010. “Attention, Uncertainty, and Free-Energy.” *Frontiers in Human Neuroscience* 4.

Fletcher, Paul C., and Chris D. Frith. 2009. “Perceiving Is Believing: A Bayesian Approach to Explaining the Positive Symptoms of Schizophrenia.” *Nature Reviews. Neuroscience* 10 (1): 48–58.

Fodor, Jerry A. 1968. *Psychological Explanation: An Introduction to the Philosophy of Psychology*. New York: Random House.

Ford, J. M., D. H. Mathalon, S. Kalba, S. Whitfield, W. O. Faustman, and W. T. Roth. 2001a. “Cortical Responsiveness during Talking and Listening in Schizophrenia: An Event-Related Brain Potential Study.” *Biological Psychiatry* 50 (7): 540–49.

Ford, Judith M., Daniel H. Mathalon, Sontine Kalba, Susan Whitfield, William O. Faustman, and Walton T. Roth. 2001b. “Cortical Responsiveness During Inner Speech in Schizophrenia: An Event-Related Potential Study.” *American Journal of Psychiatry* 158 (11): 1914–16.

Ford, Judith M., and Daniel H. Mathalon. 2004. “Electrophysiological Evidence of Corollary Discharge Dysfunction in Schizophrenia during Talking and Thinking.” *Journal of Psychiatric Research* 38 (1): 37–46.

Frankish, Keith. 2004. *Mind and Supermind*. Cambridge: Cambridge University Press.

Friston, Karl. 2005. “A Theory of Cortical Responses.” *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 360 (1456): 815–36.

Frith, Chris. 2011. “Explaining Delusions of Control: The Comparator Model 20 Years On.” *Consciousness and Cognition* 21 (July): 52–54.

Frith, Christopher D. 1992. *The Cognitive Neuropsychology of Schizophrenia*. The Cognitive Neuropsychology of Schizophrenia. Hillsdale, NJ, US: Lawrence Erlbaum Associates, Inc.

Fromkin, Victoria. 1971. “The Non-Anomalous Nature of Anomalous Utterances.” *Language* 47 (March).

Fruhholz, Sascha, and Pascal Belin. 2019. *The Oxford Handbook of Voice Perception*. Oxford Handbooks Online. Oxford, New York: Oxford University Press.

Fulford, K.W.M, Martin Davies, R.G.T Gipps, George Graham, G Stanghellini, and Tim Thornton, eds. 2013. *The Oxford Handbook of Philosophy and Psychiatry*. Oxford University Press.

Garrett, Michael, and Raul Silva. 2003. “Auditory Hallucinations, Source Monitoring, and the Belief That ‘Voices’ Are Real.” *Schizophrenia Bulletin* 29 (3): 445–57.

Gauker, Christopher. 2018. “Inner Speech as the Internalization of Outer Speech.” In *Inner Speech: New Voices*, edited by Peter Langland-Hassan and Agustin Vicente, 53–77. Oxford: Oxford University Press.

Geva, Sharon, P. Simon Jones, Jenny T. Crinion, Cathy J. Price, Jean-Claude Baron, and Elizabeth A. Warburton. 2011. “The Neural Correlates of Inner Speech Defined by Voxel-Based Lesion–Symptom Mapping.” *Brain* 134 (10): 3071–82.

Green, Lance B. 1992. “The Use of Imagery in the Rehabilitation of Injured Athletes.” *The Sport Psychologist* 6 (4): 416–28.

Grotheer, Mareike, and Gyula Kovács. 2015. “The Relationship between Stimulus Repetitions and Fulfilled Expectations.” *Neuropsychologia* 67 (January): 175–82.

———. 2016. “Can Predictive Coding Explain Repetition Suppression?” *Cortex: A Journal Devoted to the Study of the Nervous System and Behavior* 80: 113–24.

Grush, Rick. 2004. “The Emulation Theory of Representation: Motor Control, Imagery, and Perception.” *The Behavioral and Brain Sciences* 27 (3): 377–96; discussion 396-442.

Guenther, Frank H., Michelle Hampson, and Dave Johnson. 1998. “A Theoretical Investigation of Reference Frames for the Planning of Speech Movements.” *Psychological Review* 105 (4): 611–33.

Halpern, Andrea R. 1988. “Perceived and Imagined Tempos of Familiar Songs.” *Music Perception: An Interdisciplinary Journal* 6 (2): 193–202.

Heavey, Christopher L., and Russell T. Hurlburt. 2008. “The Phenomena of Inner Experience.” *Consciousness and Cognition* 17 (3): 798–810.

Heinks-Maldonado, Theda H., Daniel H. Mathalon, Max Gray, and Judith M. Ford. 2005. “Fine-Tuning of Auditory Cortex during Speech Production.” *Psychophysiology* 42 (2): 180–90.

Heinks-Maldonado, Theda H., Srikantan S. Nagarajan, and John F. Houde. 2006. “Magnetoencephalographic Evidence for a Precise Forward Model in Speech Production.” *Neuroreport* 17 (13): 1375–79.

Hemphill, R. E., and E. Stengel. 1940. “A Study on Pure Word-Deafness.” *Journal of Neurology and Psychiatry* 3 (3): 251–62.

Hoffman, R. E., M. Varanko, J. Gilmore, and A. L. Mishara. 2008. “Experiential Features Used by Patients with Schizophrenia to Differentiate ‘voices’ from Ordinary Verbal Thought.” *Psychological Medicine* 38 (8): 1167–76.

Hohwy, Jakob. 2013. *The Predictive Mind*. 1 edition. OUP Oxford.

Hohwy, Jakob, Andreas Roepstorff, and Karl Friston. 2008. “Predictive Coding Explains Binocular Rivalry: An Epistemological Review.” *Cognition* 108 (3): 687–701.

Hubbard, Edward M., and V. S. Ramachandran. 2005. “Neurocognitive Mechanisms of Synesthesia.” *Neuron* 48 (3): 509–20.

Hume, David. 1993. *An Enquiry Concerning Human Understanding: With Hume’s Abstract of A Treatise of Human Nature and A Letter from a Gentleman to His Friend in Edinburgh*. Edited by Eric Steinberg. Second Edition,2 edition. Indianapolis: Hackett Publishing Company, Inc.

Hunter, M. D., S. B. Eickhoff, T. W. R. Miller, T. F. D. Farrow, I. D. Wilkinson, and P. W. R. Woodruff. 2006. “Neural Activity in Speech-Sensitive Auditory Cortex during Silence.” *Proceedings of the National Academy of Sciences* 103 (1): 189–94.

Hunter, Michael, and Peter Woodruff. 2004. “Characteristics of Functional Auditory Hallucinations.” *The American Journal of Psychiatry* 161 (June): 923.

Hurlburt, Russell T. 2011. *Investigating Pristine Inner Experience: Moments of Truth*. New York: Cambridge University Press.

Hurlburt, Russell T., and Christopher L. Heavey. 2018. “Inner Speech as Pristine Inner Experience.” In *Inner Speech: New Voices*, edited by Peter Langland-Hassan and Agustin Vicente. Oxford: Oxford University Press.

Hurlburt, Russell T., Christopher L. Heavey, and Jason M. Kelsey. 2013. “Toward a Phenomenology of Inner Speaking.” *Consciousness and Cognition* 22 (4): 1477–94.

Intons-Pererson, Margaret Jean. 1980. “The Role of Loudness in Auditory Imagery.” *Memory & Cognition* 8 (5): 385–93.

Ishimura, G. n.d. “Visuomotor Factors for Action Capture.” *Investigative Ophtalomogy and Visual Science (Supplement)* 36.

Itti, L., and C. Koch. 2001. “Computational Modelling of Visual Attention.” *Nature Reviews. Neuroscience* 2 (3): 194–203.

Itti, Laurent, and Pierre Baldi. 2009. “Bayesian Surprise Attracts Human Attention.” *Vision Research*, Visual Attention: Psychophysics, electrophysiology and neuroimaging, 49 (10): 1295–1306.

Jack, Bradley N., Mike E. Le Pelley, Nathan Han, Anthony W. F. Harris, Kevin M. Spencer, and Thomas J. Whitford. 2019. “Inner Speech Is Accompanied by a Temporally-Precise and Content-Specific Corollary Discharge.” *NeuroImage* 198: 170–80.

Janata, Petr, and Kaivon Paroo. 2006. “Acuity of Auditory Images in Pitch and Time.” *Perception & Psychophysics* 68 (5): 829–44.

Jaspers, Karl. 1963. *Karl Jaspers’ General Psychopathology*. The University of Chicago Press.

———. 2009. “Causal and Understandable Relationships between Events and Psychosis in Dementia Praecox (Schizophrenia).” In *Anthology of German Psychiatric Texts*, edited by Henning Sass. John Wiley & Sons.

Jeannerod, M. 2006. *Motor Cognition: What Actions Tell to the Self*. Oxford Psychology Series. Oxford, New York: Oxford University Press.

Jones, Simon R. 2010. "Do We Need Multiple Models of Auditory Verbal Hallucinations? Examining the Phenomenological Fit of Cognitive and Neurological Models." *Schizophrenia Bulletin* 36 (3): 566–75.

Jones, Simon R., and Charles Fernyhough. 2007. "Thought as Action: Inner Speech, Self-Monitoring, and Auditory Verbal Hallucinations." *Consciousness and Cognition* 16 (2): 391–99.

Joyce, James. 2015. *Ulysses*. Edited by Declan Kiberd. New Ed edition. Penguin.

Junginger, J., and C. L. Frame. 1985. "Self-Report of the Frequency and Phenomenology of Verbal Hallucinations." *The Journal of Nervous and Mental Disease* 173 (3): 149–55.

Kapur, Shitij. 2003. "Psychosis as a State of Aberrant Salience: A Framework Linking Biology, Phenomenology, and Pharmacology in Schizophrenia." *The American Journal of Psychiatry* 160 (1): 13–23.

Kaya, Emine Merve, and Mounya Elhilali. 2014. "Investigating Bottom-up Auditory Attention." *Frontiers in Human Neuroscience* 8.

Kell, Christian A., Maritza Darquea, Marion Behrens, Lorenzo Cordani, Christian Keller, and Susanne Fuchs. 2017. "Phonetic Detail and Lateralization of Reading-Related Inner Speech and of Auditory and Somatosensory Feedback Processing during Overt Reading." *Human Brain Mapping* 38 (1): 493–508.

Kendler, K. S. 2012. "The Dappled Nature of Causes of Psychiatric Illness: Replacing the Organic-Functional/Hardware-Software Dichotomy with Empirically Based Pluralism." *Molecular Psychiatry* 17 (4): 377–88.

Kendler, K. S., and J. Campbell. 2014. “Expanding the Domain of the Understandable in Psychiatric Illness: An Updating of the Jasperian Framework of Explanation and Understanding.” *Psychological Medicine* 44 (1): 1–7.

Kirchhoff, Michael D. 2018. “Predictive Processing, Perceiving and Imagining: Is to Perceive to Imagine, or Something Close to It?” *Philosophical Studies* 175 (3): 751–67.

Klinger, Eric, and W. Miles Cox. 1987. “Dimensions of Thought Flow in Everyday Life.” *Imagination, Cognition and Personality* 7 (2).

Kurby, Christopher A., Joseph P. Magliano, and David N. Rapp. 2009. “Those Voices in Your Head: Activation of Auditory Images during Reading.” *Cognition* 112 (3): 457–61.

Langdon, R., S. R. Jones, E. Connaughton, and C. Fernyhough. 2009. “The Phenomenology of Inner Speech: Comparison of Schizophrenia Patients with Auditory Verbal Hallucinations and Healthy Controls.” *Psychological Medicine* 39 (4): 655–63.

Langland-Hassan, Peter. 2008. “Fractured Phenomenologies: Thought Insertion, Inner Speech, and the Puzzle of Extraneity.” *Mind & Language* 23 (4): 369–401.

———. 2014. “Inner Speech and Metacognition: In Search of a Connection.” *Mind & Language* 29 (5): 511–33.

———. 2018. “From Introspection to Essence: The Auditory Nature of Inner Speech.” In *Inner Speech: New Voices,* edited by Peter Langland-Hassan and Agustin Vicente. Oxford University Press.

———. forthcoming. “Inner Speech: The Big Future of the Little Voice in the Head.” *WIREs,* Wiley Interdisciplinary Reviews.

Langland-Hassan, Peter, and Agustin Vicente. 2018a. “Introduction.” In *Inner Speech: New Voices*, edited by Peter Langland-Hassan and Agustin Vicente. Oxford University Press.

———. 2018b. *Inner Speech: New Voices*. Oxford ; New York, NY: Oxford University Press.

Latinus, Marianne, Phil McAleer, Patricia E. G. Bestelmeyer, and Pascal Belin. 2013. "Norm-Based Coding of Voice Identity in Human Auditory Cortex." *Current Biology* 23 (12): 1075–80.

Lavan, Nadine, Sarah Knight, and Carolyn McGettigan. 2019. "Listeners Form Average-Based Representations of Individual Voice Identities." *Nature Communications* 10 (1): 2404.

Levelt, Willem J. M. 1993. *Speaking: From Intention to Articulation*. Cambridge, Mass.: MIT Press.

Levine, Joseph. 1983. "Materialism and Qualia: The Explanatory Gap." *Pacific Philosophical Quarterly* 64 (4): 354–61.

Lewis, Sinclair. 2010. *Babbitt*. 1 edition. Oxford ; New York: Oxford University Press.

Liberman, Alvin M., and Ignatius G. Mattingly. 1985. "The Motor Theory of Speech Perception Revised." *Cognition* 21 (1): 1–36.

Lim, Anastasia, Hans W. Hoek, Mathijs L. Deen, Jan Dirk Blom, Richard Bruggeman, Wiepke Cahn, Lieuwe de Haan, et al. 2016. "Prevalence and Classification of Hallucinations in Multiple Sensory Modalities in Schizophrenia Spectrum Disorders." *Schizophrenia Research* 176 (2): 493–99.

Lipton, Peter. 2009. "Understanding without Explanation." In *Scientific Understanding: Philosophical Perspectives*, edited by Henk W. de Regt, Sabina Leonelli, and Kai Eigner. University of Pittsburgh.

Lœvenbruck, H., R. Grandchamp, L. Rapin, L. Nalborczyk, M. Dohen, P. Perrier, M. Baciu, and M. Perrone-Bertolotti. 2018. "A Cognitive Neuroscience View of Inner Language: To

Predict and to Hear, See, Feel." In *Inner Speech: New Voices*, edited by Peter Langland-Hassan and Agustin Vicente. Oxford University Press.

Lu, Lingxi, Changxin Zhang, and Liang Li. 2017. "Mental Imagery of Face Enhances Face-Sensitive Event-Related Potentials to Ambiguous Visual Stimuli." *Biological Psychology* 129 (October): 16–24.

MacKay, Donald G. 1992. "Constraints on Theories of Inner Speech." In *Auditory Imagery*, 121–49. Hillsdale, NJ, US: Lawrence Erlbaum Associates, Inc.

Maiese, Michelle. 2018. "Auditory Verbal Hallucination and the Sense of Ownership." *Philosophy, Psychiatry, and Psychology* 25 (3): 183–196.

Marshall, Robert C, B. Z Rappaport, and Luis Garcia-Bunuel. 1985. "Self-Monitoring Behavior in a Case of Severe Auditory Agnosia with Aphasia." *Brain and Language* 24 (2): 297–313.

McGuigan, F. J., and A. B. Dollins. 1989. "Patterns of Covert Speech Behavior and Phonetic Coding." *The Pavlovian Journal of Biological Science* 24 (1): 19–26.

McGuire, P. K., D. Robertson, A. Thacker, A. S. David, N. Kitson, R. S. Frackowiak, and C. D. Frith. 1997. "Neural Correlates of Thinking in Sign Language." *Neuroreport* 8 (3): 695–98.

McGurk, H., and J. MacDonald. 1976. "Hearing Lips and Seeing Voices." *Nature* 264 (5588): 746–48.

McNamara, Timothy P., and Jon B. Holbrook. 2003. "Semantic Memory and Priming." In *Handbook of Psychology*, 445–74. John Wiley & Sons, Ltd.

Mintz, S., and M. Alpert. 1972. "Imagery Vividness, Reality Testing, and Schizophrenic Hallucinations." *Journal of Abnormal Psychology* 79 (3): 310–16.

Moritz, Steffen, and Frank Larøi. 2008. “Differences and Similarities in the Sensory and Cognitive Signatures of Voice-Hearing, Intrusions and Thoughts.” *Schizophrenia Research* 102 (1–3): 96–107.

Moseley, Peter, and Sam Wilkinson. 2014. “Inner Speech Is Not so Simple: A Commentary on Cho and Wu (2013).” *Frontiers in Psychiatry* 5.

Navarro Cebrian, Ana, and Petr Janata. 2010. “Electrophysiological Correlates of Accurate Mental Image Formation in Auditory Perception and Imagery Tasks.” *Brain Research* 1342 (June): 39–54.

Nayani, T. H., and A. S. David. 1996. “The Auditory Hallucination: A Phenomenological Survey.” *Psychological Medicine* 26 (1): 177–89.

Niziolek, C. A., S. S. Nagarajan, and J. F. Houde. 2013. “What Does Motor Efference Copy Represent? Evidence from Speech Production.” *Journal of Neuroscience* 33 (41): 16110–16.

Oppenheim, Gary M., and Gary S. Dell. 2008. “Inner Speech Slips Exhibit Lexical Bias, but Not the Phonemic Similarity Effect.” *Cognition* 106 (1): 528–37.

———. 2010. “Motor Movement Matters: The Flexible Abstractness of Inner Speech.” *Memory & Cognition* 38 (8): 1147–60.

Panikkath, Ragesh, Deepa Panikkath, Deb Mojumder, and Kenneth Nugent. 2014. “The Alien Hand Syndrome.” *Proceedings (Baylor University. Medical Center)* 27 (3): 219–20.

Pearson, Joel, Colin W. G. Clifford, and Frank Tong. 2008. “The Functional Impact of Mental Imagery on Conscious Perception.” *Current Biology: CB* 18 (13): 982–86.

Perrone-Bertolotti, M., L. Rapin, J. P. Lachaux, M. Baciu, and H. Lœvenbruck. 2014. “What Is That Little Voice inside My Head? Inner Speech Phenomenology, Its Role in Cognitive

Performance, and Its Relation to Self-Monitoring." *Behavioural Brain Research* 261 (March): 220–39.

Perrone-Bertolotti, Marcela, Jan Kujala, Juan R. Vidal, Carlos M. Hamame, Tomas Ossandon, Olivier Bertrand, Lorella Minotti, Philippe Kahane, Karim Jerbi, and Jean-Philippe Lachaux. 2012. "How Silent Is Silent Reading? Intracerebral Evidence for Top-Down Activation of Temporal Voice Areas during Reading." *Journal of Neuroscience* 32 (49): 17554–62.

Pickard, Hanna. 2010. "Schizophrenia and the Epistemology of Self-Knowledge." *European Journal of Analytic Philosophy* 6 (1): 55–74.

Pickering, Martin J., and Andy Clark. 2014. "Getting Ahead: Forward Models and Their Place in Cognitive Architecture." *Trends in Cognitive Sciences* 18 (9): 451–56.

Pickering, Martin J., and Simon Garrod. 2013. "An Integrated Theory of Language Production and Comprehension." *The Behavioral and Brain Sciences* 36 (4): 329–47.

Pitt, M. A., and R. G. Crowder. 1992. "The Role of Spectral and Dynamic Cues in Imagery for Musical Timbre." *Journal of Experimental Psychology. Human Perception and Performance* 18 (3): 728–38.

Postma Albert, and Kolk Herman. 1993. "The Covert Repair Hypothesis." *Journal of Speech, Language, and Hearing Research* 36 (3): 472–87.

Potochnik, Angela. 2017. *Idealization and the Aims of Science*. University of Chicago Press.

Pulvermüller, Friedemann, Martina Huss, Ferath Kherif, Fermin Moscoso del Prado Martin, Olaf Hauk, and Yury Shtyrov. 2006. "Motor Cortex Maps Articulatory Features of Speech Sounds." *Proceedings of the National Academy of Sciences* 103 (20): 7865–70.

Rapin, Isabelle. 1985. “Cortical Deafness, Auditory Agnosia, and Word-Deafness: How Distinct Are They?” *Human Communication Canada* 9 (4): 10.

Ratcliffe, Matthew. 2013. “Delusional Atmosphere and the Sense of Unreality.” In *One Century of Karl Jaspers’ General Psychopathology*, edited by Giovanni Stanghellini and Thomas Fuchs. Oxford University Press.

Regt, Henk W. de, Sabina Leonelli, and Kai Eigner, eds. 2009. *Scientific Understanding: Philosophical Perspectives*. University of Pittsburgh.

Robinson, Howard. 2001. *Perception*. 1 edition. London: Routledge.

Roelofs, A., A. S. Meyer, and W. J. Levelt. 1998. “A Case for the Lemma/Lexeme Distinction in Models of Speaking: Comment on Caramazza and Miozzo (1997).” *Cognition* 69 (2): 219–30.

Rosen, Cherise, Simon McCarthy-Jones, Kayla A. Chase, Clara Humpston, Jennifer K. Melbourne, Leah Kling, and Rajiv P. Sharma. 2018. “The Tangled Roots of Inner Speech, Voices and Delusions.” *Psychiatry Research* 264 (June): 281–89.

Rudner, Mary, Thomas Karlsson, Johan Gunnarsson, and Jerker Rönnberg. 2013. “Levels of Processing and Language Modality Specificity in Working Memory.” *Neuropsychologia* 51 (4): 656–66.

Ryle, Gilbert. 2000. *The Concept of Mind*. Chicago: University of Chicago Press.

Sass, Henning, ed. 2009. *Anthology of German Psychiatric Texts*. John Wiley & Sons.

Savariaux, C., P. Perrier, J. P. Orliaguet, and J. L. Schwartz. 1999. “Compensation Strategies for the Perturbation of French [u] Using a Lip Tube. II. Perceptual Analysis.” *The Journal of the Acoustical Society of America* 106 (1): 381–93.

Schafer, E. W., and M. M. Marcus. 1973. “Self-Stimulation Alters Human Sensory Brain Responses.” *Science (New York, N.Y.)* 181 (4095): 175–77.

Schwartz, Bennett L. 2001. *Tip-of-the-Tongue States: Phenomenology, Mechanism, and Lexical Retrieval*. Psychology Press.

Scott, Mark, H. Henny Yeung, Bryan Gick, and Janet F. Werker. 2013. “Inner Speech Captures the Perception of External Speech.” *The Journal of the Acoustical Society of America* 133 (4): EL286–92.

Seal, Marc, Andre Aleman, and Philip McGuire. 2004. “Compelling Imagery, Unanticipated Speech and Deceptive Memory: Neurocognitive Models of Auditory Verbal Hallucinations in Schizophrenia.” *Cognitive Neuropsychiatry* 9 (1–2): 43–72.

Seth, Anil K. 2014. “A Predictive Processing Theory of Sensorimotor Contingencies: Explaining the Puzzle of Perceptual Presence and Its Absence in Synesthesia.” *Cognitive Neuroscience* 5 (2): 97–118.

Shadmehr, Reza, and John W. Krakauer. 2008. “A Computational Neuroanatomy for Motor Control.” *Experimental Brain Research* 185 (3): 359–81.

Shannon, C. E. 1948. “A Mathematical Theory of Communication.” *Bell System Technical Journal* 27 (3): 379–423.

Southwell, Rosy, Anna Baumann, Cécile Gal, Nicolas Barascud, Karl Friston, and Maria Chait. 2017. “Is Predictability Salient? A Study of Attentional Capture by Auditory Patterns.” *Philosophical Transactions of the Royal Society B: Biological Sciences* 372 (1714).

Stanghellini, Giovanni, and Thomas Fuchs, eds. 2013. *One Century of Karl Jaspers’ General Psychopathology*. 1 edition. Oxford: Oxford University Press.

Sterzer, Philipp, Rick A. Adams, Paul Fletcher, Chris Frith, Stephen M. Lawrie, Lars Muckli, Predrag Petrovic, Peter Uhlhaas, Martin Voss, and Philip R. Corlett. 2018. “The Predictive Coding Account of Psychosis.” *Biological Psychiatry* 84 (9): 634–43.

Summerfield, Christopher, Jim M.P. Monti, Emily H. Trittschuh, M.-Marsel Mesulam, and Tobias Egner. 2008. “Neural Repetition Suppression Reflects Fulfilled Perceptual Expectations.” *Nature Neuroscience* 11 (9): 1004–6.

Swiney, Lauren. 2018. “Activity, Agency, and Inner Speech Pathology.” In *Inner Speech: New Voices,* edited by Peter Langland-Hassan and Agustin Vicente. Oxford University Press.

Swiney, Lauren, and Paulo Sousa. 2014a. “A New Comparator Account of Auditory Verbal Hallucinations: How Motor Prediction Can Plausibly Contribute to the Sense of Agency for Inner Speech.” *Frontiers in Human Neuroscience* 8: 675.

———. 2014b. “A New Comparator Account of Auditory Verbal Hallucinations: How Motor Prediction Can Plausibly Contribute to the Sense of Agency for Inner Speech.” *Frontiers in Human Neuroscience* 8: 675.

Szpunar, Karl K., Peggy L. St Jacques, Clifford A. Robbins, Gagan S. Wig, and Daniel L. Schacter. 2014. “Repetition-Related Reductions in Neural Activity Reveal Component Processes of Mental Simulation.” *Social Cognitive and Affective Neuroscience* 9 (5): 712–22.

Theeuwes, Jan, and Erik Van der Burg. 2013. “Priming Makes a Stimulus More Salient.” *Journal of Vision* 13 (3): 21–21.

Tian, Xing, and David Poeppel. 2010. “Mental Imagery of Speech and Movement Implicates the Dynamics of Internal Forward Models.” *Frontiers in Psychology* 1: 166.

———. 2013. “The Effect of Imagination on Stimulation: The Functional Specificity of Efference Copies in Speech Processing.” *Journal of Cognitive Neuroscience* 25 (7): 1020–36.

Titze, Ingo R. 1989. “Physiologic and Acoustic Differences between Male and Female Voices.” *The Journal of the Acoustical Society of America* 85 (4): 1699–1707.

Todorovic, Ana, and Floris P. de Lange. 2012. “Repetition Suppression and Expectation Suppression Are Dissociable in Time in Early Auditory Evoked Fields.” *Journal of Neuroscience* 32 (39): 13389–95.

Tomiczek, Caroline, and Richard Stevenson. 2009a. “Olfactory Imagery and Repetition Priming: The Effect of Odor Naming and Imagery Ability.” *Experimental Psychology* 56 (February): 397–408.

Tovar, Antonia, Paola Fuentes-Claramonte, Joan Soler-Vidal, Nuria Ramiro-Sousa, Alfonso Rodriguez-Martinez, Carmen Sarri-Closa, Salvador Sarró, et al. 2019. “The Linguistic Signature of Hallucinated Voice Talk in Schizophrenia.” *Schizophrenia Research* 206: 111–17.

Tsantani, Maria, Nikolaus Kriegeskorte, Carolyn McGettigan, and Lúcia Garrido. 2019. “Faces and Voices in the Brain: A Modality-General Person-Identity Representation in Superior Temporal Sulcus.” *NeuroImage* 201.

Vercammen, Ans, Henderikus Knegtering, Johann A. den Boer, Edith J. Liemburg, and André Aleman. 2010. “Auditory Hallucinations in Schizophrenia Are Associated with Reduced Functional Connectivity of the Temporo-Parietal Area.” *Biological Psychiatry* 67 (10): 912–18.

Vygotsky, Lev S. 1986. *Thought and Language - Revised Edition*. Edited by Alex Kozulin. Revised edition edition. Cambridge, Mass: The MIT Press.

Waters, Flavie A. V., Johanna C. Badcock, Patricia T. Michie, and Murray T. Maybery. 2006. "Auditory Hallucinations in Schizophrenia: Intrusive Thoughts and Forgotten Memories." *Cognitive Neuropsychiatry* 11 (1): 65–83.

Whitford, Thomas J, Bradley N Jack, Daniel Pearson, Oren Griffiths, David Luque, Anthony WF Harris, Kevin M Spencer, and Mike E Le Pelley. 2017. "Neurophysiological Evidence of Efference Copies to Inner Speech." Edited by Richard Ivry. *ELife* 6 (December): e28197.

Wilkenfeld, Daniel A. 2019. "Understanding as Compression." *Philosophical Studies* 176 (10): 2807–31.

Wilkinson, Sam. 2014. "Accounting for the Phenomenology and Varieties of Auditory Verbal Hallucination within a Predictive Processing Framework." *Consciousness and Cognition* 30 (November): 142–55.

Wilkinson, Sam, and Charles Fernyhough. 2017. "Auditory Verbal Hallucinations and Inner Speech : A Predictive Processing Perspective." In *Before Consciousness : In Search of the Fundamentals of Mind.*, edited by Zdravko Radman, 285–304. Exeter: Imprint Academic.

Wolpert, D. M., and J. R. Flanagan. 2001. "Motor Prediction." *Current Biology: CB* 11 (18): R729-732.

Woods, Angela, Nev Jones, Ben Alderson-Day, Felicity Callard, and Charles Fernyhough. 2015. "Experiences of Hearing Voices: Analysis of a Novel Phenomenological Survey." *The Lancet Psychiatry* 2 (4): 323–31.

Wu, Wayne. 2012. "Explaining Schizophrenia: Auditory Verbal Hallucination and Self-Monitoring." *Mind & Language* 27 (1): 86–107.

Yao, Bo, Pascal Belin, and Christoph Scheepers. 2011. “Silent Reading of Direct versus Indirect Speech Activates Voice-Selective Areas in the Auditory Cortex.” *Journal of Cognitive Neuroscience* 23 (10): 3146–52.

Ylinen, Sari, Anni Nora, Alina Leminen, Tero Hakala, Minna Huotilainen, Yury Shtyrov, Jyrki P. Mäkelä, and Elisabet Service. 2015. “Two Distinct Auditory-Motor Circuits for Monitoring Speech Production as Revealed by Content-Specific Suppression of Auditory Cortex.” *Cerebral Cortex* 25 (6): 1576–86.